On the Stochastic (Variance-Reduced) Proximal Gradient Method for Regularized Expected Reward Optimization

Anonymous authors
Paper under double-blind review

Abstract

We consider a regularized expected reward optimization problem in the non-oblivious setting that covers many existing problems in reinforcement learning (RL). In order to solve such an optimization problem, we apply and analyze the classical stochastic proximal gradient method. In particular, the method has shown to admit an $O(\epsilon^{-4})$ sample complexity to an ϵ-stationary point, under standard conditions. Since the variance of the classical stochastic gradient estimator is typically large, which slows down the convergence, we also apply an efficient stochastic variance-reduce proximal gradient method with an importance sampling based ProbAbilistic Gradient Estimator (PAGE). Our analysis shows that the sample complexity can be improved from $O(\epsilon^{-4})$ to $O(\epsilon^{-3})$ under additional conditions. Our results on the stochastic (variance-reduced) proximal gradient method match the sample complexity of their most competitive counterparts for discounted Markov decision processes under similar settings. To the best of our knowledge, the proposed methods represent a novel approach in addressing the general regularized reward optimization problem.

1 Introduction

Reinforcement learning (RL) Sutton & Barto (2018) has recently become a highly active research area of machine learning that learns to make sequential decisions via interacting with the environment. In recent years, RL has achieved tremendous success so far in many applications such as control, job scheduling, online advertising, and game-playing Zhang & Dietterich (1995); Pednault et al. (2002); Mnih et al. (2013), to mention a few. One of the central tasks of RL is to solve a certain (expected) reward optimization problem for decision-making. Following the research theme, we consider the following problem of maximizing the regularized expected reward:

$$\max_{\theta\in\mathbb{R}^n} \ \mathcal{F}(\theta) := \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\right] - \mathcal{G}(\theta), \tag{1}$$

where $\mathcal{G}:\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is a closed proper convex (possibly nonsmooth) function, $x\in\mathbb{R}^d$, $\mathcal{R}_\theta:\mathbb{R}^d\to\mathbb{R}$ is the reward function depending on the parameter θ, and π_θ denotes the probability distribution over a given subset $\mathcal{S}\subseteq\mathbb{R}^d$ parameterized by $\theta\in\mathbb{R}^n$. By adapting the convention in RL, we call π_θ a policy parameterized by θ. Moreover, for the rest of this paper, we denote $\mathcal{J}(\theta) := \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\right]$ as the expected reward function in the *non-oblivious* setting. The learning objective is to learn a decision rule via finding the policy parameter θ that maximizes the regularized expected reward. To the best of our knowledge, the study on the general model (1) has been limited in the literature. Hence, developing and analyzing algorithmic frameworks for solving the problem is of great interest.

There are large body of works in supervised learning focusing on the *oblivious* setting Zhang (2004); Hastie et al. (2009); Shapiro et al. (2021), i.e., $\mathcal{J}(\theta) := \mathbb{E}_{x\sim\pi}\left[\mathcal{R}_\theta(x)\right]$, where x is sampled from an invariant distribution π. Clearly, problem (1) can be viewed as a generalization of those machine learning problems with oblivious objective functions. In the literature, an RL problem is often formulated as a discrete-time and discounted Markov decision processes (MDPs) Sutton & Barto (2018) which aims to learn an optimal policy via optimizing the (discounted) cumulative sum of rewards. We can also see that the learning objective of an MDP can be covered by the problem (1) with the property that the function $\mathcal{R}(x)$ does not depend on θ (see Example 3.3). Recently, the application of RL for solving combinatorial optimization (CO) problems

which are typically NP-hard has attracted much attention. These CO problems may include the traveling salesman problem and related problems Bello et al. (2016); Mazyavkina et al. (2021), the reward optimization problem arising from the finite expression method Liang & Yang (2022); Song et al. (2023), and the general binary optimization problem Chen et al. (2023), to name just a few. The common key component of the aforementioned applications is the reward optimization, which could also be formulated as problem (1). There also exist problems with general reward functions that are outside the scope of cumulative sum of rewards of trajectories that are used in MDPs. An interesting example is the MDP with general utilities; see, e.g., Zhang et al. (2020a); Kumar et al. (2022); Barakat et al. (2023) and references therein.

Adding a regularizer to the objective function is a commonly used technique to impose desirable structures to the solution and/or to greatly enhance the expression power and applicability of RL Lan (2023); Zhan et al. (2023). When one considers the direct/simplex parameterization Agarwal et al. (2021) of π_θ, a regularization function using the indicator function for the standard probability simplex is needed. Moreover, by using other indicator functions for general convex sets, one is able to impose some additional constraints on the parameter θ. For the softmax parameterization, one may also enforce a bounded constraint to θ to prevent it taking values that are too large. This can avoid potential numerical issues, including the overflow error on a floating point system. On the other hand, there are incomplete parametric policy classes, such as the log-linear and neural policy classes, that are often formulated as $\{\pi_\theta | \theta \in \Theta\}$, where Θ is a closed convex set Agarwal et al. (2021). In this case, the indicator function is still necessary and useful. Some recent works (see, e.g., Ahmed et al. (2019); Agarwal et al. (2020); Mei et al. (2020); Cen et al. (2022)) have investigated the impact of the entropy regularization for MDPs. Systematic studies on general convex regularization for MDPs have been limited until the recent works Pham et al. (2020); Lan (2023); Zhan et al. (2023). Finally, problem (1) takes a similar form as the stochastic optimization problem with decision-dependent distributions and (strongly) convex loss functions considered in Drusvyatskiy & Xiao (2023) and references therein. Consequently, we can see that problem (1) is in fact quite general and has promising modeling power, as it covers many existing problems in the literature.

The purpose of this paper is to leverage existing tools and results in MDPs and nonconvex optimization for solving the general regularized expected reward optimization problem (1) with general policy parameterization, which, to the best of our knowledge, has not been formally considered in the RL literature. It is well known that the policy gradient method Williams (1992); Sutton et al. (1999); Baxter & Bartlett (2001), which lies in the heart of RL, is one of the most competitive and efficient algorithms due to its simplicity and versatility. Moreover, the policy gradient method is readily implemented and can be paired with other effective techniques. In this paper, we observe that the stochastic proximal gradient method, which shares the same spirit of the policy gradient method, can be applied directly for solving the targeted problem (1) with convergence guarantees to a stationary point. Since the classical stochastic gradient estimator typically introduces a large variance, there is also a need to consider designing advanced stochastic gradient estimators with smaller variances. To this end, we shall also look into a certain stochastic variance-reduced proximal gradient method and analyze its convergence properties. In particular, the contributions of this paper are summarized as follows.

- We consider a novel and general regularized reward optimization model (1) that covers many existing important models in the machine learning and optimization literature. Thus, problem (1) admits a promising modeling power which encourages potential applications.
- In order to solve our targeted problem, we consider applying the classical stochastic proximal gradient method and analyze its convergence properties. We first demonstrate that the gradient of $\mathcal{J}(\cdot)$ is Lipschitz continuous under standard conditions with respect to the reward function $\mathcal{R}_\theta(\cdot)$ and the parameterized policy $\pi_\theta(\cdot)$. Using the L-smoothness of $\mathcal{J}(\cdot)$, we then show that the classical stochastic proximal gradient method with a constant step-size (depending only on the Lipschitz constant for $\nabla_\theta \mathcal{J}(\cdot)$) for solving problem (1) outputs an ϵ-stationary point (see Definition 3.4) within $T := O(\epsilon^{-2})$ iterations, and the sample size for each iteration is $O(\epsilon^{-2})$, where $\epsilon > 0$ is a given tolerance. Thus, the total sample complexity becomes $O(\epsilon^{-4})$, which matches the current state-of-the-art sample complexity of the classical stochastic policy gradient for MDPs; see e.g., Williams (1992); Baxter & Bartlett (2001); Zhang et al. (2020b); Xiong et al. (2021); Yuan et al. (2022).

- Moreover, in order to further reduce the variance of the stochastic gradient estimator, we utilize an importance sampling based probabilistic gradient estimator which leads to an efficient single-looped variance reduced method. The application of this probabilistic gradient estimator is motivated by the recent progress in developing efficient stochastic variance-reduced gradient methods for solving stochastic optimization Li et al. (2021b) and (unregularized) MDPs Gargiani et al. (2022). We show that, under additional technical conditions, the total sample complexity is improved from $O(\epsilon^{-4})$ to $O(\epsilon^{-3})$. This result again matches the results of some existing competitive variance-reduced methods for MDPs Papini et al. (2018); Xu et al. (2019); Pham et al. (2020); Huang et al. (2021); Yang et al. (2022); Gargiani et al. (2022). Moreover, to the best of our knowledge, the application of the above probabilistic gradient estimator is new for solving the regularized expected reward optimization (1).

The rest of this paper is organized as follows. We first summarize some relative works in Section 2. Next, in Section 3, we present some background information that are needed for the exposition of this paper. Then, in Section 4, we describe the classical stochastic proximal gradient method for solving (1) and present the convergence properties of this method under standard technical conditions. Section 5 is dedicated to describing and analyzing the stochastic variance-reduced proximal gradient method with an importance sampling based probabilistic gradient estimator. Finally, we make some concluding remarks, and list certain limitations and future research directions of this paper in Section 6.

