

# Smoothed functional-based gradient algorithms for off-policy reinforcement learning: A non-asymptotic viewpoint

Nithia Vijayan\* and Prashanth L. A.<sup>†</sup>

*Department of Computer Science and Engineering, Indian Institute of Technology Madras, India*

## Abstract

We propose two policy gradient algorithms for solving the problem of control in an off-policy reinforcement learning (RL) context. Both algorithms incorporate a smoothed functional (SF) based gradient estimation scheme. The first algorithm is a straightforward combination of importance sampling-based off-policy evaluation with SF-based gradient estimation. The second algorithm, inspired by the stochastic variance-reduced gradient (SVRG) algorithm, incorporates variance reduction in the update iteration. For both algorithms, we derive non-asymptotic bounds that establish convergence to an approximate stationary point. From these results, we infer that the first algorithm converges at a rate that is comparable to the well-known REINFORCE algorithm in an off-policy RL context, while the second algorithm exhibits an improved rate of convergence.

## 1 Introduction

In a reinforcement learning (RL) problem, an agent learns to achieve a goal through interactions with an environment. The interactions between the agent and the environment are represented as a Markov decision process (MDP). The agent interacts with the environment through actions, and as a response the environment changes its state and provides a reward. The goal of the agent is to maximize the cumulative reward over time by learning an optimal policy to choose actions.

We consider the problem of control in an off-policy RL setting, where the agent aims to learn an optimal

policy using the data collected by executing an exploratory policy called behavior policy. Off-policy RL is useful in practical scenarios where the system may not allow execution of any policy other than a fixed behavior policy. While the behavior policy may not be optimal, it can be exploratory, and aids in the search for the optimal policy.

Policy gradient algorithms [39, 38, 25, 20, 7, 31, 40, 42, 1] are a popular approach for solving MDPs. In a few special cases such as linear systems with quadratic cost, policy gradient algorithms can be shown to be globally convergent [4, 13, 28]. In the general case, the usual convergence guarantees for a policy gradient algorithm are to a stationary point of the underlying value function (cf. [31, 42]). In [26, 1, 5], the authors analyze policy gradient methods in the idealized setting where the gradient information is made directly available, while we consider a typical off-policy RL setting where the gradient of the objective has to be estimated from a sample path of the behavior policy. Most of the previous works use the likelihood ratio method, proposed in [33], see [15, 16] for an introduction. This approach for estimating the policy gradient was first used in a policy optimization context in the REINFORCE algorithm [39]. REINFORCE style gradient estimate methods are analyzed in [41, 23]. While [41] uses log barrier regularization, [23] analyzes a natural and variance-reduced counterparts of the policy gradient algorithm. The likelihood ratio method leads to unbiased estimates of the policy gradient.

An alternative approach for gradient estimation is the simultaneous perturbation method, see [8] for a textbook introduction. This method is based on finite differences, and results in a biased estimate of the policy gradient. A popular algorithm in this class is simultaneous perturbation stochastic approximation (SPSA), proposed in [36]. Using the classic finite difference

\*nithiav@cse.iitm.ac.in

<sup>†</sup>prashla@cse.iitm.ac.in

type estimate of the policy gradient, i.e., a scheme that perturbs each co-ordinate separately, would require  $2d$  function measurements, where  $d$  is the dimension of the policy parameter. On the other hand, the SPSA scheme used random perturbations, e.g., a vector of independent Rademacher random variables (r.v.s), to simultaneously perturb all co-ordinates, and this scheme would work with two function measurements, irrespective of the dimension. SPSA has been used in a policy gradient algorithm in [7, 6]. Smoothed functional (SF) [21, 29] is another simultaneous perturbation method, where one could employ a vector of independent standard Gaussian r.v.s as random perturbations.

In this paper, we propose two policy gradient algorithms for off-policy control. For the purpose of policy evaluation, both algorithms use the importance sampling ratios — a standard scheme for unbiased off-policy evaluation. Unlike previous works on off-policy RL, our algorithms incorporate a SF-based gradient estimate scheme. We use the two function measurements variant of SF, which is equivalent to evaluating two perturbed policies. In an on-policy RL setting, SF-based approach may be restrictive owing to the fact that running two system trajectories corresponding to two perturbed policies may not be feasible in some practical applications. On the other hand, using a SF-based policy gradient scheme does not run into practical difficulties in an off-policy RL context, since the system is simulated using a single behavior policy.

The first algorithm, henceforth referred to as OffP-SF, is a straightforward combination of importance sampling-based off-policy evaluation with SF-based gradient estimation. The second algorithm is inspired by the SVRG algorithm, which was proposed in [19] for optimizing finite ‘strongly convex’ sum of smooth functions, and later adapted to a non-convex optimization setting (cf. [32, 2]). This algorithm, referred to as OffP-SF-SVRG, is the variance-reduced variant of the OffP-SF algorithm. To the best of our knowledge, a variance-reduced policy gradient algorithm inspired by SVRG has not been proposed/analyzed in an off-policy RL context in the literature, while SVRG has been explored in the context of on-policy RL in [31, 40]. Recent work in [24] explores variance reduction in an off-policy context inspired by a momentum-based method [10].

In this paper, we focus on the non-asymptotic performance of the proposed algorithms. The results for policy gradient methods employing simultaneous perturbation-based gradient estimates are asymptotic in nature (cf. [7, 6]). On the other hand, using ideas from zeroth-order optimization, policy gradient methods with

REINFORCE style gradient estimates have been shown to converge to an  $\epsilon$ -stationary point (see Definition 1 below) in the non-asymptotic regime. In this paper, we study policy gradient algorithms with the simultaneous perturbation approach, and derive non-asymptotic bounds for these algorithms — see Table 1 for a summary of our bounds in terms of iteration complexity, which is the number of policy gradient iterations required to find an  $\epsilon$ -stationary point. The primary conclusions from our non-asymptotic analysis are as follows: (i) After  $N$  iterations of OffP-SF, the value function gradient at a suitably chosen iterate, say  $\theta_R$ , satisfies an order  $O(\frac{1}{\sqrt{N}})$  bound on  $\mathbb{E}||\nabla J(\theta_R)||^2$ ; (ii) The corresponding bound for OffP-SF-SVRG is of the order  $O(\frac{1}{N})$ .

Table 1: Iteration complexity for our proposed algorithms, and the off-policy variant of REINFORCE. Here iteration complexity denotes the number of iterations required to find an  $\epsilon$ -stationary point (see Definition 1).

Algorithm	Iteration complexity
REINFORCE (off-policy variant <sup>1</sup> )	$O(1/\epsilon^2)$
OffP-SF	$O(1/\epsilon^2)$
OffP-SF-SVRG	$O(1/\epsilon)$

Our bounds have a few advantages over those in the literature for zeroth-order optimization and on-policy RL using policy gradient algorithms. To elaborate, the closest result to the non-asymptotic bound for offP-SF is Corollary 3.3 of [18]. For setting the step-size/perturbation constant in this result, one requires knowledge of quantities that are typically unknown in an RL setting. On the other hand, our non-asymptotic bound features a universal step-size/perturbation constant. In arriving at this result, we depart from the argument employed in the proof of Corollary 3.3 of [18]. Our bound for OffP-SF in Corollary 1 is comparable to the one provided in Corollary 4.4 in [42], as both results are on the size of the gradient of the objective  $J$  at a suitably chosen policy iterate. We employ smoothed functional based gradient estimation, while the authors in [42] use a REINFORCE style gradient estimate. Their result is for a diminishing step-size, while we employ a constant step-size. Next, the gradient estimates underlying the SVRG-based on-policy RL algorithms in

<sup>1</sup>This variant uses importance sampling ratios for off-policy evaluation, and the likelihood ratio method for gradient estimation.

[31, 40] use the likelihood ratio method, which result in unbiased estimates. On the other hand, our OffP-SF-SVRG algorithm employs smoothed functional-based gradient estimates, which are biased in nature. Through a careful handling of the bias terms in several steps of the proof, we are able to obtain an order  $O(\frac{1}{N})$  bound for the OffP-SF-SVRG algorithm. The corresponding results for on-policy SVRG algorithm in [31, 40] features additional terms — see the discussion below Theorem 3 for more details.

In [11, 43, 44], the authors propose actor-critic algorithms in an off-policy RL setting. In comparison, we do not incorporate function approximation in our proposed algorithms, and hence, a direct comparison is not feasible. Nevertheless, we mention that the algorithms in these references involve at least two timescales, and to the best of our knowledge, there are no non-asymptotic bounds for two timescale stochastic approximation, with a non-linear update iteration (as in the case of the actor update in the aforementioned references). In contrast, our algorithms operate on a single timescale, facilitating a non-asymptotic analysis. In [23], the authors establish global convergence results for natural and variance-reduced counterparts of the policy gradient algorithm, with REINFORCE style gradient estimates. These results are under an assumption that the underlying policy parameterization is sufficiently rich. In contrast, we study local convergence properties of the vanilla and variance-reduced variants of the policy gradient algorithm, with smoothed functional-based gradient estimates. Finally, our non-asymptotic bound for OffP-SF-SVRG shows improved dependence on the number of iterations, as compared to the bound in [24], where the authors analyze a momentum-based variance reduced policy gradient scheme in an off-policy context.

The rest of the paper is organized as follows: Section 2 describes the off-policy control problem. Section 3 introduces our algorithms, namely OffP-SF and OffP-SF-SVRG. Section 4 presents the non-asymptotic bounds for our algorithms. Section 5 provides detailed proofs of convergence. Finally, Section 7 provides the concluding remarks.

## 2 Problem formulation

We consider an MDP with a state space  $\mathcal{S}$ , and an action space  $\mathcal{A}$ , both assumed to be finite. We operate in an episodic setting with a random episode length  $T \in \mathbb{N}$ . At time  $t \in \{0, \dots, T-1\}$ , the MDP is in state  $S_t$ , and transitions to state  $S_{t+1}$  by an action  $A_t$  chosen by a behavior policy  $b$ , and receives a reward  $R_{t+1} \in \mathbb{R}$ . We assume the rewards are bounded, the start state  $S_0$  is

fixed. We also assume a special state 0 as a termination state.

Let  $\Theta$  be a compact and convex subset of  $\mathbb{R}^d$ . We consider parameterized stochastic target policies  $\{\pi_\theta, \theta \in \Theta\}$ , where  $\pi_\theta(a|s) = \mathbb{P}\{A_t = a | S_t = s, \theta_t = \theta\}$ . As an example, one may use an exponential softmax distribution, i.e.,

$$\pi_\theta(a|s) = \frac{\exp(h(s, a, \theta))}{\sum_{b \in \mathcal{A}} \exp(h(s, b, \theta))},$$

where  $h : \mathcal{S} \times \mathcal{A} \times \Theta \rightarrow \mathbb{R}$  is a parameterized user defined function (cf. Chapter 13 of [37]). We assume that each policy in the parameterized class  $\Theta$ , and the behavior policy are proper (see (A3)).

We assume that the MDP trajectory terminates under  $\pi_\theta$  w.p. 1,  $\forall \theta \in \Theta$ . The goal here is to find  $\theta^*$  such that

$$\theta^* \in \operatorname{argmax}_{\theta \in \Theta} J(\theta), \quad (1)$$

where  $J(\theta)$  is the value function, and is defined as

$$J(\theta) = \mathbb{E}_{\pi_\theta} \left[ \sum_{t=0}^{T-1} \gamma^t R_{t+1} \right], \quad (2)$$

where  $\gamma \in (0, 1]$  is the discount factor.

## 3 Off-policy gradient algorithms

A gradient-based algorithm for solving (1) would involve the following update iteration:

$$\theta_{k+1} = \Pi_\Theta(\theta_k + \alpha_k \nabla J(\theta_k)), \quad (3)$$

where  $\theta_0$  is set arbitrarily, and  $\nabla$  is with respect to  $\theta$ . In the above, the step-size  $\alpha_k \in (0, 1]$ , and  $\Pi_\Theta : \mathbb{R}^d \rightarrow \Theta$  is an operator that projects on to  $\Theta$ . The projection is required to ensure stability of the iterates in (3), and is common in the analysis of policy gradient algorithms (cf. [7]). As an example, one may define  $\Theta = \prod_{i=1}^d [\theta_{\min}^i, \theta_{\max}^i]$ . Then, the projection operator  $\Pi_\Theta(\theta) = [\Pi_\Theta^1(\theta^1), \dots, \Pi_\Theta^d(\theta^d)]$ , where  $\Pi_\Theta^i(\theta^i) = \min(\max(\theta_{\min}^i, \theta^i), \theta_{\max}^i)$ ,  $i \in \{1 \dots d\}$ . It is easy to see that such a projection operation is computationally inexpensive.

