SIAM J. CONTROL OPTIM.
Vol. 58, No. 6, pp. 3586–3612

GLOBAL CONVERGENCE OF POLICY GRADIENT METHODS TO (ALMOST) LOCALLY OPTIMAL POLICIES*

KAIQING ZHANG†, ALEC KOPPEL‡, HAO ZHU§, AND TAMER BAŞAR†

Abstract. Policy gradient (PG) methods have been one of the most essential ingredients of reinforcement learning, with application in a variety of domains. In spite of the empirical success, a rigorous understanding of the *global convergence* of PG methods appears to be relatively lacking in the literature, especially for the infinite-horizon setting with discounted factors. In this work, we close the gap by viewing PG methods from a nonconvex optimization perspective. In particular, we propose a new variant of PG methods for infinite-horizon problems that uses a random rollout horizon for the Monte Carlo estimation of the policy gradient. This method then yields an unbiased estimate of the policy gradient with bounded variance, which enables using the tools from nonconvex optimization to establish the global convergence. Employing this perspective, we first point to an alternative method to recover the convergence to *stationary-point policies* in the literature. Motivated by the recent advances in nonconvex optimization, we have modified the proposed PG method by introducing a periodically enlarged stepsize rule. More interestingly, this modified algorithm is shown to be able to escape saddle points under mild assumptions on the reward functions and the policy parameterization of the reinforcement learning (RL) problem. Specifically, we connect the correlated negative curvature condition of [H. Daneshmand et al., *Escaping saddles with stochastic gradients*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 1155–1164] to the fact that the reward must be strictly positive or negative. Under the additional assumption that all saddle points are strict, this result essentially establishes the convergence to actual *locally optimal policies* of the underlying problem and thus rigorously corroborates the overclaimed argument in the literature on the convergence of PG methods. In this aspect, our findings justify the benefit of reward-reshaping in terms of escaping saddle points from a nonconvex optimization perspective.

Key words. reinforcement learning, policy gradient methods, nonconvex optimization, global convergence

AMS subject classifications. 93E03, 93E35, 93E20, 90-08, 65K05

DOI. 10.1137/19M1288012

1. Introduction.

In reinforcement learning (RL) [6, 47], an agent moves through a state space and seeks to learn a policy which maps states to a probability distribution over actions to maximize a long-term accumulation of rewards. When the agent selects an action at a particular state, a reward is revealed and transitions to a new state according to a probability density that only depends on the current state and action. Under this setting, i.e., a Markov decision process (MDP), the agent must evaluate the merit of different actions by interacting with the environment. Two dominant approaches to RL have emerged: those based on optimizing the accumulated

*Received by the editors September 17, 2019; accepted for publication (in revised form) June 23, 2020; published electronically December 3, 2020.
https://doi.org/10.1137/19M1288012
Funding: The first and fourth authors' research was supported in part by the U.S. Army Research Office (ARO) grant W911NF-16-1-0485, in part by the U.S. Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by Office of Naval Research (ONR) MURI grant N00014-16-1-2710. The second author's research was supported by ASEE SMART Scholarship for Service. The third author's research was supported by NSF-1802319.

†Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (kzhang66@illinois.edu, basar1@illinois.edu).

‡Computational and Information Sciences Directorate, U.S. Army Research Lab, Adelphi, MD 20783 USA (aekoppel314@gmail.com).

§Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (haozhu@utexas.edu).

reward directly from the policy space, referred to as "direct policy search," and those based on finding the value function by solving the Bellman fixed point equations [5]. Our goal is to rigorously understand the former approach, namely, policy gradient (PG) methods [48], from an optimization angle. Policy search has attracted growing attention recently, with its theory being studied from a statistical aspect [60, 61]. This popularity is mainly attributed to its ability to scale gracefully to large and even continuous spaces [44, 46], to multiagent settings [21, 32, 56, 58], and to incorporate deep networks as function approximators [30, 33].