2 Related Work

The policy gradient method. One of the most influential algorithms for solving RL problems is the policy gradient method, built upon the foundations established in Williams (1992); Sutton et al. (1999); Baxter & Bartlett (2001). Motivated by the empirical success of the policy gradient method and its variants, analyzing the convergence properties for these methods has long been one of the most active research topics in RL. Since the objective function $\mathcal{J}(\theta)$ is generally nonconcave, early works Sutton et al. (1999); Pirotta et al. (2015) focused on the asymptotic convergence properties to a stationary point. By utilizing the special structure in (entropy regularized) MDPs, recent works Liu et al. (2019); Mei et al. (2020); Agarwal et al. (2021); Li et al. (2021a); Xiao (2022); Cen et al. (2022); Lan (2023); Fatkhullin et al. (2023) provided some exciting results on the global convergence. Meanwhile, since the exact gradient of the objective function can hardly be computed, sampling-based approximated/stochastic gradients have gained much attention. Therefore, many works investigated the convergence properties, including the iteration and sample complexities, for these algorithms with inexact gradients; see e.g., Zhang et al. (2020b); Liu et al. (2020); Zhang et al. (2021b); Xiong et al. (2021); Yuan et al. (2022); Lan (2023) and references therein.

Variance reduction. While the classical stochastic gradient estimator is straightforward and simple to implement, one of its most critical issues is that the variance of the inexact gradient estimator can be large, which generally slows down the convergence of the algorithm. To alleviate this issue, an attractive approach is to pair the sample-based policy gradient methods with certain variance-reduced techniques. Variance-reduced methods were originally developed for solving (oblivious) stochastic optimization problems Johnson & Zhang (2013); Nguyen et al. (2017); Fang et al. (2018); Li et al. (2021b) typically arising from supervised learning tasks. Motivated by the superior theoretical properties and practical performance of the stochastic variance-reduced gradient methods, similar algorithmic frameworks have recently been applied for solving MDPs Papini et al. (2018); Xu et al. (2019); Yuan et al. (2020); Pham et al. (2020); Huang et al. (2021); Yang et al. (2022); Gargiani et al. (2022).

Stochastic optimization with decision-dependent distributions. Stochastic optimization is the core of modern machine learning applications, whose main objective is to learn a decision rule from a limited data sample that is assumed to generalize well to the entire population Drusvyatskiy & Xiao (2023). In the classical supervised learning framework Zhang (2004); Hastie et al. (2009); Shapiro et al. (2021), the underlying data distribution is assumed to be static, which turns out to be a crucial assumption when analyzing the convergence properties of the common stochastic optimization algorithms. On the other hand, there are problems where the distribution changes over the course of iterations of a specific algorithm, and these are closely related to the concept of performative prediction Perdomo et al. (2020). In this case, understanding the convergence properties of the algorithm becomes more challenging. Toward this,

some recent progress has been made on (strongly) convex stochastic optimization with decision-dependent distributions Mendler-Dünner et al. (2020); Perdomo et al. (2020); Drusvyatskiy & Xiao (2023). Moreover, other works have also considered nonconvex problems and obtained some promising results; see Dong et al. (2023); Jagadeesan et al. (2022) and references therein. Developing theoretical foundation for these problems has become a very active field nowadays.

RL with general utilities. It is known that the goal of an agent associated with an MDP is to seek an optimal policy via maximizing the cumulative discounted reward Sutton & Barto (2018). However, there are decision problems of interest having more general forms. Beyond the scope of the expected cumulative reward in MDPs, some recent works also looked into RL problems with general utilities; see e.g., Zhang et al. (2020a); Kumar et al. (2022); Barakat et al. (2023) as mentioned previously. Global convergence results can also be derived via investigating the hidden convex structure Zhang et al. (2020a) inherited from the MDP.

3 Preliminary

In this paper, we assume that the optimal objective value for problem (1), denoted by $\mathcal{F}^*$, is finite and attained, and the reward function $\mathcal{R}_\theta(\cdot)$ satisfies the following assumption.

Assumption 3.1. *The following three conditions with respect to the function $\mathcal{R}_\theta(\cdot)$ hold:*

1. *There exists a constant $U > 0$ such that*

$$\sup_{\theta\in\mathbb{R}^n, x\in\mathbb{R}^d} |\mathcal{R}_\theta(x)| \leq U.$$

2. *$\mathcal{R}_\theta(\cdot)$ is twice continuously differentiable with respect to θ, and there exist positive constants $\widetilde{C}_g$ and $\widetilde{C}_h$ such that*

$$\sup_{\theta\in\mathbb{R}^n,\ x\in\mathbb{R}^d} \|\nabla_\theta \mathcal{R}_\theta(x)\| \leq \widetilde{C}_g, \qquad \sup_{\theta\in\mathbb{R}^n,\ x\in\mathbb{R}^d} \left\|\nabla_\theta^2 \mathcal{R}_\theta(x)\right\|_2 \leq \widetilde{C}_h.$$

The first condition on the boundedness of the function $\mathcal{R}_\theta(\cdot)$, which is commonly assumed in the literature Sutton & Barto (2018), ensures that $\mathcal{J}(\theta)$ is well-defined. And the second condition will be used to guarantee the well-definiteness and L-smoothness of the gradient $\nabla_\theta \mathcal{J}(\theta)$.

To determine the (theoretical) learning rate in our algorithmic frameworks, we also need to make some standard assumptions to establish the L-smoothness of $\mathcal{J}(\cdot)$.

Assumption 3.2 (Lipschitz and smooth policy assumption)**.** *The function $\log \pi_\theta(x)$ is twice differential with respect to $\theta \in \mathbb{R}^n$ and there exist positive constants C_g and C_h such that*

$$\sup_{x\in\mathbb{R}^d,\ \theta\in\mathbb{R}^n} \|\nabla_\theta \log \pi_\theta(x)\| \leq C_g, \qquad \sup_{x\in\mathbb{R}^d,\ \theta\in\mathbb{R}^n} \left\|\nabla_\theta^2 \log \pi_\theta(x)\right\|_2 \leq C_h.$$

This assumption is a standard one and commonly employed in the literature when studying the convergence properties of the policy gradient method for MDPs; see e.g., Pirotta et al. (2015); Papini et al. (2018); Xu et al. (2020); Pham et al. (2020); Zhang et al. (2021a); Yang et al. (2022) and references therein.

Under Assumption 3.1 and Assumption 3.2, it is easy to verify that the gradient for the expected reward function $\mathcal{J}(\theta)$ can be written as:

$$\nabla_\theta \mathcal{J}(\theta) := \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right].$$

We next present an example on the discrete-time discounted MDP, which can be covered by the general model (1).

Example 3.3 (MDP)**.** *We denote a discrete-time discounted MDP as $\mathcal{M} := \{S, A, P, R, \gamma, \rho\}$, where S and A denote the state space and the action space, respectively, $P(s'|s,a)$ is the state transition probability from*

s to s' after selecting the action a, $R : S \times A \to [0, U]$ is the reward function that is assumed to be uniformly bounded by a constant $U > 0$, $\gamma \in [0, 1)$ is the discount factor, and ρ is the initial state distribution.

The agent selects actions according to a stationary random policy $\tilde{\pi}_\theta(\cdot|\cdot) : A \times S \to [0, 1]$ parameterized by $\theta \in \mathbb{R}^n$. Given an initial state $s_0 \in S$, a trajectory $x := \{s_t, a_t, r_{t+1}\}_{t=0}^{\infty}$ can then be generated, where $s_0 \sim \rho$, $a_t \sim \tilde{\pi}_\theta(\cdot|s_t)$, $r_{t+1} = R(s_t, a_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$, and the accumulated discounted reward of the trajectory x can be defined as $\mathcal{R}(x) := \sum_{t=0}^{\infty} \gamma^t r_{t+1}$. Then, the learning objective is to compute an optimal parameter θ^* that maximizes the expected reward function $\mathcal{J}(\theta)$, i.e.,

$$\theta^* = \text{argmax}_\theta \, \mathcal{J}(\theta) := \mathbb{E}_{x \sim \pi_\theta} \left[\mathcal{R}(x) \right], \tag{2}$$

where

$$\pi_\theta(x) := \rho(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \tilde{\pi}_\theta(a_t|s_t)$$

denotes the probability distribution of a single trajectory x being sampled from π_θ.

In the special case when $S = \{s\}$ (i.e., $|S| = 1$) and $\gamma = 0$, the MDP reduced to a multi-armed bandit problem Robbins (1952) with a reward function simplified as $R : A \to \mathbb{R}$. Particularly, a trajectory $x = \{s, a\}$ with the horizon $H_x = 0$ is generated, where $a \sim \pi_\theta(\cdot) := \tilde{\pi}_\theta(\cdot|s)$, and the accumulated discounted reward reduces to $\mathcal{R}(x) = R(a)$. As a consequence, problem (2) can be simplified as

$$\min_{\theta \in \mathbb{R}^n} \; \mathcal{J}(\theta) = \mathbb{E}_{a \sim \pi_\theta} \left[R(a) \right].$$

By adding a convex regularizer $\mathcal{G}(\theta)$ to problem (2), we get the following regularized MDP:

$$\min_{\theta} \; \mathbb{E}_{x \sim \pi_\theta} \left[\mathcal{R}(x) \right] - \mathcal{G}(\theta),$$

which was considered in Pham et al. (2020). However, it is clear that $\mathcal{R}(x)$ does not depend on θ. Hence, the above regularized MDP is a special case of the proposed regularized reward optimization problem (1).