We describe two algorithms for solving (1) below.

### 3.1 OffP-SF

In an off-policy setting, the distribution of data (states/actions seen along a sample path) follows the behavior policy. The off-policy evaluation problem is to learn the value of a target policy, which is different

from the behavior one. A standard off-policy evaluation scheme is per-decision importance sampling (see Section 5.9 of [37]). Here, one scales the objective by the likelihood ratio of the target policy, say  $\pi_\theta$  to the behavior policy, say  $b$  at the current state. More precisely, we generate  $m$  episodes using  $b$  and estimate  $J(\theta)$  as follows:

$$\hat{J}_m(\theta) = \frac{1}{m} \sum_{n=1}^m \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right). \quad (4)$$

In the above,  $T^n$  is the length of the  $n^{th}$  episode, and  $R_{t+1}^n$  is the reward at time  $t+1$ . Also  $S_i^n$  is the state, and  $A_i^n$  is the action taken at time  $i$  of the  $n^{th}$  episode.

For estimating the gradient  $\nabla J(\cdot)$ , we employ the estimation scheme from [21, 29]. The idea here is to form a smoothed functional, denoted by  $J_\mu$ , of the value  $J(\cdot)$ , and use  $\nabla J_\mu$  as a proxy for  $\nabla J$ . To be more precise, the smoothed functional  $J_\mu(\theta)$  is defined by

$$J_\mu(\theta) = \mathbb{E}_{u \in \mathbb{B}^d} [J(\theta + \mu u)], \quad (5)$$

where  $\mu \in (0, 1]$  is the smoothing parameter, and  $u$  is sampled uniformly at random from a unit ball  $\mathbb{B}^d = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ . Here  $\|\cdot\|$  denotes the  $d$ -dimensional Euclidean norm.

We estimate the gradient using two randomly perturbed policies (cf. [22, 34]). We favor the ‘balanced’ estimate based on two random perturbations instead of a one-sided estimate because the bound on the second moment of the balanced estimate exhibits a linear dependence on the underlying dimension  $d$  (see Lemma 8), while the corresponding dependence in an one-sided estimate is quadratic in  $d$  (see Proposition 7.6 of [17]).

We perturb the policy parameter  $\theta$  by adding and subtracting a scalar multiple of a random unit vector  $v$ . The perturbed policy parameters lie in the set  $\Theta'$  defined as follows:

$$\Theta' = \{\theta' : \|\theta' - \theta\| \leq 1, \theta \in \Theta\}. \quad (6)$$

In order to control the variance, we average the gradient estimate over  $n$  random unit vectors. The estimate  $\hat{\nabla}_{n,\mu} \hat{J}_m(\theta)$  of the gradient  $\nabla J(\cdot)$  is formed as follows:

$$\hat{\nabla}_{n,\mu} \hat{J}_m(\theta) = \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}_m(\theta + \mu v_i) - \hat{J}_m(\theta - \mu v_i)}{2\mu} v_i, \quad (7)$$

where  $\forall i, v_i$  is sampled uniformly at random from a unit sphere  $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$ .

We collect  $m$  sample paths using the behavior policy  $b$ , and use this data to estimate the value associated with the  $2n$  perturbed policies in (7).

---

#### Algorithm 1 OffP-SF

---

- 1: **Input:** Parameterized form of target policy  $\pi$  and behavior policy  $b$ , iteration limit  $N$ , step-sizes  $\{\alpha_k\}$ , perturbation constants  $\{\mu_k\}$ , batch size  $m$ ,  $\{n_k\}$ , and probability mass function (pmf)  $P_R(\cdot)$  supported on  $\{1, \dots, N\}$ ;
  - 2: **Initialize:** Target policy parameter  $\theta_1 \in \mathbb{R}^d$ , and the discount factor  $\gamma \in (0, 1]$ ;
  - 3: **for**  $k = 0, \dots, N - 1$  **do**
  - 4:   **for**  $j = 1, \dots, m$  **do**
  - 5:     Get  $(S_0^j, A_0^j, R_1^j, \dots, S_{T_j-1}^j, A_{T_j-1}^j, R_{T_j}^j) \sim b$ ;
  - 6:   **end for**
  - 7:   **for**  $i = 1, \dots, n_k$  **do**
  - 8:     Get  $[v_i^1, \dots, v_i^d] \in \mathbb{S}^{d-1}$ ;
  - 9:     Use (4) to estimate  $\hat{J}_m(\theta_k \pm \mu_k v_i)$ ;
  - 10:   **end for**
  - 11:   Use (7) to estimate  $\hat{\nabla}_{n_k, \mu_k} \hat{J}_m(\theta_k)$ ;
  - 12:   Use (8) to calculate  $\theta_{k+1}$ ;
  - 13: **end for**
  - 14: **Output:** Policy  $\theta_R$  where  $R \sim P_R$ .
- 

We solve (1) using the following update iteration:

$$\theta_{k+1} = \Pi_\Theta(\theta_k + \alpha_k \hat{\nabla}_{n_k, \mu_k} \hat{J}_m(\theta_k)). \quad (8)$$

Algorithm 1 presents the pseudocode of OffP-SF algorithm, with the following ingredients: (i) a gradient ascent update according to (8); (ii) a SF-based gradient estimation scheme; and (iii) an importance sampling-based policy evaluation scheme.

### 3.2 OffP-SF-SVRG

Our second algorithm is a modification of the Algorithm 1 that incorporates the concept of variance reduction seen in SVRG algorithms [19, 32]. The principle of variance reduction underlying the SVRG algorithm has been explored in the context of on-policy RL in [31, 40]. The gradient estimates underlying the algorithms in the aforementioned references use the likelihood ratio method, which results in unbiased estimates. On the other hand, we employ SF-based gradient estimates, which are biased in nature.

We use nested update iterations to solve (1). Our algorithm maintains an outer loop over  $s$ , and an inner loop over  $k$ . Our policy parameters are of the form  $\theta_k^s$ , where  $\theta_0^0$  is set arbitrarily.

In the outer loop, we sample  $m$  episodes using the behavior policy  $b$ . We use a reference point  $\tilde{\theta}^s \in \Theta$ , which is initialized to  $\theta_0^0$ , and is updated as  $\tilde{\theta}^{s+1} = \theta_m^s$ . We calculate  $\hat{J}_m(\tilde{\theta}^s)$  and  $\hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s)$  using (4) and (7) respectively. Also  $\forall j \in \{1, \dots, m\}$ , we calculate  $\hat{J}^j(\tilde{\theta}^s)$

and  $\hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s)$ , where

$$\hat{J}^j(\theta) = \sum_{t=0}^{T^j-1} \gamma^t R_{t+1}^j \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^j | S_i^j)}{b(A_i^j | S_i^j)} \right), \quad (9)$$

and

$$\hat{\nabla}_{n,\mu} \hat{J}^j(\theta) = \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}^j(\theta + \mu v_i) - \hat{J}^j(\theta - \mu v_i)}{2\mu} v_i. \quad (10)$$

In the above,  $\forall i, v_i$  is sampled uniformly at random from a unit sphere  $\mathbb{S}^{d-1}$ .

In the inner loop, we pick a sample  $j$  uniformly at random from  $\{1, \dots, m\}$  and calculate  $\hat{J}^j(\theta_k^s)$  and  $\hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s)$  using (9) and (10) respectively.

We update the policy parameters as follows:

$$\theta_{k+1}^s = \Pi_\Theta(\theta_k^s + \alpha g_k^s), \quad (11)$$

where

$$g_k^s = \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) + \hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s). \quad (12)$$

Algorithm 2 presents the pseudocode of OffP-SF-SVRG algorithm.

#### 4 Main results

We make the following assumptions for the sake of analysis:

**(A1).** For any  $a \in \mathcal{A}$  and  $s \in \mathcal{S}$ ,  $\log \pi_\theta(a|s)$  exists, and is twice continuously differentiable w.r.t.  $\theta \in \Theta'$ , where  $\Theta'$  is defined in (6).

**(A2).** For every  $\theta \in \Theta'$ , the target policy  $\pi_\theta$  is absolutely continuous with respect to the behavior policy  $b$ . i.e.,

$$\forall \theta \in \Theta', b(a|s)=0 \Rightarrow \pi_\theta(a|s)=0, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}.$$

**(A3).** The behavior policy  $b$ , and the class of target policies  $\{\pi_\theta, \theta \in \Theta'\}$  are proper, i.e., there exists a positive constant  $M$  s.t.

$$\begin{aligned} &\forall \theta \in \Theta', \max_{s \in \mathcal{S}} \mathbb{P}(S_M \neq 0 \mid S_0 = s, \pi_\theta) < 1, \text{ and} \\ &\max_{s \in \mathcal{S}} \mathbb{P}(S_M \neq 0 \mid S_0 = s, b) < 1. \end{aligned}$$

An assumption like (A1) is common to the analysis of policy gradient algorithms (cf. [31, 40]), while

---

#### Algorithm 2 OffP-SF-SVRG

---

```

1: Input: Parameterized form of target policy  $\pi$  and
   behavior policy  $b$ , iteration limit  $S$ , step-size  $\alpha$ ,
   perturbation constant  $\mu$ , batch sizes  $m, n$ , and a
   joint pmf  $P_{QR}(\cdot, \cdot)$  supported on  $\{1, \dots, S\}$  and
    $\{1, \dots, m\}$  respectively;
2: Initialize: Target policy parameter  $\tilde{\theta}^0 = \theta_0^0 \in \mathbb{R}^d$ ,
   and the discount factor  $\gamma \in (0, 1]$ ;
3: for  $s = 0, \dots, S - 1$  do
4:   for  $j = 1, \dots, m$  do
5:     Get  $(S_0^j, A_0^j, R_1^j, \dots, S_{T_j-1}^j, A_{T_j-1}^j, R_{T_j}^j) \sim b$ ;
6:   end for
7:   for  $i = 1, \dots, n$  do
8:     Get  $[v_i^1, \dots, v_i^d] \in \mathbb{S}^{d-1}$ ;
9:     for  $j = 1, \dots, m$  do
10:      Use (9) to estimate  $\hat{J}^j(\tilde{\theta}^s \pm \mu v_i)$ ;
11:    end for
12:    Use (4) to estimate  $\hat{J}_m(\tilde{\theta}^s \pm \mu v_i)$ ;
13:  end for
14:  for  $j = 1, \dots, m$  do
15:    Use (10) to estimate  $\hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s)$ ;
16:  end for
17:  Use (7) to estimate  $\hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s)$ ;
18:  for  $k = 0, \dots, m - 1$  do
19:    Get  $j \in [1, m]$  uniformly and at random.
20:    for  $i = 1, \dots, n$  do
21:      Use (9) to estimate  $\hat{J}^j(\theta_k^s \pm \mu v_i)$ ;
22:    end for
23:    Use (10) to estimate  $\hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s)$ ;
24:    Use (12) to calculate  $g_k^s$ ;
25:    Use (11) to calculate  $\theta_{k+1}^s$ ;
26:  end for
27:   $\tilde{\theta}^{s+1} = \theta_0^{s+1} = \theta_m^s$ ;
28: end for
29: Output: Policy  $\theta_R^Q$  where  $Q, R \sim P_{QR}$ .

```

---

(A2) is a standard requirements for off-policy evaluation. Further, (A3) is a common requirement in the analysis of episodic MDPs, see Chapter 2 of [3]. From (A1) and (A2), we have  $\pi_\theta(a|s) > 0$  and  $b(a|s) > 0$ ,  $\forall \theta \in \mathbb{R}^d, \forall a \in \mathcal{A}$ , and  $\forall s \in \mathcal{S}$ . In other words, we consider policies that place a positive mass on every action in any state.

The objective  $J$  is not necessarily convex in a typical RL setting, and hence, several previous works (cf. [42, 31, 40, 35]) adopt convergence to an approximate stationary point, which is defined below.

**Definition 1. ( $\epsilon$ -stationary point)** Let  $\theta_R$  be the output of an algorithm. Then,  $\theta_R$  is called an  $\epsilon$ -stationary point of problem (1), if  $\mathbb{E} \|\nabla J(\theta_R)\|^2 \leq \epsilon$ .