Despite the increasing prevalence of PG methods, their global convergence in the infinite-horizon discounted setting, which is conventional in dynamic programming [6], is not yet well understood. This gap stems first from the fact that obtaining unbiased estimates of the policy gradient through sampling is often elusive. Specifically, following the Policy Gradient Theorem [48], obtaining an unbiased estimate of the policy gradient requires two significant conditions to hold: (i) the state-action pair is drawn from the discounted state-action occupancy measure of the Markov chain under the policy; (ii) the estimate of the action-value (or Q) function induced by the policy is unbiased. This gap also results from the fact that the value function to be maximized in RL is in general *nonconvex* with respect to the policy parameter [1, 7, 20, 31, 57]. In the same vein as our work, there is a surging interest in studying the global convergence of PG methods; see the recent work [12, 20, 57] and concurrent work [1, 7, 31]. However, it is worth mentioning that, orthogonal to our work, these works considered convergence to the *global optimum* in several *special* RL settings, e.g., the linear quadratic setting [12, 20], the tabular setting [1, 7], and the setting with function approximation but where the approximation error can be quantified [1]. In contrast, our focus is on a broader RL regime where the nonconvexity is general, and the approximation error cannot be quantified.

When one restricts the focus to *episodic* RL, Monte Carlo rollout may be used to obtain *unbiased* estimates of the Q-function. In particular, the rollout simulates the MDP under a certain policy to a finite horizon and then collects information along the trajectory. However, this finite-horizon rollout, though generally used in practice, is known to introduce bias in estimating an *infinite-horizon* discounted value function. Such a bias in estimating the policy gradient for infinite-horizon problems has been identified in the earlier works [3, 4] both analytically and empirically. To address this bias issue, we employ in this work random geometric time rollout horizons, a technique first proposed in [38]. Doing so allows us to obtain unbiased estimates of the Q-function using only rollouts of finite horizons. Moreover, the random rollout horizon also creates an unbiased sampling of the state-action pair from the discounted occupancy measure [48]. With these two challenges addressed, the policy gradient can be estimated unbiasedly. Consequently, the PG methods can be more naturally connected to the classical stochastic programming algorithms [45], where the unbiasedness of the stochastic search direction is critical. We refer to our algorithm as *random-horizon* policy gradient (RPG) to emphasize that the horizon of the Monte Carlo rollout is random.

Leveraging this connection, we are able to understand the effect of the policy parameterization on both the limiting and finite-iteration behavior. In particular, it is well known in nonconvex optimization that with only first-order information, convergence to a stationary point with zero gradient-norm is the best one may hope to achieve [50], except under some structured settings [17, 24, 62, 63, 64]. Indeed, current guarantees of PG methods pertain to convergence to the stationary points only, as pointed out in [1]. However, in some asymptotic analyses for PG methods

with function approximation [42], or their variant, actor-critic algorithms [8, 9, 10, 15], the limit points of the algorithms starting from any initialization supposedly constitute the *locally optimal policies*; i.e., the algorithms enjoy global convergence to the local optima. However, by the theory of stochastic approximation [11], such a claim can only be made locally; i.e., local optimality can only be obtained if the algorithm starts around a local minimum, under the assumption that a *strict* Lyapunov function exists. Therefore, *global* convergence of PG methods to the *actual locally optimal* policies, though claimed in some literature, remains open. Another line of theoretical studies of PG methods only focuses on showing the one-step policy improvement [37, 41, 42] by choosing appropriate stepsizes and/or batch data sizes. Such one-step result still does not imply any global convergence result. In summary, the misuse of the term *locally optimal policy* and the lack of studying global convergence of PG methods motivate us to further investigate this problem from a nonconvex optimization perspective. Thanks to the analytical tools from optimization, we are able to first recover the asymptotic convergence and then provide the convergence rate to *stationary-point* policies.

Encouraged by this connection between nonconvex optimization and policy search, we then tackle a related question: what implications do recent algorithms that can escape saddle points for nonconvex problems [16, 25] have on policy gradient methods in RL? To answer this question, we identify several structural properties of RL problems that can be exploited to mitigate the underlying nonconvexity which rely on some key assumptions on the policy parameterization and reward. Specifically, the reward needs to be bounded and either strictly positive or negative. In addition, the policy parameterization needs to be *regular*; i.e., its Fisher information matrix is positive definite (a conventional assumption in RL [27]). Under these mild conditions, we establish that PG methods can escape saddle points and converge to approximate *second-order* stationary points (SOSPs) with high probability, when a *periodically enlarged stepsize* strategy is employed. We refer to the resulting method as modified RPG (MRPG). The strict positivity/negativity of the reward function may amplify the variance of the gradient estimate, compared to the setting that has reward values with both signs but of smaller magnitude. This increased variance can be alleviated by introducing a *baseline* in the gradient estimate, as advocated by [9, 23, 40]. Therefore, we propose two further modified updates that include the baselines, both shown to converge to SOSPs as well.