One can check that the gradient $\nabla_\theta \mathcal{J}(\theta)$ has the following form Yuan et al. (2022):

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{x \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \sum_{t'=0}^{\infty} \nabla_\theta \log \tilde{\pi}_\theta(a_{t'}|s_{t'}) \right].$$

Being a composite optimization problem, problem (1) admits the following first-order stationary condition

$$0 \in -\nabla_\theta \mathcal{J}(\theta) + \partial \mathcal{G}(\theta). \tag{3}$$

Here, $\partial \mathcal{G}(\cdot)$ denotes the subgradient of the proper closed and convex function $\mathcal{G}(\cdot)$ which is defined as

$$\partial \mathcal{G}(\theta) := \{ g \in \mathbb{R}^n \; : \; \mathcal{G}(\theta') \geq \mathcal{G}(\theta) + \langle g, \theta' - \theta \rangle \, , \; \forall \theta \}.$$

It is well-known that $\partial \mathcal{G}(\theta)$ is a nonempty closed convex subset of $\mathbb{R}^n$ for any $\theta \in \mathbb{R}^n$ such that $\mathcal{G}(\theta) < \infty$ (see e.g., Rockafellar (1997)). Note that any optimal solution of problem (1) satisfies the condition (3), while the reverse statement is generally not valid for nonconcave problems, including the problem (1). The condition (3) leads to the following concept of stationary points for problem (1).

Definition 3.4. *A point $\theta \in \mathbb{R}^n$ is called a stationary point for problem (1) if it satisfies the condition (3). Given a tolerance $\epsilon > 0$, a stochastic optimization method attains an (expected) ϵ-stationary point, denoted as $\theta \in \mathbb{R}^n$, if*

$$\mathbb{E}_T \left[\text{dist} \left(0, -\nabla_\theta \mathcal{J}(\theta) + \partial \mathcal{G}(\theta) \right)^2 \right] \leq \epsilon^2,$$

where the expectation is taken with respect to all the randomness caused by the algorithm, after running it T iterations.

Remark 3.5 (Gradient mapping)**.** *Note that the optimality condition* (3) *can be rewritten as*

$$0 = G_\eta(\theta) := \frac{1}{\eta}\left[\mathrm{Prox}_{\eta\mathcal{G}}\left(\theta + \eta\nabla_\theta \mathcal{J}(\theta)\right) - \theta\right],$$

for some $\eta > 0$*, where*

$$\mathrm{Prox}_{\eta\mathcal{G}}(\theta) := \mathrm{argmin}_{\theta'}\left\{\mathcal{G}(\theta') + \frac{1}{2\eta}\left\|\theta' - \theta\right\|^2\right\}$$

denotes the proximal mapping of the function $\mathcal{G}(\cdot)$*. The mapping* $G_\eta(\cdot)$ *is called the gradient mapping in the field of optimization Beck (2017). It is easy to verify that if for a* $\theta \in \mathbb{R}^n$*, it holds that*

$$\mathrm{dist}\left(0, -\nabla_\theta \mathcal{J}(\theta) + \partial\mathcal{G}(\theta)\right) \leq \epsilon,$$

then there exists a vector d *satisfying* $\|d\| \leq \epsilon$ *such that*

$$d + \nabla_\theta \mathcal{J}(\theta) \in \partial\mathcal{G}(\theta),$$

which is equivalent to saying that

$$\theta = \mathrm{Prox}_{\eta\mathcal{G}}\left(\eta d + \theta + \eta\nabla_\theta \mathcal{J}(\theta)\right).$$

Moreover, we can verify that (by using the firm nonexpansiveness of $\mathrm{Prox}_{\eta\mathcal{G}}(\cdot)$*; see e.g., Beck (2017))*

$$\|G_\eta(\theta)\| = \frac{1}{\eta}\left\|\mathrm{Prox}_{\eta\mathcal{G}}\left(\theta + \eta\nabla_\theta \mathcal{J}(\theta)\right) - \theta\right\| \leq \|d\| \leq \epsilon.$$

Therefore, we can also characterize an (expected) ϵ*-stationary point by using the following condition*

$$\mathbb{E}_T\left[\left\|G_\eta(\theta)\right\|^2\right] \leq \epsilon^2.$$

The main objective of this paper is to study the convergence properties, including iteration and sample complexities, of the stochastic (variance-reduced) proximal gradient method to a ϵ-stationary point with a pre-specified $\epsilon > 0$. Note that all proofs of our results are presented in the appendix.

4 The stochastic proximal gradient method

In this section, we present and analyze the stochastic proximal gradient method for solving the problem (1). The fundamental idea of the algorithm is to replace the true gradient $\nabla_\theta \mathcal{J}(\theta)$, which are not available for most of the time, with a stochastic gradient estimator in the classical proximal gradient method Beck (2017). The method can be viewed as extensions to the projected policy gradient method with direct parameterization Agarwal et al. (2021) and the stochastic policy gradient method for unregularized MDPs Williams (1992). The detailed description of the algorithm is presented in Algorithm 1.

For notational simplicity, we denote

$$g(x, \theta) := \mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x).$$

From Algorithm 1, we see that at each iteration, N data points, namely $\{x^{t,1}, \ldots, x^{t,N}\}$, are sample according to the current probability distribution π_{θ^t}. Using these data points, we can construct a REINFORCE-type stochastic gradient estimator g^t. Then, the algorithm just performs a proximal gradient ascent updating. Let $T > 0$ be the maximal number of iterations, then a sequence $\{\theta^t\}_{t=1}^T$ can be generated, and the output solution is selected randomly from this sequence. Next, we shall proceed to answer the questions that how to choose the learning rate $\eta > 0$, how large the sample size N should be, and how many iterations for the algorithm to output an ϵ-stationary point for a given $\epsilon > 0$, theoretically. The next lemma establishes the L-smoothness of $\mathcal{J}(\cdot)$ whose proof is given at Appendix A.1.

Algorithm 1 The stochastic proximal gradient method

1: **Input:** initial point θ^0, sample size N and the learning rate $\eta > 0$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Compute the stochastic gradient estimator:

$$g^t := \frac{1}{N} \sum_{j=1}^{N} g(x^{t,j}, \theta^t),$$

where $\{x^{t,1}, \ldots, x^{t,N}\}$ are sampled independently according to π_{θ^t}.
4: Update

$$\theta^{t+1} = \mathrm{Prox}_{\eta\mathcal{G}}\left(\theta^t + \eta g^t\right).$$

5: **end for**
6: **Output:** $\hat{\theta}^T$ selected randomly from the generated sequence $\{\theta^t\}_{t=1}^T$.

Lemma 4.1. *Under Assumptions 3.1 and 3.2, the gradient of $\mathcal{J}$ is L-smooth, i.e.,*

$$\|\nabla_\theta \mathcal{J}(\theta) - \nabla_\theta \mathcal{J}(\theta')\| \leq L \|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^n,$$

with $L := U(C_g^2 + C_h) + \widetilde{C}_h + 2C_g\widetilde{C}_g > 0$.

Remark 4.2 (L-smoothness in MDPs)**.** *For an MDP with finite action space and state space as in Example 3.3, the Lipschitz constant of $\nabla_\theta \mathcal{J}(\cdot)$ can be expressed in terms of $|A|$, $|S|$ and γ. We refer the reader to Agarwal et al. (2021); Xiao (2022) for more details.*

As a consequence of the L-smoothness of the function $\mathcal{J}(\cdot)$, we next show that the learning rate can be chosen as a positive constant upper bounded by a constant depends only on the Lipschitz constant of $\nabla_\theta \mathcal{J}(\cdot)$. For notational complicity, we denote $\Delta := \mathcal{F}^* - \mathcal{F}(\theta^0) > 0$ for the rest of this paper.

Theorem 4.3. *Under Assumptions 3.1 and 3.2, if we set $\eta \in \left(0, \frac{1}{2L}\right)$, then Algorithm 1 outputs a point $\hat{\theta}^T$ satisfying*

$$\begin{aligned} &\mathbb{E}_T\left[\mathrm{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \\ \leq &\left(2 + \frac{2}{\eta L(1-2\eta L)}\right) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] + \frac{\Delta}{T}\left(\frac{2}{\eta} + \frac{4}{\eta(1-2\eta L)}\right), \end{aligned}$$

where $\mathbb{E}_T$ is defined in Definition 3.4.

The proof of the above theorem is provided in Appendix A.2. From this theorem, if one sets $g^t = \nabla_\theta \mathcal{J}(\theta^t)$, i.e., $\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2 = 0$, then there is no randomness along the iterations and the convergence property is reduced to

$$\min_{1 \leq t \leq T} \mathrm{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^t) + \partial\mathcal{G}(\hat{\theta}^t)\right) = O\left(\frac{1}{\sqrt{T}}\right),$$

which is implied by classical results on proximal gradient method (see e.g., Beck (2017)). However, since the exact full gradient $\nabla_\theta \mathcal{J}(\theta)$ is rarely computable, it is common to require the variance (i.e., the trace of the covariance matrix) of the stochastic estimator to be bounded. The latter condition plays an essential role in analyzing stochastic first-order methods for solving nonconvex optimization problems, including RL applications; see e.g., Beck (2017); Papini et al. (2018); Shen et al. (2019); Lan (2020); Yang et al. (2022).