The non-asymptotic bounds for Algorithms 1–2 that we present below establish convergence to an  $\epsilon$ -stationary point.

For the non-asymptotic analysis, we rewrite the update rule in (8) as follows:

$$\theta_{k+1} = \theta_k + \alpha_k \mathcal{P}_\Theta(\theta_k, \hat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k), \alpha_k), \quad (13)$$

where

$$\mathcal{P}_\Theta(\theta, f(\theta), \alpha) = \frac{1}{\alpha} [\Pi_\Theta(\theta + \alpha f(\theta)) - \theta]. \quad (14)$$

**Theorem 1 (OffP-SF: Non-asymptotic bound).** Assume (A1)–(A3). Let  $P_R(k) = \mathbb{P}(R = k) = \frac{\alpha_k}{\sum_{k=0}^{N-1} \alpha_k}$ ,  $\forall N \in \mathbb{N}$ , and  $J^* = \max_{\theta \in \Theta} J(\theta)$ . Then,

$$\begin{aligned} & \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha_R)\|^2 \right] \\ & \leq \frac{(J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k}{\sum_{k=0}^{N-1} \alpha_k} \\ & \quad + \frac{\frac{2dL^3}{c_0} \sum_{k=0}^{N-1} \frac{\alpha_k^2}{n_k} + \frac{2\sqrt{d}L^2}{\sqrt{c_0}} \sum_{k=0}^{N-1} \frac{\alpha_k}{\sqrt{n_k}}}{\sum_{k=0}^{N-1} \alpha_k}, \end{aligned} \quad (15)$$

where  $c_0$  is an absolute positive constant, and  $L$  is the Lipschitz constant of  $J$  as well as  $\nabla J$  (see Lemma 4 in Section 5 below).

*Proof.* See Section 5.  $\square$

The result above holds for any choice of step-sizes  $\{\alpha_k\}$ , perturbation constants  $\{\mu_k\}$ , and batch sizes  $m$ ,  $\{n_k\}$ . We specialize the bound in (15) for a particular choice of the aforementioned parameters in the corollary below.

**Corollary 1 (OffP-SF: Non-asymptotic bound).** Let  $\forall k$ ,  $\alpha_k = \frac{c_1}{\sqrt{N}}$ ,  $\mu_k = \frac{c_2}{\sqrt{N}}$ ,  $n_k = c_3 N$ , and  $m < \infty$  for some absolute constants  $c_1, c_2, c_3 > 0$ . Then, under conditions of Theorem 1, we have

$$\begin{aligned} & \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha_R)\|^2 \right] \\ & \leq \frac{(J^* - J(\theta_0)) + c' L^2 (d + \sqrt{d})}{\sqrt{N}} + \frac{c'' d L^3}{N \sqrt{N}}, \end{aligned}$$

for some constants  $c', c'' > 0$ .

*Proof.* See Section 5.  $\square$

**Remark 1.** Ignoring the error due to projection, i.e., assuming  $\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha_R) = \nabla J(\theta_R)$ , the bound above can be read as follows: after  $N$  iterations of (3), offP-SF returns an iterate that satisfies  $\mathbb{E} \|\nabla J(\theta_R)\|^2 = O\left(\frac{1}{\sqrt{N}}\right)$ . The closest result in a zeroth-order smooth

non-convex optimization context is Corollary 3.3 of [18]. In comparison to this result, our bound has a few advantages. First, the step-size in Corollary 1 is set using a universal constant, while they require the knowledge of the smoothness parameter  $L$ . Second, the perturbation constant in Corollary 1 is set using a universal constant, while the corresponding choice in [18] requires the knowledge of  $J^* - J(\theta_0)$ . In a typical RL setting, one could possibly approximate  $L$ , but  $J^* - J(\theta_0)$  is usually unknown.

In REINFORCE, which is a well-known policy gradient algorithm, the gradient estimation scheme is based on the likelihood ratio method. In principle, one could employ importance sampling-based policy evaluation together with a REINFORCE style gradient estimate.

**Theorem 2 (REINFORCE (off-policy variant): Non-asymptotic bound).** Assume (A1)–(A3). Let  $\mathbb{P}(R = k) = \frac{\alpha_k}{\sum_{k=0}^{N-1} \alpha_k}$ ,  $\forall N \in \mathbb{N}$ ,  $J^* = \max_{\theta \in \Theta} J(\theta)$ , and  $\alpha = \frac{1}{\sqrt{N}}$ . Then,

$$\begin{aligned} & \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha)\|^2 \right] \\ & \leq \frac{(J^* - J(\theta_0)) + \frac{L^2}{2}}{\sqrt{N}} + L^2 \end{aligned}$$

**Remark 2.** It is apparent that the result that we derived in Corollary 1 is comparable to REINFORCE in an off-policy RL framework, which lets us conclude that SF-based gradient estimation is a viable alternative to the likelihood ratio method.

Now, we present a non-asymptotic bound for Algorithm 2. For the analysis, we rewrite the update rule in (11) as follows:

$$\theta_{k+1}^s = \theta_k^s + \alpha \mathcal{P}_\Theta(\theta_k, g_k^s, \alpha), \quad (16)$$

where  $\mathcal{P}_\Theta(\cdot, \cdot, \cdot)$  is as defined in (14).

**Theorem 3 (OffP-SF-SVRG: Non-asymptotic bound).** Assume (A1)–(A3). Let  $P_{QR}(s, k) = \mathbb{P}(Q = s, R = k) = \frac{\alpha}{\sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha}$ ,  $\forall (S, m) \in \mathbb{N} \times \mathbb{N}$ , and  $J^* = \max_{\theta \in \Theta} J(\theta)$ . Let  $\alpha = \frac{1}{4dL}$ ,  $\mu = \frac{1}{\sqrt{Sm}}$ , and  $n = Sm^2$ . Then,

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_R^Q, \nabla J(\theta_R^Q), \alpha) \right\|^2 \right] \\ & \leq \frac{4dL (J^* - J(\theta_0^0))}{Sm} + \frac{L^2 (15c_0 d^2 + 40ed + 2e)}{10c_0 Sm} \end{aligned}$$

$$+ \frac{20dL^2(8+e) + L^2(5+e)}{5c_0Sm^2}, \quad (17)$$

where  $c_0$  is an absolute positive constant,  $e$  is the Euler's number, and  $L$  is the Lipschitz constant of  $J$  as well as  $\nabla J$ .

*Proof.* See Section 5.  $\square$

**Remark 3.** As mentioned earlier, SVRG has been employed in an on-policy RL context in [31, 40]. Unlike these works, we operate in an off-policy RL setting, and more importantly, use a biased gradient estimation scheme that is based on the idea of smoothed functionals. Through a careful handling of the bias terms in several steps of the proof, we are able to obtain an order  $O(\frac{1}{Sm})$  bound for the OffP-SF-SVRG algorithm. The bound in [31] is of the form  $O(1/Sm) + O(1/n) + O(1/B)$ , where  $B$  is the mini-batch size used for averaging in their inner-loop. In comparison, we obtain an order  $O(1/Sm)$  without additional terms, and our algorithm does not require simulation of system trajectories for mini-batching owing to the fact that we operate in the off-policy setting. In other words, the on-policy setting of [31, 40] implies  $n$  system trajectories are simulated in the outer loop, while we obtain an  $n$ -sample average of the gradient estimate using off-policy evaluation. Next, the bound in [40] is of the form  $O(1/Sm) + O(1/n)$ , while our bound is without the additional  $O(1/n)$  term, since we can choose  $n = Sm^2$  without requiring additional simulations.

**Remark 4.** In [24], the authors explore an alternative approach to variance reduction of a policy gradient algorithm in an off-policy context. The authors obtain a non-asymptotic bound of the order  $O(1/T^{2/3})$ , where  $T$  is the number of iterations of the policy gradient algorithm. In comparison, we obtain an improved bound of  $O(1/T)$  in Theorem 3 above.

## 5 Convergence analysis

### 5.1 Proofs for OffP-SF

Our analysis proceeds through a sequence of lemmas. We begin with a result that is well-known in the context of off-policy RL (cf. Chapter 5 of [37]). We have provided the proof for the sake of completeness.

**Lemma 1.**  $\mathbb{E}_b [\hat{J}_m(\theta)] = J(\theta)$ .

*Proof.* Notice that

$$\mathbb{E}_b [\hat{J}_m(\theta)]$$

$$\begin{aligned} &= \mathbb{E}_b \left[ \frac{1}{m} \sum_{n=1}^m \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \right] \\ &= \frac{1}{m} \sum_{n=1}^m \mathbb{E}_b \left[ \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \right] \\ &= \frac{1}{m} \sum_{n=1}^m \mathbb{E}_{\pi_\theta} \left[ \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \right] = J(\theta). \end{aligned}$$

$\square$

**Lemma 2.**  $\Theta' = \{\theta' : \|\theta' - \theta\| \leq 1, \theta \in \Theta\}$  is compact.

*Proof.* Since  $\Theta$  is compact,  $\exists \theta_c \in \Theta$ , and  $r \in \mathbb{R}$  such that  $\Theta \subseteq B(\theta_c, r)$ , where  $B(\theta_c, r)$  is an open ball centered at  $\theta_c$  with radius  $r$ . The set  $B[\theta_c, r+1]$  is a closed and bounded subset of  $\mathbb{R}^d$ , and hence compact. It is easy to see that  $\Theta' \subseteq B[\theta_c, r+1]$ . Using the fact that  $\Theta$  is closed, and the definition of  $\Theta'$ , it is easy to see that  $\Theta'$  is closed. Since every closed subset of a compact set is compact,  $\Theta'$  is compact.  $\square$

**Lemma 3.** For any  $m \geq 1$ , there exists a constant  $L > 0$  such that the following conditions hold w.p. 1 for any  $\theta_1, \theta_2 \in \Theta'$ :

$$\begin{aligned} \|\hat{J}_m(\theta_1) - \hat{J}_m(\theta_2)\| &\leq L\|\theta_1 - \theta_2\|, \\ \|\nabla \hat{J}_m(\theta_1) - \nabla \hat{J}_m(\theta_2)\| &\leq L\|\theta_1 - \theta_2\|. \end{aligned}$$

*Proof.* For any twice differentiable function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^+$ , the Hessian  $\nabla^2 f(\cdot)$  can be defined as

$$\nabla^2 f(x) = f(x) [\nabla^2 \log f(x) + \nabla \log f(x) \nabla \log f(x)^\top].$$

Using the above equation and (A1), we obtain

$$\begin{aligned} &\nabla^2 \prod_{i=0}^t \pi_\theta(A_i | S_i) \\ &= \left( \prod_{i=0}^t \pi_\theta(A_i | S_i) \right) \left[ \nabla^2 \log \prod_{i=0}^t \pi_\theta(A_i | S_i) \right. \\ &\quad \left. + \left[ \nabla \log \prod_{i=0}^t \pi_\theta(A_i | S_i) \right] \left[ \nabla \log \prod_{i=0}^t \pi_\theta(A_i | S_i) \right]^\top \right] \\ &= \left( \prod_{i=0}^t \pi_\theta(A_i | S_i) \right) \left[ \sum_{i=0}^t \nabla^2 \log \pi_\theta(A_i | S_i) \right. \\ &\quad \left. + \left[ \sum_{i=0}^t \nabla \log \pi_\theta(A_i | S_i) \right] \left[ \sum_{i=0}^t \nabla \log \pi_\theta(A_i | S_i) \right]^\top \right]. \end{aligned} \quad (18)$$

From (4), we obtain

$$\begin{aligned}
& \nabla^2 \hat{J}_m(\theta) \\
&= \frac{1}{m} \sum_{n=1}^m \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \left( \prod_{i=0}^t \frac{1}{b(A_i^n | S_i^n)} \right) \\
&\quad \times \nabla^2 \left( \prod_{i=0}^t \pi_\theta(A_i^n | S_i^n) \right) \\
&= \frac{1}{m} \sum_{n=1}^m \sum_{t=0}^{T^n-1} \gamma^t R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \\
&\quad \times \left[ \sum_{i=0}^t \nabla^2 \log \pi_\theta(A_i^n | S_i^n) \right] \\
&\quad + \left[ \sum_{i=0}^t \nabla \log \pi_\theta(A_i^n | S_i^n) \right] \left[ \sum_{i=0}^t \nabla \log \pi_\theta(A_i^n | S_i^n) \right]^\top,
\end{aligned}$$

where the last equality follows from (18). Observe that the RHS above is a finite sum of continuous functions, since  $\nabla^2 \log \pi_\theta(\cdot | \cdot)$  is continuous w.r.t  $\theta$  (see (A1)), the rewards  $R_{t+1}^n$  are bounded, the policy  $b$  is proper (see (A3)), and  $m$  is finite. Thus,  $\nabla^2 \hat{J}_m(\theta)$  is continuous which in turn implies  $\nabla \hat{J}_m(\theta)$  is continuous. Further, since  $\Theta'$  is compact, from Lemma 2, we have

$$\begin{aligned}
& \|\nabla^2 \hat{J}_m(\theta)\| \leq \|\nabla^2 \hat{J}_m(\theta)\|_F \leq L_1, \text{ and} \\
& \|\nabla \hat{J}_m(\theta)\| \leq L_2, \forall \theta \in \Theta',
\end{aligned}$$

for some constants  $L_1, L_2 < \infty$ . In the above,  $\|A\|$  and  $\|A\|_F$  denote the operator and Frobenius norm of a  $d \times d$  matrix  $A$ .