Main contribution. Our main contribution is threefold. (i) We propose a series of RPG methods that unbiasedly estimate the true policy gradient for *infinite-horizon discounted* MDPs (section 3). These proposed methods facilitate the use of analytical tools from nonconvex optimization to establish their convergence to *stationary-point* policies (spotlighted in a companion report [54]). (ii) By virtue of such a connection of PG methods and nonconvex optimization, we propose modified RPG methods using periodically enlarged stepsizes, with guaranteed convergence to actual *locally optimal policies* under mild conditions on the reward functions and parameterization of the policies (section 4). (iii) We connect the condition on the reward function to the reward-reshaping technique advocated in empirical RL studies, while justifying its benefit from a nonconvex optimization perspective. The proof technique underlying this result (Theorem 4.8) is a unique contribution of this work, and hinges on reworking and fixing the flaws in the analysis of [16]. Overall, we believe this perspective opens the door to exploiting more advances in nonconvex optimization to improve the convergence of PG methods in RL.

Notation. We denote the probability measure over the space $\mathcal{S}$ by $\mathcal{P}(\mathcal{S})$, and the

set of integers $\{1,\dots,N\}$ by $[N]$. We use $\mathbb{R}$ to denote the set of real numbers, and $\mathbb{E}$ to denote the expectation operator. We let $\|\cdot\|$ denote the 2-norm of a vector in $\mathbb{R}^d$, or the spectral norm of a matrix in $\mathbb{R}^{d\times d}$. We use $|\mathcal{A}|$ to denote the cardinality of a finite set $\mathcal{A}$, or the area of a region $\mathcal{A}$, i.e., $|\mathcal{A}| = \int_{\mathcal{A}} da$. For any symmetric matrix $A \in \mathbb{R}^{d\times d}$, we use $A \succ 0$ and $A \succeq 0$ to denote that A is positive definite and positive semidefinite, respectively. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote, respectively, the smallest and largest eigenvalues of some square symmetric matrix A. We use $\mathbb{E}_X$ or $\mathbb{E}_{X\sim f(x)}$ to denote the expectation with respect to random variable X. Unless otherwise specified, we use $\mathbb{E}$ to denote the full expectation with respect to all random variables.

2. Problem formulation. RL is generally studied upon the Markov decision process (MDP) model, which is characterized by a tuple $(\mathcal{S},\mathcal{A},\mathbb{P},R,\gamma)$. $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces of the agent. $\mathbb{P}(s' \,|\, s,a) : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the Markov kernel that determines the transition probability from (s,a) to state s'. $\gamma \in (0,1)$ is the discount factor. $R : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the reward function of s and a.

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

$$V_\pi(s) = \mathbb{E}_{a_t\sim\pi(\cdot\,|\,s_t),s_{t+1}\sim\mathbb{P}(\cdot\,|\,s_t,a_t)}\bigg(\sum_{t=0}^{\infty}\gamma^t r_t \,\bigg|\, s_0 = s\bigg),$$

which quantifies the long-term expected accumulation of rewards discounted by γ. We can further define the value $Q_\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ conditioned on a given initial state-action pair as the action-value, or Q-function, as $Q_\pi(s,a) = \mathbb{E}\big(\sum_{t=0}^{\infty}\gamma^t r_t \,\big|\, s_0 = s, a_0 = a\big)$. We also define $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$ for any s,a to be the *advantage function*. Given any initial state s_0, the goal is to find the optimal policy π that maximizes the long-term return $V_\pi(s_0)$, i.e., to solve the optimization problem $\max_{\pi\in\Pi} V_\pi(s_0)$. Particularly, an RL agent seeks to solve the problem when the model on the transition probability $\mathbb{P}$ and the reward function R is unknown to the agent. Note that Π is the function class of π that can have infinite dimensions. In this work, we investigate *policy-search* methods to solve this optimization problem. To make the policy search tractable, one can *parameterize* the policies π in Π by a vector $\theta \in \mathbb{R}^d$ for some integer $d > 0$, i.e., $\pi = \pi_\theta$, which yields RL algorithms called *policy gradient* (PG) *methods* [10, 13, 29]. For notational convenience, we define $J(\theta) := V_{\pi_\theta}(s_0)$; then the vector-valued optimization problem can be written as