Lemma 4.4. *Under Assumptions 3.1 and 3.2, there exists a constant $\sigma > 0$ such that for any θ,*

$$\mathbb{E}_{x \sim \pi_\theta}\left[\|g(x, \theta) - \nabla_\theta \mathcal{J}(\theta)\|^2\right] \leq \sigma^2.$$

The proof of Lemma 4.4 is given in Appendix A.3. By choosing a suitable sample size N, we can rely on Lemma 4.4 to make the term $\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right]$ in Theorem 4.3 small, for every $t \leq T$. Then, Theorem

4.3 implies that Algorithm 1 admits an expected $O(T^{-1})$ convergence rate to a stationary point. These results are summarized in the following theorem; see Appendix A.4 for a proof.

Theorem 4.5. *Suppose that Assumptions 3.1 and 3.2 hold. Let $\epsilon > 0$ be a given accuracy. Running the Algorithm 1 for*

$$T := \left\lceil \frac{\Delta}{\epsilon^2}\left(\frac{4}{\eta} + \frac{8}{\eta(1-2\eta L)}\right)\right\rceil = O(\epsilon^{-2})$$

iterations with the learning rate $\eta < \frac{1}{2L}$ and the sample size

$$N := \left\lceil \frac{\sigma^2}{\epsilon^2}\left(4 + \frac{4}{\eta L(1-2\eta L)}\right)\right\rceil = O(\epsilon^{-2})$$

outputs a point $\hat{\theta}^T$ satisfying

$$\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \leq \epsilon^2.$$

Moreover, the sample complexity is $O(\epsilon^{-4})$.

As already mentioned in the introduction, the total sample complexity of Algorithm 1 to an ϵ-stationary point is shown to be $O(\epsilon^{-4})$, which matches the most competitive sample complexity of the classical stochastic policy gradient for MDPs Williams (1992); Baxter & Bartlett (2001); Zhang et al. (2020b); Xiong et al. (2021); Yuan et al. (2022).

Remark 4.6 (Sample size)**.** *Note that the current state-of-the-art iteration complexity for the (small-batch) stochastic gradient descent method is $T := O(\epsilon^{-2})$ with $\eta_t := \min\{O(L^{-1}), O(T^{-1/2})\}$; see, e.g., Ghadimi & Lan (2013). The reason for requiring larger batch-size in Theorem 4.5 is to allow a constant learning rate. To the best of our knowledge, to get the same convergence properties as Theorem 4.5 under the same conditions for problem* (1)*, the large batch-size is required.*

Remark 4.7 (Global convergence)**.** *As mentioned in introduction, some recent progress has been made for analyzing the global convergence properties of the policy gradient methods for MDPs, which greatly rely on the concepts of gradient domination and its extensions Agarwal et al. (2021); Mei et al. (2020); Xiao (2022); Yuan et al. (2022); Gargiani et al. (2022). This concept is also highly related to the classical PŁ-condition Polyak (1963) and KŁ-condition Bolte et al. (2007) in the field of optimization. The key idea is to assume or verify that the difference between the optimal objective function value, namely $\mathcal{F}^*$, and $\mathcal{F}(\theta)$ can be bounded by the quantity depending on the norm of the gradient mapping at an arbitrary point. In particular, suppose that there exists a positive constant ω such that*

$$\|G_\eta(\theta))\| \geq 2\sqrt{\omega}\left(\mathcal{F}^* - \mathcal{F}(\theta)\right), \quad \forall\, \theta \in \mathbb{R}^n,$$

where G_η is defined in Remark 3.5 (see e.g., Xiao (2022)). Then, after running Algorithm 2 for $T = O(\epsilon^{-2})$ iterations, one can easily check that

$$\mathbb{E}_T\left[\mathcal{F}^* - \mathcal{F}(\hat{\theta}^T)\right] \leq \frac{1}{2\sqrt{\omega}}\epsilon.$$

As a conclusion, by assuming or verifying stronger conditions, one can typically show that any stationary point of the problem (1) *is also a globally optimal solution. This shares the same spirit of Zhang et al. (2020a) for MDPs with general utilities. We leave it as a future research to analyze the global convergence of the problem* (1)*.*

5 Variance reduction via PAGE

Recall from Theorem 4.3 that, there is a trade-off between the sample complexity and the iteration complexity of Algorithm 1. In particular, while there is little room for us to improve the term $\frac{\Delta}{T}\left(\frac{2}{\eta} + \frac{4}{\eta(1-2\eta L)}\right)$ which corresponds to the iteration complexity, it is possible to construct g^t in an advanced manner to

improve the sample complexity. Therefore, our main goal in this section is to reduce the expected sample complexity while keeping the term $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\nabla_\theta \mathcal{J}(\theta^t)-g^t\right\|^2\right]$ small. We achieve this goal by considering the stochastic variance-reduced gradient methods that have recently attracted much attention. Among these variance-reduced methods, as argued in Gargiani et al. (2022), the ProbAbilistic Gradient Estimator (PAGE) proposed in Li et al. (2021b) has a simple and single-looped structure, and can lead to optimal convergence properties. These appealing features make it attractive in machine learning applications. Therefore, in this section, we also consider the stochastic variance-reduced proximal gradient method with PAGE for solving the problem (1). To the best of our knowledge, the application of this technique for solving the general regularized reward optimization problem in the non-oblivious setting considered in this paper is new.

For notational simplicity, for the rest of this section, we denote

$$g_w(x,\theta,\theta') = \frac{\pi_\theta(x)}{\pi_{\theta'}(x)} g(x,\theta),$$

for $\theta,\theta'\in\mathbb{R}^n,\ x\in\mathbb{R}^d$, where $\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}$ denotes the importance weight between π_θ and $\pi_{\theta'}$. Note also that

$$\mathbb{E}_{x\sim\pi_{\theta'}}\left[\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\right]=1.$$

The description of the proposed PAGE variance-reduced stochastic proximal gradient method is given in Algorithm 2.

Algorithm 2 The variance-reduced stochastic proximal gradient method with PAGE

1: **Input:** initial point θ^0, sample sizes N_1 and N_2, a probability $p\in(0,1]$, and the learning rate $\eta>0$.
2: Compute

$$g^0 := \frac{1}{N_1}\sum_{j=1}^{N_1} g(x^{0,j},\theta^0),$$

where $\{x^{0,j}\}_j$ are sampled independently according to π_{θ^0}.
3: **for** $t=0,\ldots,T-1$ **do**
4: Update

$$\theta^{t+1} = \mathrm{Prox}_{\eta\mathcal{G}}\left(\theta^t+\eta g^t\right).$$

5: Compute

$$g^{t+1} = \begin{cases} \frac{1}{N_1}\sum_{j=1}^{N_1} g(x^{t+1,j},\theta^{t+1}), & \text{with probability } p,\\ \frac{1}{N_2}\sum_{j=1}^{N_2} g(x^{t+1,j},\theta^{t+1}) - \frac{1}{N_2}\sum_{j=1}^{N_2} g_w(x^{t+1,j},\theta^t,\theta^{t+1}) + g^t, & \text{with probability } 1-p,\end{cases}$$

where $\{x^{t+1,j}\}_j$ are sampled independently according to $\pi_{\theta^{t+1}}$.
6: **end for**
7: **Output:** $\hat{\theta}^T$ selected randomly from the generated sequence $\{\theta^t\}_{t=1}^T$.

It is clear that the only difference between Algorithm 1 and Algorithm 2 is the choice of the gradient estimator. At each iteration of the latter algorithm, we have two choices for the gradient estimator, where, with probability p, one chooses the same estimator as in Algorithm 1 with a sample size N_1, and with probability $1-p$, one constructs the estimator in a clever way which combines the information of the current iterate and the previous one. Since the data set $\{x^{t+1,1},\ldots,x^{t+1,N_2}\}$ is sampled according to the current probability distribution $\pi_{\theta^{t+1}}$, we need to rely on the importance weight between θ^t and θ^{t+1} and construct

the gradient estimator $\frac{1}{N_2}\sum_{j=1}^{N_2} g_w(x^{t+1},\theta^t,\theta^{t+1})$, which is an unbiased estimator for $\nabla_\theta \mathcal{J}(\theta^t)$, so that g^{t+1} becomes an unbiased estimator of $\nabla_\theta \mathcal{J}(\theta^{t+1})$. Indeed, one can easily verify that for any $\theta,\theta' \in \mathbb{R}^n$, it holds that

$$\mathbb{E}_{x\sim\pi_{\theta'}}\left[g_w(x,\theta,\theta')\right] = \nabla_\theta \mathcal{J}(\theta), \tag{4}$$

i.e., $g(x,\theta,\theta')$ is an unbiased estimator for $\nabla_\theta \mathcal{J}(\theta)$ provided that $x \sim \pi_{\theta'}$.

Next, we shall analyze the convergence properties of Algorithm 2. Our analysis relies on the following assumption on the importance weight, which essentially controls the change of the distributions.