Let  $L = \max(L_1, L_2)$ . Then the result follows by Lemma 1.2.2 in [30].  $\square$

**Lemma 4.**  $J(\theta)$  and  $\nabla J(\theta)$  are  $L$ -Lipschitz w.r.t.  $\theta \in \Theta'$ .

*Proof.* Notice that

$$\begin{aligned}
& \|J(\theta_1) - J(\theta_2)\| \\
&= \left\| \mathbb{E}_b \left[ \hat{J}_m(\theta_1) \right] - \mathbb{E}_b \left[ \hat{J}_m(\theta_2) \right] \right\| \quad (\text{from Lemma 1}) \\
&\leq \mathbb{E}_b \left[ \left\| \hat{J}_m(\theta_1) - \hat{J}_m(\theta_2) \right\| \right] \\
&\leq L \|\theta_1 - \theta_2\|, \quad (\text{from Lemma 3}).
\end{aligned}$$

This proves the first claim. For the second claim, notice that

$$\begin{aligned}
& \|\nabla J(\theta_1) - \nabla J(\theta_2)\| \\
&= \left\| \nabla \mathbb{E}_b \left[ \hat{J}_m(\theta_1) \right] - \nabla \mathbb{E}_b \left[ \hat{J}_m(\theta_2) \right] \right\| \quad (\text{from Lemma 1})
\end{aligned}$$

$$\begin{aligned}
&= \left\| \mathbb{E}_b \left[ \nabla \hat{J}_m(\theta_1) \right] - \mathbb{E}_b \left[ \nabla \hat{J}_m(\theta_2) \right] \right\| \quad (19) \\
&\leq \mathbb{E}_b \left[ \left\| \nabla \hat{J}_m(\theta_1) - \nabla \hat{J}_m(\theta_2) \right\| \right] \\
&\leq L \|\theta_1 - \theta_2\|, \quad (\text{from Lemma 3}).
\end{aligned}$$

In the above, the equality in (19) follows by an application of the dominated convergence theorem to interchange the differentiation and integration operations. For this application, we use the following facts:

- (i)  $\mathbb{E}_b \left[ \hat{J}_m(\theta) \right] < \infty$  holds for any  $\theta \in \mathbb{R}^d$  because the state and actions spaces are finite, the rewards are bounded,  $\pi_\theta(a|s) > 0$  and  $b(a|s) > 0$ ,  $\forall \theta \in \mathbb{R}^d, \forall a \in \mathcal{A}$ , and  $\forall s \in \mathcal{S}$  (from (A1) and (A2)), and  $\mathbb{P}(S_M \neq 0 | S_0, b) < 1$  from (A3);
- (ii)  $\|\nabla \hat{J}_m(\theta)\| \leq L$  from Lemma 3; and
- (iii)  $\mathbb{E}_b[L] < \infty$  since the state as well as action spaces are finite, and  $\mathbb{P}(S_M \neq 0 | S_0, b) < 1$  from (A3).  $\square$

Next, we recall a result from [14], which will be used to establish unbiasedness of the gradient estimate in (7).

**Lemma 5.**  $\nabla J_\mu(\theta) = \mathbb{E}_{v \in \mathbb{S}^{d-1}} \left[ \frac{d}{\mu} J(\theta + \mu v) v \right]$ .

*Proof.* See Lemma 2.1 in [14].  $\square$

**Lemma 6.**  $\nabla J_\mu(\theta) = \mathbb{E} \left[ \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) | \theta \right]$ .

*Proof.* We follow the technique from [34].

$$\begin{aligned}
& \mathbb{E} \left[ \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) | \theta \right] \\
&= \mathbb{E}_{b, v_{1:n}} \left[ \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right] = \mathbb{E}_b \left[ \mathbb{E}_{v_{1:n}} \left[ \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right] \right] \\
&= \mathbb{E}_b \left[ \frac{d}{n} \mathbb{E}_{v_{1:n}} \left[ \sum_{i=1}^n \frac{\hat{J}_m(\theta + \mu v_i) - \hat{J}_m(\theta - \mu v_i)}{2\mu} v_i \right] \right] \\
&= \frac{d}{2\mu} \mathbb{E}_b \left[ \mathbb{E}_v \left[ \left( \hat{J}_m(\theta + \mu v) - \hat{J}_m(\theta - \mu v) \right) v \right] \right] \\
&= \frac{d}{2\mu} \mathbb{E}_v \left[ \mathbb{E}_b \left[ \left( \hat{J}_m(\theta + \mu v) - \hat{J}_m(\theta - \mu v) \right) v \right] \right] \\
&= \frac{d}{2\mu} \mathbb{E}_v \left[ (J(\theta + \mu v) - J(\theta - \mu v)) v \right] \\
&\quad (\text{from Lemma 1}) \\
&= \frac{d}{2\mu} \mathbb{E}_v \left[ J(\theta + \mu v) v \right] + \mathbb{E}_v \left[ J(\theta - \mu v) (-v) \right] \\
&= \frac{d}{\mu} \mathbb{E}_v \left[ J(\theta + \mu v) v \right] \\
&\quad (\text{since } v \text{ has symmetric distribution}) \\
&= \nabla J_\mu(\theta), \quad (\text{from Lemma 5}).
\end{aligned}$$

$\square$



The claim below bounds the bias in the gradient estimate in (7), and can be inferred from [17]. For the sake of completeness, we provide the detailed proof.

**Lemma 7.**  $\|\nabla J_\mu(\theta) - \nabla J(\theta)\| \leq \frac{\mu d L}{2}$ .

*Proof.* Notice that

$$\begin{aligned}
& \|\nabla J_\mu(\theta) - \nabla J(\theta)\| \\
&= \left\| \mathbb{E}_v \left[ \frac{d}{\mu} J(\theta + \mu v) v \right] - \nabla J(\theta) \right\| \quad (\text{from Lemma 5}) \\
&= \left\| \mathbb{E}_v \left[ \frac{d}{\mu} J(\theta + \mu v) v \right] - \frac{d}{\mu} J(\theta) \mathbb{E}_v[v] \right. \\
&\quad \left. - \frac{d}{\mu} \langle \nabla J(\theta), \mu \mathbb{E}_v[v v^\top] \rangle \right\| \\
&\quad (\text{since } \mathbb{E}_{v \in \mathbb{S}^{d-1}}[v] = 0 \text{ and } \mathbb{E}_{v \in \mathbb{S}^{d-1}}[v v^\top] = \frac{\mathbf{1}_{d \times d}}{d}, \\
&\quad \text{cf. Theorem 2.7 in [12]}) \\
&= \frac{d}{\mu} \|\mathbb{E}_v[J(\theta + \mu v) v - J(\theta) v - \langle \nabla J(\theta), \mu v \rangle v]\| \\
&\leq \frac{d}{\mu} \mathbb{E}_v[\|J(\theta + \mu v) - J(\theta) - \langle \nabla J(\theta), \mu v \rangle\| \|v\|] \\
&\leq \frac{d}{\mu} \mathbb{E}_v[\|J(\theta + \mu v) - J(\theta) - \langle \nabla J(\theta), \mu v \rangle\|] \\
&\quad (\text{since } v \in \mathbb{S}^{d-1}, \|v\| = 1) \\
&\leq \frac{d}{\mu} \mathbb{E}_v \left[ \frac{L}{2} \mu^2 \|v\|^2 \right] \quad (\text{from Lemma 4}) \\
&\leq \frac{\mu d L}{2}, \quad (\text{since } v \in \mathbb{S}^{d-1}, \|v\| = 1).
\end{aligned}$$

□

**Lemma 8.**  $\mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right\|^2 \right] \leq \frac{4dL^2}{c_0 n}$ , for some absolute constant  $c_0 > 0$ .

*Proof.* Our proof proceeds in a similar manner to [34]. From Lemma 3, we obtain

$$\left\| \hat{J}_m(\theta + \mu v_1) - \hat{J}_m(\theta + \mu v_2) \right\| \leq L \mu \|v_1 - v_2\|. \quad (20)$$

Let  $f : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$  be an  $M$ -Lipschitz function. Then,

$$\mathbb{P} \left( \left| f - \int_{\mathbb{S}^{d-1}} f d\mathbb{P} \right| > \epsilon \right) \leq 4e^{-\frac{c_0 \epsilon^2 d}{M^2}}, \quad (21)$$

where  $c_0 > 0$  is an absolute constant [see 27, Appendix V.2].

From (20) and (21), we obtain

$$\mathbb{P} \left( \left| \hat{J}_m(\theta + \mu v) - \mathbb{E}_{v \in \mathbb{S}^{d-1}} [\hat{J}_m(\theta + \mu v)] \right| > \epsilon \right)$$

$$\leq 4e^{-\frac{c_0 \epsilon^2 d}{\mu^2 L^2}}. \quad (22)$$

Using (22), we obtain

$$\begin{aligned}
& \mathbb{E}_{v \in \mathbb{S}^{d-1}} \left[ \left( \hat{J}_m(\theta + \mu v) - \mathbb{E}_{v \in \mathbb{S}^{d-1}} [\hat{J}_m(\theta + \mu v)] \right)^2 \right] \\
&= \int_0^\infty \mathbb{P} \left( \left| \hat{J}_m(\theta + \mu v) - \mathbb{E}_{v \in \mathbb{S}^{d-1}} [\hat{J}_m(\theta + \mu v)] \right| > \sqrt{\epsilon} \right) d\epsilon \\
&\leq \int_0^\infty 4e^{-\frac{c_0 \epsilon^2 d}{\mu^2 L^2}} d\epsilon \leq \frac{4L^2 \mu^2}{c_0 d}. \quad (23)
\end{aligned}$$

Now,

$$\begin{aligned}
& \mathbb{E}_{v_{1:n}} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right\|^2 \right] \\
&= \mathbb{E}_{v_{1:n}} \left[ \left\| \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}_m(\theta + \mu v_i) - \hat{J}_m(\theta - \mu v_i)}{2\mu} v_i \right\|^2 \right] \quad (\text{from (7)}) \\
&\leq \frac{d^2}{n^2} \sum_{i=1}^n \mathbb{E}_v \left[ \left\| \frac{\hat{J}_m(\theta + \mu v) - \hat{J}_m(\theta - \mu v)}{2\mu} v \right\|^2 \right] \\
&\leq \frac{d^2}{4\mu^2 n} \mathbb{E}_v \left[ \left( \hat{J}_m(\theta + \mu v) - \hat{J}_m(\theta - \mu v) \right)^2 \|v\|^2 \right] \\
&\leq \frac{d^2}{4\mu^2 n} \mathbb{E}_v \left[ \left( \hat{J}_m(\theta + \mu v) - \hat{J}_m(\theta - \mu v) \right)^2 \right] \quad (\text{since } v \in \mathbb{S}^{d-1}, \|v\| = 1) \\
&\leq \frac{d^2}{4\mu^2 n} \left( \mathbb{E}_v \left[ \left( \left( \hat{J}_m(\theta + \mu v) - \mathbb{E}_v [\hat{J}_m(\theta + \mu v)] \right) \right. \right. \right. \\
&\quad \left. \left. \left. - \left( \hat{J}_m(\theta - \mu v) - \mathbb{E}_v [\hat{J}_m(\theta - \mu v)] \right) \right)^2 \right] \right) \\
&\leq \frac{d^2}{2\mu^2 n} \left( \mathbb{E}_v \left[ \left( \hat{J}_m(\theta + \mu v) - \mathbb{E}_v [\hat{J}_m(\theta + \mu v)] \right)^2 \right] \right. \\
&\quad \left. + \mathbb{E}_v \left[ \left( \hat{J}_m(\theta - \mu v) - \mathbb{E}_v [\hat{J}_m(\theta - \mu v)] \right)^2 \right] \right) \quad (\text{since } (a - b)^2 \leq 2a^2 + 2b^2) \\
&\leq \frac{d^2}{\mu^2 n} \mathbb{E}_v \left[ \left( \hat{J}_m(\theta + \mu v) - \mathbb{E}_v [\hat{J}_m(\theta + \mu v)] \right)^2 \right] \quad (\text{since } v \text{ has symmetric distribution}) \\
&\leq \frac{d^2}{\mu^2 n} \left( \frac{4L^2 \mu^2}{c_0 d} \right) \quad (\text{from (23)}) \\
&\leq \frac{4dL^2}{c_0 n}. \quad (24)
\end{aligned}$$