$$\max_{\theta\in\mathbb{R}^d} \quad J(\theta). \tag{2.1}$$

Generally, the value function is nonconvex with respect to the parameter θ, meaning that obtaining a globally optimal solution to (2.1) is NP-hard, except in the case of several special RL settings that have been identified very recently [1, 7, 20]. In fact, the conventional limit point of most approaches to nonconvex optimization is a stationary-point solution, which could be either a saddle point or a local optimum. Usually the local optima achieve reasonably good performance, in some cases comparable to the global optima, whereas saddle points are undesirable and can stall training procedures. Therefore, it is beneficial to design methods that may escape saddle points—see recent efforts on escaping saddle points in nonconvex optimization with both first-order [16, 22, 25] and second-order methods [18, 51].

Our goal in this paper is to develop PG methods to maximize $J(\theta)$ and to rigorously understand the interplay between its limiting properties and the necessity of augmenting the algorithmic update, reward function, and policy parameterization, all toward escaping undesirable limit points. This issue was first observed and addressed in [28] by adding random perturbations in the actor-critic-type RL algorithms, based on the *asymptotic* convergence results in [39]. Here we provide a modern perspective via the latest developments in nonconvex optimization.

3. Policy gradient methods. In this section, we connect stochastic gradient ascent, as it is called in stochastic optimization, with the PG methods, a kind of direct policy search in reinforcement learning. We start with the following standard assumption on the regularity of the MDP and the parameterized policy π_θ.

ASSUMPTION 3.1. *Suppose the reward function R and the parameterized policy π_θ satisfy the following conditions:*

(i) *The absolute value of the reward R is uniformly bounded, say by U_R, i.e., $|R(s,a)| \in [0, U_R]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

(ii) *The policy π_θ is differentiable with respect to θ, and $\nabla \log \pi_\theta(a \,|\, s)$, known as the* score function *corresponding to the distribution $\pi_\theta(\cdot \,|\, s)$, exists. Moreover, it is L_Θ-Lipschitz and has bounded norm, i.e., for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\nabla \log \pi_{\theta^1}(a \,|\, s) - \nabla \log \pi_{\theta^2}(a \,|\, s)\| \leq L_\Theta \cdot \|\theta^1 - \theta^2\| \quad \textit{for any} \quad \theta^1, \theta^2, \tag{3.1}$$

$$\|\nabla \log \pi_\theta(a \,|\, s)\| \leq B_\Theta \quad \textit{for some constant } B_\Theta \quad \textit{for any} \quad \theta. \tag{3.2}$$

Note that the boundedness of the reward function in Assumption 3.1(i) is standard in the literature of policy gradient/actor-critic algorithms [9, 10, 13, 55, 58]. By definition, the uniform boundedness of R also implies that the absolute value of the Q-function is upper-bounded by $U_R/(1-\gamma)$. The same bound also applies to $V_{\pi_\theta}(s)$ for any π_θ and $s \in \mathcal{S}$, and thus to the objective $J(\theta)$, which is defined as $V_{\pi_\theta}(s_0)$

In addition, the conditions (3.1) and (3.2) have also been adopted in several recent works on the convergence analysis of policy gradient algorithms [13, 14, 36, 42]. Both conditions can be readily satisfied by common parameterized policies such as the Gibbs policy [28, 48] and the Gaussian policy [19, 36, 41]. For example, for the Gaussian policy[1] in continuous spaces, $\pi_\theta(\cdot \,|\, s)$ is parameterized as $\mathcal{N}(\phi(s)^\top \theta, \sigma^2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2, and $\phi(s)$ is the feature vector that incorporates some domain knowledge, such that $\phi(s)^\top \theta$ estimates the desired (continuous) action to take at state s. The variance dictated by σ^2 enables certain exploration around the mean action. One general option without relying on particular domain knowledge is to choose the feature $\phi(s)$ such that $\phi(s)^\top \theta$ is a linear combination of several radial basis functions [19]. Then the score function has the form $[a - \phi(s)^\top \theta]\phi(s)/\sigma^2$, which satisfies (3.1) and (3.2) if the norm of the feature $\|\phi(s)\|$ is bounded; the parameter θ lies in some bounded set; and the actions $a \in \mathcal{A}$ are bounded. Moreover, for discrete-action spaces, the Gibbs policies $\pi(a \,|\, s) \propto \exp(\theta^\top \phi_{s,a})$ for some feature vector $\phi_{s,a}$, i.e., $\nabla \log \pi(a \,|\, s) \propto \phi_{s,a}$. As a consequence, the boundedness in (3.2) holds as long as the feature vector $\phi_{s,a}$ is bounded.