Assumption 5.1. *Let $\theta,\theta' \in \mathbb{R}^n$, the importance weight between π_θ and $\pi_{\theta'}$ is well-defined and there exists a constant $C_w > 0$ such that*

$$\mathbb{E}_{x\sim\pi_{\theta'}}\left[\left(\frac{\pi_\theta(x)}{\pi_{\theta'}(x)} - 1\right)^2\right] \le C_w^2.$$

Note that Assumption 5.1 is also employed in many existing works Papini et al. (2018); Xu et al. (2019); Pham et al. (2020); Yuan et al. (2022); Gargiani et al. (2022). However, this assumption could be too strong, and it is not checkable in general. It is out of scope of this paper on how to relax this assumption. Some related works on MDPs include Zhang et al. (2021a); Salehkaleybar et al. (2022).

The bounded variance of the importance weight implies that the (expected) distance between $g(x,\theta')$ and $g_w(x,\theta,\theta')$ is controlled by the distance between θ and θ', for any given $\theta,\theta' \in \mathbb{R}^d$. In particular, we have the following lemma, whose proof is provided in Appendix A.5.

Lemma 5.2. *Under Assumption 3.1, Assumption 3.2 and Assumption 5.1, then it holds that*

$$\mathbb{E}_{x\sim\pi_{\theta'}}\left[\left\|g(x,\theta') - g_w(x,\theta,\theta')\right\|^2\right] \le C\left\|\theta-\theta'\right\|^2,$$

where $C > 0$ is a constant defined as

$$C := 6U^2C_h^2 + 6C_g^2\widetilde{C}_g^2 + 6\widetilde{C}_h^2 + \left(4U^2C_g^2 + 4\widetilde{C}_g^2\right)(2C_g^2 + C_h)(C_w^2+1).$$

Under the considered assumptions, we are able to provide an estimate for the term $\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta\mathcal{J}(\theta^t)\right\|^2\right]$, which plays an essential role in deriving an improved sample complexity of Algorithm 2. The results are summarized in the following Lemma 5.3; see Appendix A.6 for a proof.

Lemma 5.3. *Suppose that Assumption 3.1, Assumption 3.2, and Assumption 5.1 hold. Let $\{g^t\}$ and $\{\theta^t\}$ be the sequences generated by Algorithm 2, then it holds that*

$$\left(1 - \frac{(1-p)C\eta}{pN_2L(1-2\eta L)}\right)\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta\mathcal{J}(\theta^t)\right\|^2\right] \le \frac{p\sigma^2T + \sigma^2}{pN_1} + \frac{2\eta(1-p)C\Delta}{pN_2(1-2\eta L)}.$$

We are now ready to present the main result on the convergence property of the Algorithm 2 by showing how to select the sample sizes N_1 and N_2, probability p, and the learning rate η. Intuitively, N_1 is typically a large number and one does not want to perform samplings with N_1 samples frequently, thus the probability p and the sample size N_2 should both be small. Given N_1, N_2 and p, we can then determine the value of η such that $\eta < \frac{1}{2L}$. Consequently, the key estimate in Theorem 4.3 can be applied directly. Our results are summarized in the following theorem. Reader is referred to Appendix A.7 for the proof of this result.

Theorem 5.4. *Suppose that Assumption 3.1, Assumption 3.2 and Assumption 5.1 hold. For a given $\epsilon \in (0,1)$, we set $p := \frac{N_2}{N_1+N_2}$ with $N_1 := O(\epsilon^{-2})$ and $N_2 := \sqrt{N_1} = O(\epsilon^{-1})$. Choose a learning rate η satisfying $\eta \in \left(0, L/(2C+2L^2)\right]$. Then, running Algorithm 2 for $T := O(\epsilon^{-2})$ iterations outputs a point $\hat{\theta}^T$ satisfying*

$$\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta\mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \le \epsilon^2.$$

Moreover, the total expected sample complexity is $O(\epsilon^{-3})$.

By using the stochastic variance-reduce gradient estimator with PAGE and the importance sampling technique, we have improved the total sample complexity from $O(\epsilon^{-4})$ to $O(\epsilon^{-3})$, under the considered conditions. This result matches with the current competitive results established in Xu et al. (2019); Yuan et al. (2020); Pham et al. (2020); Gargiani et al. (2022) for solving MDPs. Finally, as mentioned in Remark 4.7, by assuming or verifying stronger conditions, such as the gradient domination and its extensions, it is also possible to derive some global convergence results. Again, such a possibility is left to a future research direction.

6 Conclusions

We have studied the stochastic (variance-reduced) proximal gradient method addressing a general regularized expected reward optimization problem which covers many existing important problem in reinforcement learning. We have established the $O(\epsilon^{-4})$ sample complexity of the classical stochastic proximal gradient method and the $O(\epsilon^{-3})$ sample complexity of the stochastic variance-reduced proximal gradient method with an importance sampling based probabilistic gradient estimator. Our results match the sample complexity of their most competitive counterparts under similar settings for Markov decision processes.

Meanwhile, we have also suspected some limitations in the current paper. First, due to the nonconcavity of the objective function, we found it challenging to derive global convergence properties of the stochastic proximal gradient method and its variants without imposing additional conditions. Second, the bounded variance condition for the importance weight turns out to be quite strong and can not be verified in general. How to relax this condition deserves further investigation. Last but not least, since we focus more on the theoretical analysis in this paper and due to the space constraint, we did not conduct any numerical simulation to examine the practical efficiency of the proposed methods. We shall try to delve into these challenges and get better understandings of the proposed problem and algorithms in a future research.

Finally, this paper has demonstrated the possibility of pairing the stochastic proximal gradient method with efficient variance reduction techniques Li et al. (2021b) for solving the reward optimization problem (1). Beyond variance-reduced methods, there are other possibilities that allow one deriving more sophisticated algorithms. For instance, one can also pair the stochastic proximal gradient method with the ideas of the actor-critic method Konda & Tsitsiklis (1999), the natural policy gradient method Kakade (2001), policy mirror descent methods Tomar et al. (2020); Lan (2023), trust-region methods Schulman et al. (2015); Shani et al. (2020), and the variational policy gradient methods Zhang et al. (2020a). We think that these possible generalizations can lead to more exciting results and make further contributions to the literature.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.

Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. *arXiv preprint arXiv:2306.01854*, 2023.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *journal of artificial intelligence research*, 15:319–350, 2001.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4): 1205–1223, 2007.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Cheng Chen, Ruitao Chen, Tianyou Li, Ruichen Ao, and Zaiwen Wen. Monte carlo policy gradient method for binary optimization. *arXiv preprint arXiv:2307.00783*, 2023.

Roy Dong, Heling Zhang, and Lillian Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 11172–11184. PMLR, 2023.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. *arXiv preprint arXiv:2302.01734*, 2023.

Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, and John Lygeros. Page-pg: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *International Conference on Machine Learning*, pp. 7223–7240. PMLR, 2022.

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

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. *arXiv preprint arXiv:2106.12112*, 2021.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pp. 9760–9785. PMLR, 2022.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Navdeep Kumar, Kaixin Wang, Kfir Levy, and Shie Mannor. Policy gradient for reinforcement learning with general utilities. *arXiv preprint arXiv:2210.00991*, 2022.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pp. 3107–3110. PMLR, 2021a.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021b.

Senwei Liang and Haizhao Yang. Finite expression method for solving high-dimensional partial differential equations. *arXiv preprint arXiv:2206.10121*, 2022.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33: 7624–7636, 2020.

Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400, 2021.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Edwin Pednault, Naoki Abe, and Bianca Zadrozny. Sequential cost-sensitive decision making with reinforcement learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 259–268, 2002.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Nhan Pham, Lam Nguyen, Dzung Phan, Phuong Ha Nguyen, Marten Dijk, and Quoc Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 374–385. PMLR, 2020.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100:255–283, 2015.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Saber Salehkaleybar, Sadegh Khorasani, Negar Kiyavash, Niao He, and Patrick Thiran. Momentum-based policy gradient with second-order information. *arXiv preprint arXiv:2205.08253*, 2022.

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

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International conference on machine learning*, pp. 5729–5738. PMLR, 2019.

Zezheng Song, Maria K Cameron, and Haizhao Yang. A finite expression method for solving high-dimensional committor problems. *arXiv preprint arXiv:2306.12268*, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Lin Xiao. On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922, 2022.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10460–10468, 2021.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pp. 541–551. PMLR, 2020.

Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8823–8831, 2022.

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

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 3332–3380. PMLR, 2022.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020a.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10887–10895, 2021b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116, 2004.

Wei Zhang and Thomas G Dietterich. A reinforcement learning approach to job-shop scheduling. In *IJCAI*, volume 95, pp. 1114–1120. Citeseer, 1995.