Using (24) we obtain,

$$\mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right\|^2 \right]$$

$$= \mathbb{E}_\theta \left[ \mathbb{E}_b \left[ \mathbb{E}_{v_{1:n}} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right\|^2 \right] \right] \right] \leq \frac{4dL^2}{c_0 n}.$$

□

**Lemma 9.** For some absolute constant  $c_0 > 0$ ,  $\mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) - \mathbb{E} \left[ \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) \mid \theta \right] \right\|^2 \right] \leq \frac{4dL^2}{c_0 n}.$

*Proof.* Notice that

$$\begin{aligned} & \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) - \mathbb{E} \left[ \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) \mid \theta \right] \right\|^2 \right] \\ &= \sum_{i=1}^d \mathbb{E} \left[ \left( \widehat{\nabla}_{n,\mu}^i \hat{J}_m(\theta) - \mathbb{E} \left[ \widehat{\nabla}_{n,\mu}^i \hat{J}_m(\theta) \mid \theta \right] \right)^2 \right] \\ &\leq \sum_{i=1}^d \mathbb{E} \left[ \left( \widehat{\nabla}_{n,\mu}^i \hat{J}_m(\theta) \right)^2 \right] = \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}_m(\theta) \right\|^2 \right] \\ &\quad (\text{since } \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] \leq \mathbb{E}[X^2]) \\ &\leq \frac{4dL^2}{c_0 n}, \quad (\text{from Lemma 8}). \end{aligned}$$

□

The claim below is well-known in the context of projections on to convex sets. We have provided the proof for the sake of completeness.

**Lemma 10.** The projection operator  $\mathcal{P}_\Theta$  defined in (14) satisfies

$$\begin{aligned} (i) \quad & \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha)\| \leq \|f(\theta)\|, \\ (ii) \quad & \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha) - \mathcal{P}_\Theta(\theta, g(\theta), \alpha)\| \\ & \leq \|f(\theta) - g(\theta)\|, \text{ and} \\ (iii) \quad & \langle f(\theta), \mathcal{P}_\Theta(\theta, f(\theta), \alpha) \rangle \geq \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha)\|^2. \end{aligned}$$

*Proof.*

$$\begin{aligned} (i) \quad & \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha)\| \\ &= \frac{1}{\alpha} \|\Pi_\Theta(\theta + \alpha f(\theta)) - \theta\| \quad (\text{from (14)}) \\ &\leq \frac{1}{\alpha} \|\theta + \alpha f(\theta) - \theta\| = \|f(\theta)\|, \\ &\quad (\text{since } \|\Pi_\Theta(x) - y\| \leq \|x - y\|, \forall y \in \Theta). \\ (ii) \quad & \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha) - \mathcal{P}_\Theta(\theta, g(\theta), \alpha)\| \\ &= \left\| \frac{1}{\alpha} [\Pi_\Theta(\theta + \alpha f(\theta)) - \theta] \right. \\ &\quad \left. - \frac{1}{\alpha} [\Pi_\Theta(\theta + \alpha g(\theta)) - \theta] \right\| \quad (\text{from (14)}) \\ &= \frac{1}{\alpha} \|\Pi_\Theta(\theta + \alpha f(\theta)) - \Pi_\Theta(\theta + \alpha g(\theta))\| \end{aligned}$$

$$\begin{aligned} & \leq \frac{1}{\alpha} \|\theta + \alpha f(\theta) - \theta + \alpha g(\theta)\| \\ & \quad (\text{since } \|\Pi_\Theta(x) - \Pi_\Theta(y)\| \leq \|x - y\|, \forall x, y) \\ & \leq \|f(\theta) - g(\theta)\|. \end{aligned}$$

$$\begin{aligned} (iii) \quad & \langle f(\theta), \mathcal{P}_\Theta(\theta, f(\theta), \alpha) \rangle - \|\mathcal{P}_\Theta(\theta, f(\theta), \alpha)\|^2 \\ &= \langle f(\theta), \mathcal{P}_\Theta(\theta, f(\theta), \alpha) \rangle \\ & \quad - \langle \mathcal{P}_\Theta(\theta, f(\theta), \alpha), \mathcal{P}_\Theta(\theta, f(\theta), \alpha) \rangle \\ &= \langle f(\theta) - \mathcal{P}_\Theta(\theta, f(\theta), \alpha), \mathcal{P}_\Theta(\theta, f(\theta), \alpha) \rangle \\ &= \left\langle f(\theta) - \frac{1}{\alpha} [\Pi_\Theta(\theta + \alpha f(\theta)) - \theta], \right. \\ & \quad \left. \frac{1}{\alpha} [\Pi_\Theta(\theta + \alpha f(\theta)) - \theta] \right\rangle \\ &= -\frac{1}{\alpha^2} \langle \Pi_\Theta(\theta + \alpha f(\theta)) - (\theta + \alpha f(\theta)), \\ & \quad \Pi_\Theta(\theta + \alpha f(\theta)) - \theta \rangle, \\ & \quad (\text{since } \langle \Pi_\Theta(x) - x, \Pi_\Theta(x) - y \rangle \leq 0, \forall y \in \Theta). \end{aligned}$$

□

**Proof of Theorem 1.** Using the fundamental theorem of calculus, we obtain

$$\begin{aligned} & J(\theta_k) - J(\theta_{k+1}) \\ &= \langle \nabla J(\theta_k), \theta_k - \theta_{k+1} \rangle \\ & \quad + \int_0^1 \langle \nabla J(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})) - \nabla J(\theta_k), \\ & \quad \theta_k - \theta_{k+1} \rangle d\tau \\ &\leq \langle \nabla J(\theta_k), \theta_k - \theta_{k+1} \rangle \\ & \quad + \int_0^1 \|\nabla J(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})) - \nabla J(\theta_k)\| \\ & \quad \|\theta_k - \theta_{k+1}\| d\tau \quad (\text{Cauchy-Schwarz inequality}) \\ &\leq \langle \nabla J(\theta_k), \theta_k - \theta_{k+1} \rangle \\ & \quad + L \|\theta_k - \theta_{k+1}\|^2 \int_0^1 (1 - \tau) d\tau \quad (\text{from Lemma 4}) \\ &\leq \langle \nabla J(\theta_k), \theta_k - \theta_{k+1} \rangle + \frac{L}{2} \|\theta_k - \theta_{k+1}\|^2 \\ &\leq \alpha_k \left\langle \nabla J(\theta_k), -\mathcal{P}_\Theta(\theta_k, \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k), \alpha_k) \right\rangle \\ & \quad + \frac{L\alpha_k^2}{2} \left\| \mathcal{P}_\Theta(\theta_k, \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k), \alpha_k) \right\|^2 \\ & \quad (\text{from (13)}) \\ &\leq \alpha_k \langle \nabla J(\theta_k), \mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k) \\ & \quad - \mathcal{P}_\Theta(\theta_k, \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k), \alpha_k) \rangle \\ & \quad - \alpha_k \langle \nabla J(\theta_k), \mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k) \rangle \\ & \quad + \frac{L\alpha_k^2}{2} \left\| \mathcal{P}_\Theta(\theta_k, \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k), \alpha_k) \right\|^2 \end{aligned}$$

$$\begin{aligned}
&\leq \alpha_k \|\nabla J(\theta_k)\| \left\| \nabla J(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad - \alpha_k \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \\
&\quad + \frac{L\alpha_k^2}{2} \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2 \quad (\text{from Lemma 10}) \\
&\leq L\alpha_k \left\| \nabla J(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad - \alpha_k \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \\
&\quad + \frac{L\alpha_k^2}{2} \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2 \quad (\text{from Lemma 4}) \\
&\leq L\alpha_k \|\nabla J(\theta_k) - \nabla J_{\mu_k}(\theta_k)\| \\
&\quad + L\alpha_k \left\| \nabla J_{\mu_k}(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad - \alpha_k \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \\
&\quad + \frac{L\alpha_k^2}{2} \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2 \\
&\leq \frac{dL^2}{2} \alpha_k \mu_k + L\alpha_k \left\| \nabla J_{\mu_k}(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad - \alpha_k \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \\
&\quad + \frac{L\alpha_k^2}{2} \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2, \quad (25)
\end{aligned}$$

where the final inequality follows from Lemma 7. Summing up (25) for  $k = 0, \dots, N-1$ , we obtain

$$\begin{aligned}
&\sum_{k=0}^{N-1} \alpha_k \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \\
&\leq (J(\theta_N) - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k \\
&\quad + L \sum_{k=0}^{N-1} \alpha_k \left\| \nabla J_{\mu_k}(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad + \frac{L}{2} \sum_{k=0}^{N-1} \alpha_k^2 \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2 \\
&\leq (J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k \\
&\quad + L \sum_{k=0}^{N-1} \alpha_k \left\| \nabla J_{\mu_k}(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \\
&\quad + \frac{L}{2} \sum_{k=0}^{N-1} \alpha_k^2 \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2. \quad (26)
\end{aligned}$$

Taking expectations on both sides of (26), we obtain

$$\sum_{k=0}^{N-1} \alpha_k \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \right]$$

$$\begin{aligned}
&\leq (J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k \\
&\quad + L \sum_{k=0}^{N-1} \alpha_k \mathbb{E} \left[ \left\| \nabla J_{\mu_k}(\theta_k) - \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\| \right] \\
&\quad + \frac{L}{2} \sum_{k=0}^{N-1} \alpha_k^2 \mathbb{E} \left[ \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right\|^2 \right] \\
&\leq (J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k + \frac{2dL^3}{c_0} \sum_{k=0}^{N-1} \frac{\alpha_k^2}{n_k} \\
&\quad + L \sum_{k=0}^{N-1} \alpha_k \mathbb{E} \left[ \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right. \right. \\
&\quad \quad \left. \left. - \mathbb{E} \left[ \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \mid \theta_k \right] \right\| \right] \quad (\text{from Lemmas 6 and 8}) \\
&\leq (J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k + \frac{2dL^3}{c_0} \sum_{k=0}^{N-1} \frac{\alpha_k^2}{n_k} \\
&\quad + L \sum_{k=0}^{N-1} \alpha_k \left( \mathbb{E} \left[ \left\| \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \right. \right. \right. \\
&\quad \quad \left. \left. - \mathbb{E} \left[ \widehat{\nabla}_{n_k, \mu_k} \hat{J}_{m_k}(\theta_k) \mid \theta_k \right] \right\|^2 \right] \right)^{\frac{1}{2}} \\
&\leq (J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k + \frac{2dL^3}{c_0} \sum_{k=0}^{N-1} \frac{\alpha_k^2}{n_k} \\
&\quad + \frac{2\sqrt{d}L^2}{\sqrt{c_0}} \sum_{k=0}^{N-1} \frac{\alpha_k}{\sqrt{n_k}}, \quad (\text{from Lemma 9}).
\end{aligned}$$

Since  $\mathbb{P}(R = k) = \frac{\alpha_k}{\sum_{k=0}^{N-1} \alpha_k}$ , we obtain

$$\begin{aligned}
&\mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha_R)\|^2 \right] \\
&= \frac{\sum_{k=0}^{N-1} \alpha_k \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha_k)\|^2 \right]}{\sum_{k=0}^{N-1} \alpha_k} \\
&\leq \frac{(J^* - J(\theta_0)) + \frac{dL^2}{2} \sum_{k=0}^{N-1} \alpha_k \mu_k}{\sum_{k=0}^{N-1} \alpha_k} \\
&\quad + \frac{\frac{2dL^3}{c_0} \sum_{k=0}^{N-1} \frac{\alpha_k^2}{n_k} + \frac{2\sqrt{d}L^2}{\sqrt{c_0}} \sum_{k=0}^{N-1} \frac{\alpha_k}{\sqrt{n_k}}}{\sum_{k=0}^{N-1} \alpha_k}.
\end{aligned}$$

□

**Proof of Corollary 1.** In (15), we substitute  $\alpha_k = \frac{c_1}{\sqrt{N}}$ ,  $\mu_k = \frac{c_2}{\sqrt{N}}$ , and  $n_k = c_3 N$ ,  $\forall k$ , for some absolute constants  $c_1, c_2, c_3 > 0$ , to obtain

$$\mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_R, \nabla J(\theta_R), \alpha_R)\|^2 \right]$$

$$\leq \frac{(J^* - J(\theta_0)) + c' L^2(d + \sqrt{d})}{\sqrt{N}} + \frac{c'' d L^3}{N \sqrt{N}},$$

for some constants  $c', c'' > 0$ .  $\square$

## 5.2 Proofs for OffP-SF-SVRG

**Lemma 11.**  $\mathbb{E}_b [\hat{J}^j(\theta)] = J(\theta), \forall j$ .

*Proof.* Notice that

$$\begin{aligned} \mathbb{E}_b [\hat{J}^j(\theta)] &= \mathbb{E}_{\substack{[1,m] \sim b \\ j \in [1,m]}} \left[ \sum_{t=0}^{T^j-1} \gamma^t R_{t+1}^j \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^j | S_i^j)}{b(A_i^j | S_i^j)} \right) \right] \\ &= \mathbb{E}_{\substack{[1,m] \sim \pi_\theta \\ j \in [1,m]}} \left[ \sum_{t=0}^{T^j-1} \gamma^t R_{t+1}^j \right] = J(\theta). \end{aligned}$$

$\square$

Note that Lemmas 4, 6 hold for OffP-SF-SVRG, and the proof follows by using Lemma 11 in place of Lemma 1.

**Lemma 12.**

$$\begin{aligned} \mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_1) - \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_2) \right\|^2 \right] \\ \leq \frac{d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_1 - \theta_2\|^2 \right]. \end{aligned}$$

*Proof.* Notice that

$$\begin{aligned} &\mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_1) - \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_2) \right\|^2 \right] \\ &\leq \frac{d^2}{4\mu^2 n^2} \sum_{i=1}^n \mathbb{E} \left[ \left\| \hat{J}^j(\theta_1 + \mu v_i) - \hat{J}^j(\theta_2 + \mu v_i) \right. \right. \\ &\quad \left. \left. + \hat{J}^j(\theta_2 - \mu v_i) - \hat{J}^j(\theta_1 - \mu v_i) \right\|^2 \|v_i\|^2 \right] \\ &\leq \frac{d^2}{4\mu^2 n} \mathbb{E} \left[ \left\| \hat{J}^j(\theta_1 + \mu v) - \hat{J}^j(\theta_2 + \mu v) \right. \right. \\ &\quad \left. \left. + \hat{J}^j(\theta_2 - \mu v) - \hat{J}^j(\theta_1 - \mu v) \right\|^2 \right] \quad (\text{since } \|v\| = 1) \\ &\leq \frac{d^2}{2\mu^2 n} \left[ \mathbb{E} \left[ \left\| \hat{J}^j(\theta_1 + \mu v) - \hat{J}^j(\theta_2 + \mu v) \right\|^2 \right] \right. \\ &\quad \left. + \mathbb{E} \left[ \left\| \hat{J}^j(\theta_2 - \mu v) - \hat{J}^j(\theta_1 - \mu v) \right\|^2 \right] \right] \\ &\leq \frac{d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_1 - \theta_2\|^2 \right], \quad (\text{by Lemma 3 with } m=1). \end{aligned}$$

$\square$

**Lemma 13.**  $\hat{\nabla}_{n,\mu} \hat{J}_m(\theta) = \mathbb{E}_{j \in [1,m]} [\hat{\nabla}_{n,\mu} \hat{J}^j(\theta)]$ .

*Proof.* Notice that

$$\begin{aligned} \hat{\nabla}_{n,\mu} \hat{J}_m(\theta) &= \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}_m(\theta + \mu v_i) - \hat{J}_m(\theta - \mu v_i)}{2\mu} v_i \\ &= \frac{1}{m} \sum_{j=1}^m \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}^j(\theta + \mu v_i) - \hat{J}^j(\theta - \mu v_i)}{2\mu} v_i \\ &\quad (\text{from (4) and (9)}) \\ &= \mathbb{E}_{j \in [1,m]} \left[ \frac{d}{n} \sum_{i=1}^n \frac{\hat{J}^j(\theta + \mu v_i) - \hat{J}^j(\theta - \mu v_i)}{2\mu} v_i \right] \\ &= \mathbb{E}_{j \in [1,m]} [\hat{\nabla}_{n,\mu} \hat{J}^j(\theta)], \quad (\text{from (10)}). \end{aligned}$$

$\square$

**Lemma 14.**  $\mathbb{E}[g_k^s] = \mathbb{E}[\hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s)]$ .

*Proof.* Notice that

$$\begin{aligned} &\mathbb{E}[g_k^s] \\ &= \mathbb{E} \left[ \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) + \hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s) \right] \\ &\quad (\text{from (12)}) \\ &= \mathbb{E} \left[ \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right] \\ &\quad + \mathbb{E} \left[ \mathbb{E}_{j \in [1,m]} [\hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s) - \hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s)] \right] \\ &= \mathbb{E} \left[ \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right] \quad (\text{from Lemma 13}). \end{aligned}$$

$\square$

**Lemma 15.** For some absolute constant  $c_0 > 0$ ,

$$\mathbb{E}[\|g_k^s\|^2] \leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E}[\|\theta_k^s - \tilde{\theta}^s\|^2] + \frac{8dL^2}{c_0 n}.$$

*Proof.* Notice that

$$\begin{aligned} &\mathbb{E}[\|g_k^s\|^2] \\ &= \mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) + \hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s) \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \hat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right\|^2 \right] \\ &\quad + 2\mathbb{E} \left[ \left\| \hat{\nabla}_{n,\mu} \hat{J}_m(\tilde{\theta}^s) \right\|^2 \right] \\ &\leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E}[\|\theta_k^s - \tilde{\theta}^s\|^2] + \frac{8dL^2}{c_0 n}, \\ &\quad (\text{from Lemmas 8 and 12}). \end{aligned}$$

$\square$

**Lemma 16.** For some absolute constant  $c_0 > 0$ ,

$$\begin{aligned} & \mathbb{E} \left[ \|g_k^s - \nabla J(\theta_k^s)\|^2 \right] \\ & \leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_k^s - \tilde{\theta}^s\|^2 \right] + \mu^2 d^2 L^2 + \frac{16dL^2}{c_0 n}. \end{aligned}$$

*Proof.* Notice that

$$\begin{aligned} & \mathbb{E} \left[ \|g_k^s - \nabla J(\theta_k^s)\|^2 \right] \\ & = \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right. \right. \\ & \quad \left. \left. + \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) - \nabla J(\theta_k^s) \right\|^2 \right] \\ & \leq 2\mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right\|^2 \right] \\ & \quad - \mathbb{E}_{j \in [1,m]} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right\|^2 \right] \\ & \quad + 2\mathbb{E} \left[ \left\| \mathbb{E}_{j \in [1,m]} \left[ \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right] - \nabla J(\theta_k^s) \right\|^2 \right] \\ & \quad \quad \quad \text{(from Lemma 13)} \\ & \leq 2\mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right\|^2 \right] \\ & \quad + 2\mathbb{E} \left[ \left\| \frac{1}{m} \sum_{j=1}^m \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \nabla J(\theta_k^s) \right\|^2 \right] \\ & \quad \quad \quad \text{(since } \mathbb{E}[\|X - \mathbb{E}[X|Y]\|^2] \leq \mathbb{E}[\|X\|^2]) \\ & \leq 2\mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \widehat{\nabla}_{n,\mu} \hat{J}^j(\tilde{\theta}^s) \right\|^2 \right] \\ & \quad + \frac{2}{m} \sum_{j=1}^m \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \nabla J(\theta_k^s) \right\|^2 \right] \\ & \quad \quad \quad \text{(since } \|\sum_{i=1}^m X_i\|^2 \leq m \sum_{i=1}^m \|X_i\|^2) \\ & \leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_k^s - \tilde{\theta}^s\|^2 \right] \\ & \quad + \frac{4}{m} \sum_{j=1}^m \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) - \mathbb{E} \left[ \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) | \theta_k^s \right] \right\|^2 \right] \\ & \quad + \frac{4}{m} \sum_{j=1}^m \mathbb{E} \left[ \|\nabla J_\mu(\theta_k^s) - \nabla J(\theta_k^s)\|^2 \right] \\ & \quad \quad \quad \text{(from Lemma 12 and Lemma 6 with } m=1) \\ & \leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_k^s - \tilde{\theta}^s\|^2 \right] \\ & \quad + \frac{4}{m} \sum_{j=1}^m \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right\|^2 \right] \end{aligned}$$

$$\begin{aligned} & + \frac{4}{m} \sum_{j=1}^m \mathbb{E} \left[ \|\nabla J_\mu(\theta_k^s) - \nabla J(\theta_k^s)\|^2 \right] \\ & \quad \quad \quad \text{(since } \mathbb{E}[\|X - E[X|Y]\|^2] \leq \mathbb{E}[\|X\|^2]) \\ & \leq \frac{2d^2 L^2}{\mu^2 n} \mathbb{E} \left[ \|\theta_k^s - \tilde{\theta}^s\|^2 \right] + \mu^2 d^2 L^2 + \frac{16dL^2}{c_0 n}, \\ & \quad \quad \quad \text{(from Lemma 7 and Lemma 8 with } m=1). \end{aligned}$$

□

**Proof of Theorem 3.** By using a completely parallel argument to the initial passage in the proof of Theorem 1 leading up to (25), we obtain

$$\begin{aligned} & J(\theta_k^s) - J(\theta_{k+1}^s) \\ & \leq \alpha \langle \nabla J(\theta_k^s), \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) - \mathcal{P}_\Theta(\theta_k^s, g_k^s, \alpha) \rangle \\ & \quad - \alpha \langle \nabla J(\theta_k^s), \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \rangle \\ & \quad + \frac{L\alpha^2}{2} \|\mathcal{P}_\Theta(\theta_k^s, g_k^s, \alpha)\|^2 \\ & \leq \frac{\alpha}{2} \|\nabla J(\theta_k^s)\|^2 + \frac{\alpha}{2} \|\nabla J(\theta_k^s) - g_k^s\|^2 \\ & \quad - \alpha \|\mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha)\|^2 \\ & \quad + \frac{L\alpha^2}{2} \|g_k^s\|^2 \quad \quad \quad \text{(from Lemma 10)} \\ & \leq \frac{3\alpha}{2} \|\nabla J(\theta_k^s) - g_k^s\|^2 + \left( \frac{L\alpha^2}{2} + \alpha \right) \|g_k^s\|^2 \\ & \quad - \alpha \|\mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha)\|^2. \end{aligned} \quad (27)$$

Taking expectations on both sides of (27), we obtain

$$\begin{aligned} & \mathbb{E} [J(\theta_{k+1}^s)] \\ & \geq \mathbb{E} [J(\theta_k^s)] + \alpha \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha)\|^2 \right] \\ & \quad - \frac{3\alpha}{2} \mathbb{E} \left[ \|\nabla J(\theta_k^s) - g_k^s\|^2 \right] \\ & \quad - \left( \frac{L\alpha^2}{2} + \alpha \right) \mathbb{E} \left[ \|g_k^s\|^2 \right] \\ & \geq \mathbb{E} [J(\theta_k^s)] + \alpha \mathbb{E} \left[ \|\mathcal{P}_\Theta(\theta_k, \nabla J(\theta_k), \alpha)\|^2 \right] \\ & \quad - \left( \frac{5\alpha d^2 L^2}{\mu^2 n} + \frac{\alpha^2 d^2 L^3}{\mu^2 n} \right) \mathbb{E} \left[ \|\theta_k^s - \tilde{\theta}^s\|^2 \right] \\ & \quad - \frac{3\alpha \mu^2 d^2 L^2}{2} - \left( \frac{\alpha^2 L}{2} + 4\alpha \right) \frac{8dL^2}{c_0 n}, \end{aligned} \quad (28)$$

where the final inequality follows from Lemmas 15–16.