A Proofs

A.1 Proof of Lemma 4.1

Proof of Lemma 4.1. One could establish the L-smoothness of $\mathcal{J}(\cdot)$ via bounding the spectral norm of the Hessian $\nabla_\theta^2 \mathcal{J}(\cdot)$. To this end, we first calculate the Hessian of $\mathcal{J}$ as follows:

$$\begin{aligned}
\nabla_\theta^2 \mathcal{J}(\theta) &= \nabla_\theta \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right] \\
&= \nabla_\theta \int \left(\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)\pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\pi_\theta(x)\right) \mathrm{d}x \\
&= \int \mathcal{R}_\theta(x)\pi_\theta(x)\left(\nabla_\theta^2 \log \pi_\theta(x) + \nabla_\theta \log \pi_\theta(x)\nabla_\theta \log \pi_\theta(x)^\top\right) \mathrm{d}x \\
&\quad + \int \nabla_\theta^2 \mathcal{R}_\theta(x)\pi_\theta(x) + 2\nabla_\theta \mathcal{R}_\theta(x)\nabla_\theta \pi_\theta(x)^\top \mathrm{d}x \\
&= \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\nabla_\theta^2 \log \pi_\theta(x)\right] + \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)\nabla_\theta \log \pi_\theta(x)^\top\right] \\
&\quad + \mathbb{E}_{x\sim\pi_\theta}\left[\nabla_\theta^2 \mathcal{R}_\theta(x)\right] + 2\mathbb{E}_{x\sim\pi_\theta}\left[\nabla_\theta \mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)^\top\right].
\end{aligned}$$

Then, by the triangular inequality, it holds that

$$\begin{aligned}
\left\|\nabla_\theta^2 \mathcal{J}(\theta)\right\|_2 \leq & \sup_{x\in\mathbb{R}^d,\, \theta\in\mathbb{R}^n} \left\|\mathcal{R}_\theta(x)\nabla_\theta^2 \log \pi_\theta(x)\right\|_2 + \sup_{x\in\mathbb{R}^d,\, \theta\in\mathbb{R}^n} \left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)\nabla_\theta \log \pi_\theta(x)^\top\right\|_2 \\
& + \sup_{x\in\mathbb{R}^d,\, \theta\in\mathbb{R}^n} \left\|\nabla_\theta^2 \mathcal{R}_\theta(x)\right\|_2 + 2\sup_{x\in\mathbb{R}^d,\, \theta\in\mathbb{R}^n} \left\|\nabla_\theta \mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)^\top\right\|_2 \\
\leq & \, U(C_g^2 + C_h) + \widetilde{C}_h + 2C_g\widetilde{C}_g.
\end{aligned}$$

Thus, $\mathcal{J}$ is L-smooth with $L := U(C_g^2 + C_h) + \widetilde{C}_h + 2C_g\widetilde{C}_g$, and the proof is completed. □

A.2 Proof of Theorem 4.3

Proof of Theorem 4.3. From Lemma 4.1, we see that

$$\mathcal{J}(\theta^{t+1}) \geq \mathcal{J}(\theta^t) + \left\langle \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle - \frac{L}{2}\left\|\theta^{t+1} - \theta^t\right\|^2. \tag{5}$$

By the updating rule of θ^{t+1}, we see that

$$-\left\langle g^t, \theta^{t+1} - \theta^t \right\rangle + \frac{1}{2\eta}\left\|\theta^{t+1} - \theta^t\right\|^2 + \mathcal{G}(\theta^{t+1}) \leq \mathcal{G}(\theta^t), \tag{6}$$

$$g^t - \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right) \in \partial\mathcal{G}(\theta^{t+1}). \tag{7}$$

Combining (5) and (6), we see that

$$\begin{aligned}&\mathcal{J}(\theta^{t+1}) + \left\langle g^t, \theta^{t+1} - \theta^t \right\rangle - \frac{1}{2\eta}\left\|\theta^{t+1} - \theta^t\right\|^2 - \mathcal{G}(\theta^{t+1})\\ \geq\ &\mathcal{J}(\theta^t) + \left\langle \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle - \frac{L}{2}\left\|\theta^{t+1} - \theta^t\right\|^2 - \mathcal{G}(\theta^t).\end{aligned}$$

Rearranging terms, we can rewrite the above inequality as

$$\frac{1-\eta L}{2\eta}\left\|\theta^{t+1} - \theta^t\right\|^2 \leq \mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t) + \left\langle g^t - \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle. \tag{8}$$

By the Cauchy-Schwarz inequality, we see that

$$\left\langle g^t - \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle \leq \frac{1}{2L}\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2 + \frac{L}{2}\left\|\theta^{t+1} - \theta^t\right\|^2,$$

which together with (8) implies that

$$\frac{1-2\eta L}{2\eta}\left\|\theta^{t+1} - \theta^t\right\|^2 \leq \mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t) + \frac{1}{2L}\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2.$$

Summing the above inequality across $t = 0, \ldots, T-1$, we get

$$\begin{aligned}\frac{1-2\eta L}{2\eta}\sum_{t=0}^{T-1}\left\|\theta^{t+1} - \theta^t\right\|^2 &\leq \mathcal{F}(\theta^T) - \mathcal{F}(\theta^0) + \frac{1}{2L}\sum_{t=0}^{T-1}\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\\ &\leq \Delta + \frac{1}{2L}\sum_{t=0}^{T-1}\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2.\end{aligned} \tag{9}$$

Here, we recall that $\Delta := \mathcal{F}^* - \mathcal{F}(\theta^0) > 0$.

On the other hand, (8) also implies that

$$\begin{aligned}&2\left\langle \nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t, \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\rangle + \frac{1-\eta L}{\eta^2}\left\|\theta^{t+1} - \theta^t\right\|^2\\ \leq\ &\frac{2}{\eta}\left(\mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t)\right) + \frac{2}{\eta}\left\langle \nabla_\theta \mathcal{J}(\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle.\end{aligned} \tag{10}$$

Notice that

$$\begin{aligned}&2\left\langle \nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t, \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\rangle\\ =\ &\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t + \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\|^2 - \left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t\right\|^2 - \frac{1}{\eta^2}\left\|\theta^{t+1} - \theta^t\right\|^2.\end{aligned}$$

Then by substituting the above equality into (10) and rearranging terms, we see that

$$\begin{aligned}&\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t + \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\|^2\\ \leq\ &\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t\right\|^2 + \frac{1}{\eta^2}\left\|\theta^{t+1} - \theta^t\right\|^2 - \frac{1-\eta L}{\eta^2}\left\|\theta^{t+1} - \theta^t\right\|^2\\ &+ \frac{2}{\eta}\left(\mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t)\right) + \frac{2}{\eta}\left\langle \nabla_\theta \mathcal{J}(\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^t), \theta^{t+1} - \theta^t \right\rangle\\ \leq\ &2\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2 + 2\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2 + \frac{L}{\eta}\left\|\theta^{t+1} - \theta^t\right\|^2\\ &+ \frac{2}{\eta}\left(\mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t)\right) + \frac{2}{\eta}\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^t)\right\|\left\|\theta^{t+1} - \theta^t\right\|\end{aligned}$$

$$\leq 2\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2 + \left(2L^2 + \frac{3L}{\eta}\right)\left\|\theta^{t+1} - \theta^t\right\|^2 + \frac{2}{\eta}\left(\mathcal{F}(\theta^{t+1}) - \mathcal{F}(\theta^t)\right),$$

where the second inequality is due to the Cauchy-Schwarz inequality and fact that

$$\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t\right\|^2 \leq 2\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2 + 2\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2,$$

and the third inequality is implied by Lemma 4.1.

Summing the above inequality across $t = 0, 1 \ldots, T-1$, we get

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t + \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\|^2 \\
&\leq 2\sum_{t=0}^{T-1}\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2 + \left(2L^2 + \frac{3L}{\eta}\right)\sum_{t=0}^{T-1}\left\|\theta^{t+1} - \theta^t\right\|^2 + \frac{2}{\eta}\left(\mathcal{F}(\theta^T) - \mathcal{F}(\theta^0)\right) \\
&\leq 2\sum_{t=0}^{T-1}\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2 + \frac{2}{\eta^2}\sum_{t=0}^{T-1}\left\|\theta^{t+1} - \theta^t\right\|^2 + \frac{2\Delta}{\eta}, \qquad (11)
\end{aligned}$$

where the last inequality is obtained from the fact that $L < \frac{1}{2\eta}$ as a consequence of the choice of the learning rate.

Consequently, we have that

$$\begin{aligned}
&\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta \mathcal{J}(\theta^{t+1}) + \partial\mathcal{G}(\theta^{t+1})\right)^2\right] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\nabla_\theta \mathcal{J}(\theta^{t+1}) - g^t + \frac{1}{\eta}\left(\theta^{t+1} - \theta^t\right)\right\|^2\right] \\
&\leq \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2\right] + \frac{2}{\eta^2 T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\theta^{t+1} - \theta^t\right\|^2\right] + \frac{2\Delta}{\eta T} \\
&\leq \frac{4}{\eta T(1-2\eta L)}\left(\Delta + \frac{1}{2L}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right]\right) + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2\right] + \frac{2\Delta}{\eta T} \\
&= \left(2 + \frac{2}{\eta L(1-2\eta L)}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\nabla_\theta \mathcal{J}(\theta^t) - g^t\right\|^2\right] + \frac{\Delta}{T}\left(\frac{2}{\eta} + \frac{4}{\eta(1-2\eta L)}\right),
\end{aligned}$$

where the first inequality is because of (7), the second inequality is due to (11) and the third inequality is derived from (9). Thus, the proof is completed. □

A.3 Proof of Lemma 4.4

Proof of Lemma 4.4. We first estimate $\mathbb{E}_{x\sim\pi_\theta}\left[\left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right\|^2\right]$ as follows

$$\begin{aligned}
\mathbb{E}_{x\sim\pi_\theta}\left[\left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right\|^2\right] &\leq 2\mathbb{E}_{x\sim\pi_\theta}\left[\left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x)\right\|^2\right] + 2\mathbb{E}_{x\sim\pi_\theta}\left[\left\|\nabla_\theta \mathcal{R}_\theta(x)\right\|^2\right] \\
&\leq 2U^2 C_g^2 + 2\widetilde{C}_g^2.
\end{aligned}$$