Now,

$$\begin{aligned} & \mathbb{E} \left[ \|\theta_{k+1}^s - \tilde{\theta}^s\|^2 \right] = \mathbb{E} \left[ \|\Pi_\Theta(\theta_k^s + \alpha g_k^s) - \tilde{\theta}^s\|^2 \right] \\ & \leq \mathbb{E} \left[ \|\theta_k^s + \alpha g_k^s - \tilde{\theta}^s\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 + \alpha^2 \|g_k^s\|^2 + 2\alpha \left\langle \theta_k^s - \tilde{\theta}^s, g_k^s \right\rangle \right] \\
&\leq \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] + 2\alpha \mathbb{E} \left[ \left\langle \theta_k^s - \tilde{\theta}^s, \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right\rangle \right] \\
&\quad + \alpha^2 \mathbb{E} \left[ \|g_k^s\|^2 \right] \quad (\text{from Lemma 14}) \\
&\leq \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] + \alpha^2 \mathbb{E} \left[ \|g_k^s\|^2 \right] \\
&\quad + \frac{\alpha}{2\alpha m} \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] + 2\alpha^2 m \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right\|^2 \right] \\
&\quad (\text{since } \langle a, b \rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\|b\|^2\beta}{2}, \beta > 0) \\
&\leq \left( 1 + \frac{1}{2m} \right) \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] + \alpha^2 \mathbb{E} \left[ \|g_k^s\|^2 \right] \\
&\quad + 2\alpha^2 m \mathbb{E} \left[ \left\| \widehat{\nabla}_{n,\mu} \hat{J}^j(\theta_k^s) \right\|^2 \right] \\
&\leq \left( 1 + \frac{1}{2m} + \frac{2\alpha^2 d^2 L^2}{\mu^2 n} \right) \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] \\
&\quad + \frac{8\alpha^2 d L^2 (1+m)}{c_0 n}, \quad (29)
\end{aligned}$$

where the final inequality follows from Lemma 15 and the result in Lemma 8 with  $m = 1$ . Let

$$R_k^s = \mathbb{E} [J(\theta_k^s)] - b_k \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right]. \quad (30)$$

Now,

$$\begin{aligned}
&R_{k+1}^s \\
&\geq \mathbb{E} [J(\theta_k^s)] + \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\
&\quad - \left( \frac{5\alpha d^2 L^2}{\mu^2 n} + \frac{\alpha^2 d^2 L^3}{\mu^2 n} \right) \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] \\
&\quad - \frac{3\alpha \mu^2 d^2 L^2}{2} - \left( \frac{L\alpha^2}{2} + 4\alpha \right) \frac{8dL^2}{c_0 n} \\
&\quad - b_{k+1} \mathbb{E} \left[ \left\| \theta_{k+1}^s - \tilde{\theta}^s \right\|^2 \right] \quad (\text{from (28)}) \\
&\geq \mathbb{E} [J(\theta_k^s)] + \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\
&\quad - \left( \frac{5\alpha d^2 L^2}{\mu^2 n} + \frac{\alpha^2 d^2 L^3}{\mu^2 n} + b_{k+1} \left( 1 + \frac{1}{2m} \right) \right. \\
&\quad \left. + \frac{2\alpha^2 d^2 L^2}{\mu^2 n} \right) \mathbb{E} \left[ \left\| \theta_k^s - \tilde{\theta}^s \right\|^2 \right] \\
&\quad - \left( \frac{L\alpha^2}{2} + 4\alpha + b_{k+1} \alpha^2 (1+m) \right) \frac{8dL^2}{c_0 n} \\
&\quad - \frac{3\alpha \mu^2 d^2 L^2}{2}, \quad (\text{from (29)}). \\
&\quad (31)
\end{aligned}$$

Let

$$b_k = \begin{cases} x + b_{k+1} (1+y) & \text{for } k \in \{0, m-1\} \\ 0 & \text{for } k \geq m \end{cases} \quad (32)$$

where

$$x = \frac{5\alpha d^2 L^2}{\mu^2 n} + \frac{\alpha^2 d^2 L^3}{\mu^2 n}, y = \frac{1}{2m} + \frac{2\alpha^2 d^2 L^2}{\mu^2 n}.$$

By solving the recursion (32), we obtain

$$b_k = \frac{x}{y} \left( (1+y)^{m-k} - 1 \right). \quad (33)$$

It is easy to see that

$$b_k \leq b_0 \leq \frac{x}{y} (1+y)^m, \quad \forall k. \quad (34)$$

From (31), (32) and (34) we obtain

$$\begin{aligned}
&\alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\
&\leq R_{k+1}^s - R_k^s + \frac{3\alpha \mu^2 d^2 L^2}{2} \\
&\quad + \left( \frac{L\alpha^2}{2} + 4\alpha + b_0 \alpha^2 (1+m) \right) \frac{8dL^2}{c_0 n}. \quad (35)
\end{aligned}$$

Now, from (30) we obtain

$$\begin{aligned}
R_m^s &= \mathbb{E} [J(\theta_m^s)] - b_m \mathbb{E} \left[ \left\| \theta_m^s - \tilde{\theta}^s \right\|^2 \right] \\
&= \mathbb{E} [J(\theta_m^s)] = \mathbb{E} [J(\tilde{\theta}^{s+1})], \quad (\text{from (32)}) \\
R_0^s &= \mathbb{E} [J(\theta_0^s)] - b_0 \mathbb{E} \left[ \left\| \theta_0^s - \tilde{\theta}^s \right\|^2 \right] \\
&= \mathbb{E} [J(\theta_0^s)] = \mathbb{E} [J(\tilde{\theta}^s)], \quad (\text{since } \theta_0^s = \tilde{\theta}^s), \\
&\quad (36)
\end{aligned}$$

Summing up (35) from  $k = 0, \dots, m-1$ , we obtain

$$\begin{aligned}
&\sum_{k=0}^{m-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\
&\leq R_m^s - R_0^s + \frac{3\alpha \mu^2 d^2 L^2 m}{2} \\
&\quad + \left( \frac{L\alpha^2}{2} + 4\alpha + b_0 \alpha^2 (1+m) \right) \frac{8dL^2 m}{c_0 n} \\
&\leq \mathbb{E} [J(\tilde{\theta}^{s+1})] - \mathbb{E} [J(\tilde{\theta}^s)] + \frac{3\alpha \mu^2 d^2 L^2 m}{2} \\
&\quad + \left( \frac{L\alpha^2}{2} + 4\alpha + b_0 \alpha^2 (1+m) \right) \frac{8dL^2 m}{c_0 n}, \\
&\quad (\text{from (36)}).
\end{aligned}$$

Summing the RHS above from  $s = 0, \dots, S-1$ , we obtain

$$\begin{aligned} & \sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\ & \leq \mathbb{E} \left[ J(\tilde{\theta}^S) \right] - \mathbb{E} \left[ J(\tilde{\theta}^0) \right] + \frac{3\alpha\mu^2 d^2 L^2 S m}{2} \\ & \quad + \left( \frac{L\alpha^2}{2} + 4\alpha + b_0\alpha^2(1+m) \right) \frac{8dL^2 S m}{c_0 n}. \quad (37) \end{aligned}$$

From the definition of  $\alpha, \mu$ , and  $n$  in the theorem statement, we have

$$\begin{aligned} y &= \frac{1}{2m} + \frac{1}{8m} \leq \frac{1}{m}, \frac{x}{y} = 2dL + \frac{L}{10}, \text{ and} \\ b_0 &\leq \left( 2dL + \frac{L}{10} \right) (1 + 1/m)^m \leq \left( 2dL + \frac{L}{10} \right) e. \end{aligned}$$

Using the bound on  $b_0$  in (37), we have

$$\begin{aligned} & \sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right] \\ & \leq \mathbb{E} \left[ J(\tilde{\theta}^S) \right] - \mathbb{E} \left[ J(\tilde{\theta}^0) \right] + \frac{3dL}{8} \\ & \quad + \left( 1 + \frac{1}{32d} + \frac{(20d+1)e(1+m)}{160d} \right) \frac{8L}{c_0 m}. \quad (38) \end{aligned}$$

Since  $\mathbb{P}(Q = s, R = k) = \frac{\alpha}{\sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha}$ , we obtain

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_R^Q, \nabla J(\theta_R^Q), \alpha) \right\|^2 \right] \\ &= \frac{\sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_\Theta(\theta_k^s, \nabla J(\theta_k^s), \alpha) \right\|^2 \right]}{\sum_{s=0}^{S-1} \sum_{k=0}^{m-1} \alpha} \\ &\leq \frac{4dL (J^* - J(\theta_0^0))}{Sm} + \frac{L^2 (15c_0 d^2 + 40ed + 2e)}{10c_0 Sm} \\ &\quad + \frac{20dL^2(8+e) + L^2(5+e)}{5c_0 Sm^2}. \quad (39) \end{aligned}$$

□

### 5.3 Proofs for REINFORCE (off-policy variant)

**Proof of Theorem 2.** The REINFORCE (off-policy variant) algorithm solves the following update iterate

$$\theta_{k+1} = \Pi_\Theta(\theta_k + \alpha \hat{\nabla} J(\theta)), \quad (40)$$

where  $\hat{\nabla} J(\theta)$  is defined as below:

$$\hat{\nabla} J(\theta) = \frac{1}{m} \sum_{n=1}^m \left[ \sum_{t=0}^{T^n-1} \nabla \log \pi_\theta(A_t^n | S_t^n) \right]$$

$$\left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \left( \sum_{i=t}^{T^n-1} \gamma^i R_{i+1}^n \right) \quad (41)$$

The policy gradient estimate  $\hat{\nabla} J(\theta)$  is an unbiased estimate of  $\nabla J(\theta)$ , where

$$\begin{aligned} & \nabla J(\theta) \\ &= \mathbb{E}_{\pi_\theta} \left[ \left( \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t | S_t) \right) \left( \sum_{t=0}^{T-1} \gamma^t R_{t+1} \right) \right] \\ &= \mathbb{E}_b \left[ \left( \prod_{t=0}^{T-1} \frac{\pi_\theta(A_t | S_t)}{b(A_t | S_t)} \right) \left( \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t | S_t) \right) \left( \sum_{t=0}^{T-1} \gamma^t R_{t+1} \right) \right] \\ &= \mathbb{E}_b \left[ \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t | S_t) \left( \prod_{i=0}^t \frac{\pi_\theta(A_i | S_i)}{b(A_i | S_i)} \right) \left( \sum_{i=t}^{T-1} \gamma^i R_{i+1} \right) \right], \quad (42) \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E}_b \left[ \hat{\nabla} J(\theta) \right] \\ &= \mathbb{E}_b \left[ \frac{1}{m} \sum_{n=1}^m \left[ \sum_{t=0}^{T^n-1} \nabla \log \pi_\theta(A_t^n | S_t^n) \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \left( \sum_{i=t}^{T^n-1} \gamma^i R_{i+1}^n \right) \right] \right] \\ &= \mathbb{E}_b \left[ \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t | S_t) \left( \prod_{i=0}^t \frac{\pi_\theta(A_i | S_i)}{b(A_i | S_i)} \right) \left( \sum_{i=t}^{T-1} \gamma^i R_{i+1} \right) \right] = \nabla J(\theta) \quad (43) \end{aligned}$$

Now,

$$\begin{aligned} & \nabla \hat{J}_m(\theta) \\ &= \nabla \left[ \sum_{t=0}^{T^n-1} \gamma^t \frac{1}{m} \sum_{n=1}^m R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \right] \\ &= \sum_{t=0}^{T^n-1} \gamma^t \frac{1}{m} \sum_{n=1}^m R_{t+1}^n \left( \prod_{i=0}^t \frac{\pi_\theta(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \\ &\quad \left( \sum_{i=0}^t \nabla \log \pi_\theta(A_i^n | S_i^n) \right) \\ &= \frac{1}{m} \sum_{n=1}^m \left[ \sum_{t=0}^{T^n-1} \nabla \log \pi_\theta(A_t^n | S_t^n) \right] \end{aligned}$$

$$\begin{aligned} & \left( \prod_{i=0}^t \frac{\pi_{\theta}(A_i^n | S_i^n)}{b(A_i^n | S_i^n)} \right) \left( \sum_{i=t}^{T^n-1} \gamma^i R_{i+1}^n \right) \\ &= \widehat{\nabla} J(\theta) \end{aligned} \quad (44)$$

It is easy to see that  $\nabla J(\theta)$  is  $L$ -Lipschitz w.r.t  $\theta$  and  $\|\widehat{\nabla} J(\theta)\| \leq L$  using (43), (44) and Lemmas 3 and 4.

By using a completely parallel argument to the initial passage in the proof of Theorem 1 leading up to (25), we obtain

$$\begin{aligned} & J(\theta_k) - J(\theta_{k+1}) \\ & \leq L\alpha \left\| \nabla J(\theta_k) - \widehat{\nabla} J(\theta_k) \right\| + \frac{L\alpha^2}{2} \left\| \widehat{\nabla} J(\theta_k) \right\|^2 \\ & \quad - \alpha \left\| \mathcal{P}_{\Theta}(\theta_k, \nabla J(\theta_k), \alpha) \right\|^2. \end{aligned} \quad (45)$$

Summing up (45) from  $k = 0, \dots, N-1$ , we obtain

$$\begin{aligned} & \sum_{k=0}^{N-1} \alpha \left\| \mathcal{P}_{\Theta}(\theta_k, \nabla J(\theta_k), \alpha) \right\|^2 \\ & \leq (J^* - J(\theta_0)) + L\alpha \sum_{k=0}^{N-1} \left\| \nabla J(\theta_k) - \widehat{\nabla} J(\theta_k) \right\| \\ & \quad + \frac{L\alpha^2}{2} \sum_{k=0}^{N-1} \left\| \widehat{\nabla} J(\theta_k) \right\|^2. \end{aligned} \quad (46)$$

Taking expectations on both sides of (46), we obtain

$$\begin{aligned} & \sum_{k=0}^{N-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_{\Theta}(\theta_k, \nabla J(\theta_k), \alpha) \right\|^2 \right] \\ & \leq (J^* - J(\theta_0)) + \frac{L\alpha^2}{2} \sum_{k=0}^{N-1} \mathbb{E} \left[ \left\| \widehat{\nabla} J(\theta_k) \right\|^2 \right] \\ & \quad + L\alpha \sum_{k=0}^{N-1} \mathbb{E} \left[ \left\| \nabla J(\theta_k) - \widehat{\nabla} J(\theta_k) \right\| \right] \quad (\text{from (43)}) \\ & \leq (J^* - J(\theta_0)) + \frac{L\alpha^2}{2} \sum_{k=0}^{N-1} \mathbb{E} \left[ \left\| \widehat{\nabla} J(\theta_k) \right\|^2 \right] \\ & \quad + L\alpha \sum_{k=0}^{N-1} \left( \mathbb{E} \left[ \left\| \mathbb{E}_b \left[ \widehat{\nabla} J(\theta_k) \right] - \widehat{\nabla} J(\theta_k) \right\|^2 \right] \right)^{\frac{1}{2}} \\ & \quad \quad \quad (\text{from (43)}) \\ & \leq (J^* - J(\theta_0)) + \frac{L\alpha^2}{2} \sum_{k=0}^{N-1} \mathbb{E} \left[ \left\| \widehat{\nabla} J(\theta_k) \right\|^2 \right] \\ & \quad + L\alpha \sum_{k=0}^{N-1} \left( \mathbb{E} \left[ \left\| \widehat{\nabla} J(\theta_k) \right\|^2 \right] \right)^{\frac{1}{2}} \\ & \quad \quad \quad (\text{since } \mathbb{E}[(X - \mathbb{E}[X|Y])^2] \leq \mathbb{E}[X^2]) \end{aligned}$$

$$\leq (J^* - J(\theta_0)) + \frac{L^3}{2} \alpha^2 N + L^2 \alpha N \quad (47)$$

Since  $\mathbb{P}(R = k) = \frac{\alpha}{\sum_{k=0}^{N-1} \alpha}$ , we obtain

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathcal{P}_{\Theta}(\theta_R, \nabla J(\theta_R), \alpha) \right\|^2 \right] \\ &= \frac{\sum_{k=0}^{N-1} \alpha \mathbb{E} \left[ \left\| \mathcal{P}_{\Theta}(\theta_k, \nabla J(\theta_k), \alpha) \right\|^2 \right]}{\sum_{k=0}^{N-1} \alpha} \\ &\leq \frac{(J^* - J(\theta_0)) + \frac{L^3}{2} \alpha^2 N + L^2 \alpha N}{\sum_{k=0}^{N-1} \alpha}. \end{aligned}$$

Since  $\alpha = \frac{1}{\sqrt{N}}$ , we obtain

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathcal{P}_{\Theta}(\theta_R, \nabla J(\theta_R), \alpha) \right\|^2 \right] \\ &\leq \frac{(J^* - J(\theta_0)) + \frac{L^3}{2}}{\sqrt{N}} + L^2. \end{aligned}$$

□

## 6 Simulation analysis

We conducted experiments on an control problem called CartPole from OpenAI Gym toolkit [9]. The problem is to balance a pole which is attached to a moving cart. The state space is continuous and each state is a quadruple (cart position, cart velocity, pole angle, pole velocity at tip) and the action space is discrete (push cart to the left and push cart to the right). We fixed the initial state. The problem is reset to the initial state either after 200 steps, the pole tilt more than 15 degrees from vertical, or the cart moves more than 2.4 units from the centre. We receive a reward of +1 for each timestep in which the pole is upright.

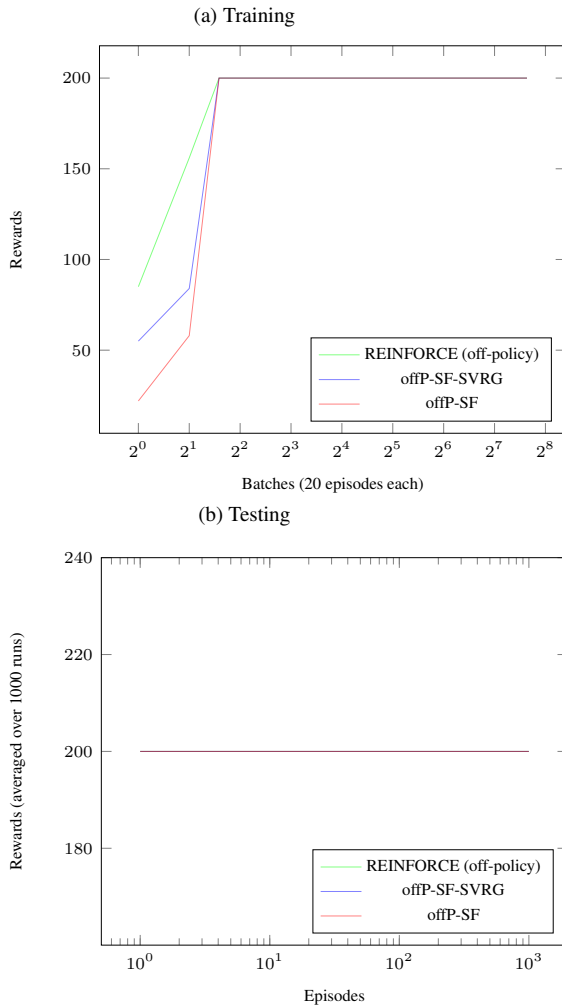
We have used the samples collected using an  $\epsilon$ -greedy behavior policy and a target policy which follows an exponential softmax distribution. We have compared the performance of OffP-SF, OffP-SF-SVRG and REINFORCE (off-policy variant) algorithms. In Figure 1 we plot the performance of the aforementioned algorithms.

## 7 Conclusions and future work

We proposed two policy gradient algorithms for off-policy control in a RL context. Both algorithms incorporated a smoothed functional scheme for gradient estimation. For both algorithms, we provided non-asymptotic bounds that establish convergence to an approximate stationary point.



Figure 1: CartPole with fixed initial state



As future work, it would be interesting to study the global convergence properties of our algorithms under additional assumptions such as those used in [41, 42]. An orthogonal research direction is to incorporate feature-based representations and function approximation together with smoothed functional gradient estimation, and study the non-asymptotic performance of the resulting actor-critic algorithms. Another direction of future work is to check if our algorithms are globally convergent under additional assumptions such as those in [23].

## References

- [1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, volume
- 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.
- [2] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.
- [3] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108.
- [4] J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *CoRR*, abs/1906.01786, 2019.
- [5] J. Bhandari and D. Russo. A note on the linear convergence of policy gradient methods. *CoRR*, abs/2007.11120, 2020.
- [6] S. Bhatnagar. An actor-critic algorithm with function approximation for discounted cost constrained markov decision processes. *Systems & Control Letters*, 59:760–766, 2010.
- [7] S. Bhatnagar and S. Kumar. A simultaneous perturbation stochastic approximation-based actor-critic algorithm for Markov decision processes. *IEEE Transactions on Automatic Control*, 49(4):592–598, 2004.
- [8] S. Bhatnagar, H. Prasad, and L. A. Prashanth. *Stochastic recursive algorithms for optimization. Simultaneous perturbation methods*, volume 434. Springer-Verlag London, 01 2013.
- [9] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016.
- [10] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [11] T. Degris, M. White, and R. S. Sutton. Off-policy actor-critic. In *International Conference on Machine Learning*, pages 179–186, 2012.
- [12] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions*. Number 36 in Monographs on statistics and applied probability. Chapman & Hall, 1990. ISBN 0412314304.
- [13] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.
- [14] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005. ISBN 0898715857.
- [15] M. C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 19, pages 575–616. Elsevier, 2006.
- [16] M. C. Fu. Stochastic gradient estimation. In M. C. Fu, editor, *Handbook on Simulation Optimization*, chapter 5. Springer, 2015.
- [17] X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: An iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018. ISSN 1573-7691.
- [18] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23: 2341–2368, 2013.
- [19] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Neural Information Processing Systems*, pages 315–323. Curran Associates Inc., 2013.
- [20] S. M. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.
- [21] V. Katkovnik and Y. Kulchitsky. Convergence of a class of random search algorithms. *Automation and Remote Control*, 33: 1321–1326, 1972.
- [22] S. Liu, X. Li, P. Chen, J. Haupt, and L. Amini. Zeroth-order

- stochastic projected gradient descent for nonconvex optimization. In *IEEE Global Conference on Signal and Information Processing*, pages 1179–1183, 2018.
- [23] Y. Liu, K. Zhang, T. Basar, and W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.
- [24] D. Lyu, Q. Qi, M. Ghavamzadeh, H. Yao, T. Yang, and B. Liu. Variance-reduced off-policy memory-efficient policy search, 2020.
- [25] P. Marbach and J. N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.
- [26] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020.
- [27] V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Springer-Verlag, 1986.
- [28] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović. On the linear convergence of random search for discrete-time lqr. *IEEE Control Systems Letters*, 5(3):989–994, 2021.
- [29] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017. ISSN 1615-3383.
- [30] Y. E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004. ISBN 978-1-4613-4691-3.
- [31] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035. PMLR, 10–15 Jul 2018.
- [32] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323. JMLR.org, 2016.
- [33] R. Y. Rubinstein. *Some problems in monte carlo optimization*. PhD thesis, 1969.
- [34] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, 2017. ISSN 1532-4435.
- [35] Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR, 2019.
- [36] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.
- [37] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition, 2018.
- [38] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pages 1057–1063, 1999.
- [39] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992. ISSN 0885-6125.
- [40] P. Xu, F. Gao, and Q. Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 541–551, 2020.
- [41] J. Zhang, J. Kim, B. O’Donoghue, and S. Boyd. Sample efficient reinforcement learning with reinforce, 2020.
- [42] K. Zhang, A. Koppel, H. Zhu, and T. Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control. Optim.*, 58(6):3586–3612, 2020.
- [43] S. Zhang, W. Boehmer, and S. Whiteson. Generalized off-policy actor-critic. In *Advances in Neural Information Processing Systems*, volume 32, pages 2001–2011, 2019.
- [44] S. Zhang, B. Liu, H. Yao, and S. Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11204–11213. PMLR, 13–18 Jul 2020.