Then, by the fact that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \leq \mathbb{E}\left[X^2\right]$ for all random variable X, we have

$$\begin{aligned}
\mathbb{E}_{x\sim\pi_\theta}\left[\left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x) - \nabla_\theta \mathcal{J}(\theta)\right\|^2\right] &\leq \mathbb{E}_{x\sim\pi_\theta}\left[\left\|\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right\|^2\right] \\
&\leq 2U^2 C_g^2 + 2\widetilde{C}_g^2,
\end{aligned}$$

which completes the proof. □

A.4 Proof of Theorem 4.5

Proof of Theorem 4.5. From Theorem 4.3, in order to ensure that $\hat{\theta}^T$ is a ϵ-stationary point, we can require

$$\left(2+\frac{2}{\eta L(1-2\eta L)}\right)\mathbb{E}_T\left[\left\|g^t-\nabla_\theta\mathcal{J}(\theta^t)\right\|^2\right]\leq\frac{1}{2}\epsilon^2,\quad\forall\, t=0,\ldots,T-1, \tag{12}$$

$$\frac{\Delta}{T}\left(\frac{2}{\eta}+\frac{4}{\eta(1-2\eta L)}\right)\leq\frac{1}{2}\epsilon^2. \tag{13}$$

It is easy to verify that g^t is an unbiased estimator of $\nabla_\theta\mathcal{J}(\theta^t)$. Then, Lemma 4.4 implies that

$$\mathbb{E}_T\left[\left\|g^t-\nabla_\theta\mathcal{J}(\theta^t)\right\|^2\right]\leq\frac{\sigma^2}{N}.$$

As a consequence, if one chooses $N=\left\lceil\frac{\sigma^2}{\epsilon^2}\left(4+\frac{4}{\eta L(1-2\eta L)}\right)\right\rceil$, then (12) holds.

On the other hand, (13) holds if one sets $T=\left\lceil\frac{\Delta}{\epsilon^2}\left(\frac{4}{\eta}+\frac{8}{\eta(1-2\eta L)}\right)\right\rceil$. Moreover, we see that the sample complexity can be computed as $TN=O(\epsilon^{-4})$. Therefore, the proof is completed. □

A.5 Proof of Lemma 5.2

Proof of Lemma 5.2. First, recall that

$$\mathbb{E}_{x\sim\pi_{\theta'}}\left[\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\right]=1.$$

Then, by the definitions of g and g_w, we can verify that

$$\begin{aligned}
&\mathbb{E}_{x\sim\pi_{\theta'}}\left[\|g(x,\theta')-g_w(x,\theta,\theta')\|^2\right]\\
&\leq 2\mathbb{E}_{x\sim\pi_{\theta'}}\left[\|g(x,\theta')-g(x,\theta)\|^2\right]+2\mathbb{E}_{x\sim\pi_{\theta'}}\left[\|g(x,\theta)-g_w(x,\theta,\theta')\|^2\right]\\
&=2\int\|\mathcal{R}_{\theta'}(x)\nabla_\theta\log\pi_{\theta'}(x)-\mathcal{R}_\theta(x)\nabla_\theta\log\pi_\theta(x)+\nabla_\theta\mathcal{R}_{\theta'}(x)-\nabla_\theta\mathcal{R}_\theta(x)\|^2\pi_{\theta'}(x)\mathrm{d}x\\
&\quad+2\int\left\|\mathcal{R}_\theta(x)\left(\nabla_\theta\log\pi_\theta(x)-\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\nabla_\theta\log\pi_\theta(x)\right)+\nabla_\theta\mathcal{R}_\theta(x)-\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\nabla_\theta\mathcal{R}_\theta(x)\right\|^2\pi_{\theta'}(x)\mathrm{d}x\\
&\leq 6\int\|\mathcal{R}_{\theta'}(x)\left(\nabla_\theta\log\pi_{\theta'}(x)-\nabla_\theta\log\pi_\theta(x)\right)\|^2\pi_{\theta'}(x)\mathrm{d}x+6\int\|\left(\mathcal{R}_{\theta'}(x)-\mathcal{R}_\theta(x)\right)\nabla_\theta\log\pi_\theta(x)\|^2\pi_{\theta'}(x)\mathrm{d}x\\
&\quad+6\int\|\nabla_\theta\mathcal{R}_{\theta'}(x)-\nabla_\theta\mathcal{R}_\theta(x)\|^2\pi_{\theta'}(x)\mathrm{d}x+4\int\left\|\mathcal{R}_\theta(x)\left(1-\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\right)\nabla_\theta\log\pi_\theta(x)\right\|^2\pi_{\theta'}(x)\mathrm{d}x\\
&\quad+4\int\left\|\nabla_\theta\mathcal{R}_\theta(x)\left(1-\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}\right)\right\|^2\pi_{\theta'}(x)\mathrm{d}x\\
&\leq\left(6U^2C_h^2+6C_g^2\widetilde{C}_g^2+6\widetilde{C}_h^2\right)\|\theta-\theta'\|^2+\left(4U^2C_g^2+4\widetilde{C}_g^2\right)\mathbb{E}_{x\sim\pi_{\theta'}}\left[\left(\frac{\pi_\theta(x)}{\pi_{\theta'}(x)}-1\right)^2\right]\\
&=\left(6U^2C_h^2+6C_g^2\widetilde{C}_g^2+6\widetilde{C}_h^2\right)\|\theta-\theta'\|^2+\left(4U^2C_g^2+4\widetilde{C}_g^2\right)\left(\int\frac{(\pi_\theta(x))^2}{\pi_{\theta'}(x)}\mathrm{d}x-1\right).
\end{aligned}$$

We next consider the function $f(\theta):=\int\frac{(\pi_\theta(x))^2}{\pi_{\theta'}(x)}\mathrm{d}x$. Taking the derivative of f with respect to θ, we get

$$\nabla_\theta f(\theta)=\int\frac{2\pi_\theta(x)\nabla_\theta\pi_\theta(x)}{\pi_{\theta'}(x)}\mathrm{d}x.$$

Moreover, since

$$\nabla_\theta^2\log\pi_\theta(x)=\frac{1}{(\pi_\theta(x))^2}\left(\pi_\theta(x)\nabla_\theta^2\pi_\theta(x)-\nabla_\theta\pi_\theta(x)\nabla\theta\pi_\theta(x)^\top\right)$$

$$= \frac{1}{\pi_\theta(x)} \nabla_\theta^2 \pi_\theta(x) - \nabla_\theta \log \pi_\theta(x) \nabla_\theta \log \pi_\theta(x)^\top,$$

we see that the Hessian of f with respect to θ can be computed as

$$\begin{aligned}
\nabla_\theta^2 f(\theta) &= \int \frac{2}{\pi_{\theta'}(x)} \left(\nabla_\theta \pi_\theta(x) \nabla_\theta \pi_\theta(x)^\top + \pi_\theta(x) \nabla_\theta^2 \pi_\theta(x) \right) \mathrm{d}x \\
&= \int \frac{2(\pi_\theta(x))^2}{\pi_{\theta'}(x)} \left(2\nabla_\theta \log \pi_\theta(x) \nabla_\theta \log \pi_\theta(x)^\top + \nabla_\theta^2 \log \pi_\theta(x) \right) \mathrm{d}x.
\end{aligned}$$

Notice that $f(\theta') = 1$ and $\nabla_\theta f(\theta') = 0$. Therefore, by the Mean Value Theorem, we get

$$f(\theta) = 1 + \frac{1}{2} \left\langle \nabla_\theta^2 f(\tilde{\theta})(\theta - \theta'), \theta - \theta' \right\rangle,$$

where $\tilde{\theta}$ is a point between θ and θ'. Now, from the expression of the Hessian matrix, we see that for any $\theta \in \mathbb{R}^n$,

$$\begin{aligned}
\left\| \nabla_\theta^2 f(\theta) \right\|_2 &\leq \int \frac{2(\pi_\theta(x))^2}{\pi_{\theta'}(x)} \left\| 2\nabla_\theta \log \pi_\theta(x) \nabla_\theta \log \pi_\theta(x)^\top + \nabla_\theta^2 \log \pi_\theta(x) \right\|_2 \mathrm{d}x \\
&\leq 2(2C_g^2 + C_h) \int \frac{(\pi_\theta(x))^2}{\pi_{\theta'}(x)} \mathrm{d}x \\
&= 2(2C_g^2 + C_h) \left(1 + \mathbb{E}_{x \sim \pi_{\theta'}} \left[\left(\frac{\pi_\theta(x)}{\pi_{\theta'}(x)} - 1 \right)^2 \right] \right) \\
&\leq 2(2C_g^2 + C_h)(C_w^2 + 1).
\end{aligned}$$

As a consequence, we have

$$\begin{aligned}
&\mathbb{E}_{x \sim \pi_{\theta'}} \left[\left\| g(x, \theta') - g_w(x, \theta, \theta') \right\|^2 \right] \\
\leq\ & \left(6U^2 C_h^2 + 6C_g^2 \widetilde{C}_g^2 + 6\widetilde{C}_h^2 \right) \left\| \theta - \theta' \right\|^2 + \left(4U^2 C_g^2 + 4\widetilde{C}_g^2 \right) \left(\int \frac{(\pi_\theta(x))^2}{\pi_{\theta'}(x)} \mathrm{d}x - 1 \right) \\
\leq\ & \left(6U^2 C_h^2 + 6C_g^2 \widetilde{C}_g^2 + 6\widetilde{C}_h^2 + \left(4U^2 C_g^2 + 4\widetilde{C}_g^2 \right) (2C_g^2 + C_h)(C_w^2 + 1) \right) \left\| \theta - \theta' \right\|^2,
\end{aligned}$$

which completes the proof. □

A.6 Proof of Lemma 5.3

Proof of Lemma 5.3. By the definition of the stochastic gradient estimator given in Algorithm 2, we can see that for $t \geq 0$,

$$\begin{aligned}
&\mathbb{E}_{t+1} \left[\left\| g^{t+1} - \nabla_\theta \mathcal{J}(\theta^{t+1}) \right\|^2 \right] \\
&= p\mathbb{E}_{t+1} \left[\left\| \frac{1}{N_1} \sum_{j=1}^{N_1} g(x^{t+1,j}, \theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^{t+1}) \right\|^2 \right] \\
&\quad + (1-p)\mathbb{E}_{t+1} \left[\left\| \frac{1}{N_2} \sum_{j=1}^{N_2} \left(g(x^{t+1,j}, \theta^{t+1}) - g_w(x^{t+1,j}, \theta^t, \theta^{t+1}) \right) + g^t - \nabla_\theta \mathcal{J}(\theta^{t+1}) \right\|^2 \right] \\
&= p\mathbb{E}_{t+1} \left[\left\| \frac{1}{N_1} \sum_{j=1}^{N_1} g(x^{t+1,j}, \theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^{t+1}) \right\|^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&+ (1-p)\mathbb{E}_{t+1}\left[\left\|\frac{1}{N_2}\sum_{j=1}^{N_2}\left(g(x^{t+1,j},\theta^{t+1}) - g_w(x^{t+1,j},\theta^t,\theta^{t+1})\right) + \nabla_\theta \mathcal{J}(\theta^t) - \nabla_\theta \mathcal{J}(\theta^{t+1}) + g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] \\
&\leq p\mathbb{E}_{t+1}\left[\left\|\frac{1}{N_1}\sum_{j=1}^{N_1} g(x^{t+1,j},\theta^{t+1}) - \nabla_\theta \mathcal{J}(\theta^{t+1})\right\|^2\right] \\
&\quad + (1-p)\mathbb{E}_{t+1}\left[\left\|\frac{1}{N_2}\sum_{j=1}^{N_2}\left(g(x^{t+1,j},\theta^{t+1}) - g_w(x^{t+1,j},\theta^t,\theta^{t+1})\right) + g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] \\
&\leq \frac{p\sigma^2}{N_1} + (1-p)\mathbb{E}_{t+1}\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] + (1-p)\frac{1}{N_2^2}\sum_{j=1}^{N_2}\mathbb{E}_{t+1}\left[\left\|\left(g(x^{t+1,j},\theta^{t+1}) - g_w(x^{t+1,j},\theta^t,\theta^{t+1})\right)\right\|^2\right] \\
&\leq \frac{p\sigma^2}{N_1} + (1-p)\mathbb{E}_{t+1}\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] + \frac{(1-p)C}{N_2}\left\|\theta^{t+1} - \theta^t\right\|^2,
\end{aligned}
$$

where in the first inequality, we use the facts that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \leq \mathbb{E}\left[X^2\right]$ for all random variable X and g^t is unbiased estimator for $\nabla_\theta \mathcal{J}(\theta^t)$ for all $t \geq 0$, in the second inequality, we rely on the fact that $\{x^{t+1,j}\}$ is independent, and the last inequality is due to Lemma 5.2. By summing the above relation across $t = 0, \ldots, T-2$, we see that

$$
\begin{aligned}
&\sum_{t=1}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] \\
&\leq \frac{p\sigma^2(T-1)}{N_1} + (1-p)\sum_{t=0}^{T-2}\mathbb{E}_{t+1}\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] + \frac{(1-p)C}{N_2}\sum_{t=0}^{T-2}\mathbb{E}_T\left[\left\|\theta^{t+1} - \theta^t\right\|^2\right],
\end{aligned}
$$

which implies that

$$
\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] \leq \frac{p\sigma^2 T + \sigma^2}{pN_1} + \frac{(1-p)C}{pN_2}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|\theta^{t+1} - \theta^t\right\|^2\right]. \tag{14}
$$

Recall from (9) that

$$
\sum_{t=0}^{T-1}\left\|\theta^{t+1} - \theta^t\right\|^2 \leq \frac{2\eta\Delta}{1-2\eta L} + \frac{\eta}{L(1-2\eta L)}\sum_{t=0}^{T-1}\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2,
$$

which together with (14) implies that

$$
\left(1 - \frac{(1-p)C\eta}{pN_2L(1-2\eta L)}\right)\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] \leq \frac{p\sigma^2 T + \sigma^2}{pN_1} + \frac{2\eta(1-p)C\Delta}{pN_2(1-2\eta L)}.
$$

Thus, the proof is completed. □

A.7 Proof Theorem 5.4

Proof Theorem 5.4. Since $p = \frac{N_2}{N_1+N_2} \in (0,1)$ and

$$
\eta \leq \frac{pN_2L}{2(1-p)C + 2pN_2L^2} = \frac{N_2^2L}{2N_1C + 2N_2^2L^2},
$$

we can readily check that

$$
\eta \in \left(0, \frac{1}{2L}\right), \quad 1 - \frac{(1-p)C\eta}{N_2L(1-2\eta L)} \geq \frac{1}{2}. \tag{15}
$$

Then, we can see that

$$
\begin{aligned}
&\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \\
\leq\ & \left(2 + \frac{2}{\eta L(1-2\eta L)}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_T\left[\left\|g^t - \nabla_\theta \mathcal{J}(\theta^t)\right\|^2\right] + \frac{1}{T}\left(\frac{2\Delta}{\eta} + \frac{4\Delta}{\eta(1-2\eta L)}\right) \\
\leq\ & \frac{1}{T}\left(2 + \frac{2}{\eta L(1-2\eta L)}\right)\left(1 - \frac{(1-p)C\eta}{pN_2 L(1-2\eta L)}\right)^{-1}\left(\frac{p\sigma^2 T + \sigma^2}{pN_1} + \frac{2\eta(1-p)C\Delta}{pN_2(1-2\eta L)}\right) \\
&+ \frac{1}{T}\left(\frac{2\Delta}{\eta} + \frac{4\Delta}{\eta(1-2\eta L)}\right) \\
\leq\ & \frac{4}{T}\left(1 + \frac{1}{\eta L(1-2\eta L)}\right)\left(\frac{T\sigma^2}{N_1} + \frac{(N_1+N_2)\sigma^2}{N_1 N_2} + \frac{2\eta N_1 C\Delta}{N_2^2(1-2\eta L)}\right) + \frac{2\Delta}{T}\left(\frac{1}{\eta} + \frac{2}{\eta(1-2\eta L)}\right)
\end{aligned}
$$

where $\Delta := \mathcal{F}^* - \mathcal{F}(\theta^0) > 0$ is a constant, the first inequality is due to Theorem 4.3, the second inequality is derived from Lemma 5.3, and the third inequality is implied by (15).

Then, in order to have $\mathbb{E}_T\left[\operatorname{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial\mathcal{G}(\hat{\theta}^T)\right)^2\right] \leq \epsilon^2$ for a given tolerance $\epsilon > 0$, we can simply set $N_2 = \sqrt{N_1}$,

$$
\eta \leq \frac{N_2^2 L}{2N_1 C + 2N_2^2 L^2} = \frac{L}{2C + 2L^2},
$$

and require that

$$
\begin{aligned}
4\left(1 + \frac{1}{\eta L(1-2\eta L)}\right)\frac{\sigma^2}{N_1} &\leq \frac{\epsilon^2}{3}, \\
\frac{4}{T}\left(1 + \frac{1}{\eta L(1-2\eta L)}\right)\frac{(N_1+N_2)\sigma^2}{N_1 N_2} &\leq \frac{\epsilon^2}{3}, \\
\frac{2\Delta}{T}\left[\left(1 + \frac{1}{\eta L(1-2\eta L)}\right)\frac{4\eta N_1 C}{N_2^2(1-2\eta L)} + \frac{1}{\eta} + \frac{2}{\eta(1-2\eta L)}\right] &\leq \frac{\epsilon^2}{3}.
\end{aligned}
$$

Therefore, it suffices to set $N_1 = O(\epsilon^{-2})$, $N_2 = \sqrt{N_1} = O(\epsilon^{-1})$ and $T = O(\epsilon^{-2})$. (We ignore deriving the concrete expressions of T, N_1 and N_2, in terms of ϵ and other constants, but only give the big-O notation here for simplicity.)

Finally, we can verify that the sample complexity can be bounded as

$$
N_1 + T\left(pN_1 + (1-p)N_2\right) = N_1 + T\frac{2N_1 N_2}{N_1 + N_2} \leq N_1 + 2TN_2 = O(\epsilon^{-3}).
$$

Therefore, the proof is completed. □