Under Assumption 3.1, the gradient of $J(\theta)$ with respect to the policy parameter

[1] In practice, the action space $\mathcal{A}$ is bounded; thus a truncated Gaussian policy over $\mathcal{A}$ is often used; see [36].

Algorithm 3.1 EstQ: Unbiasedly Estimating Q-function

Input: s, a, and θ. Initialize $\hat{Q} \leftarrow 0$, $s_0 \leftarrow s$, and $a_0 \leftarrow a$.
Draw $T \sim \text{Geom}(1-\gamma^{1/2})$, i.e., $P(T=t) = (1-\gamma^{1/2})\gamma^{t/2}$.
for all $t = 0, \ldots, T-1$ **do**
  Collect and add the instantaneous reward $R(s_t, a_t)$ to $\hat{Q}$,

$$\hat{Q} \leftarrow \hat{Q} + \gamma^{t/2} \cdot R(s_t, a_t).$$

  Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \,|\, s_t, a_t)$ and action $a_{t+1} \sim \pi(\cdot \,|\, s_{t+1})$.
end for
Collect reward $R(s_T, a_T)$ by $\hat{Q} \leftarrow \hat{Q} + \gamma^{T/2} \cdot R(s_T, a_T)$.
return $\hat{Q}$.

θ, given by the Policy Gradient Theorem [48], has the following form:[2]

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\big[\nabla \log \pi_\theta(a \,|\, s) \cdot Q_{\pi_\theta}(s,a)\big] \tag{3.3}$$

$$= \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\big\{\nabla \log \pi_\theta(a \,|\, s) \cdot A_{\pi_\theta}(s,a)\big\}. \tag{3.4}$$

Here, we denote by $\rho_\theta(s,a) = \rho_{\pi_\theta}(s) \cdot \pi_\theta(a \,|\, s)$ the *discounted state-action occupancy measure*, where $\rho_{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(s_t = s \,|\, s_0, \pi_\theta)$ is a valid probability measure over the state $\mathcal{S}$ [48], known as the *discounted state-occupancy measure*, and $p(s_t = s \,|\, s_0, \pi_\theta)$ is the probability that state $s_t = s$ given initial state s_0 and policy π_θ. Also, recall that the advantage function in (3.4) is defined as $A_{\pi_\theta}(s,a) = Q_{\pi_\theta}(s,a) - V_{\pi_\theta}(s)$, with $V_{\pi_\theta}(s)$ usually referred to as the *baseline*. Derivation of (3.4) from (3.3) uses the fact that

$$\mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\big[\nabla \log \pi_\theta(a \,|\, s) \cdot V_{\pi_\theta}(s)\big] = \int_{\mathcal{S}\times\mathcal{A}} \rho_{\pi_\theta}(s) \cdot \nabla\pi_\theta(a \,|\, s) \cdot V_{\pi_\theta}(s) ds da$$
$$= \int_{\mathcal{S}} \rho_{\pi_\theta}(s) \cdot V_{\pi_\theta}(s) \cdot \Big(\nabla \int_{\mathcal{A}} \pi_\theta(a \,|\, s) da\Big) ds = \int_{\mathcal{S}} \rho_{\pi_\theta}(s) \cdot V_{\pi_\theta}(s) \cdot \big(\nabla 1\big) ds = 0.$$

Next, we discuss how (3.3) and (3.4) can be used to develop first-order stochastic algorithms to address (2.1). Unbiased samples of the gradient $\nabla J(\theta)$ are required to perform the stochastic gradient ascent, which hopefully converges to a stationary-point solution of the nonconvex optimization problem. Moreover, through the addition of carefully designed perturbations, we aim to attain a local optimum.

Sampling the policy gradient. To obtain an *unbiased* sample of $\nabla J(\theta)$, it is necessary to (i) draw state-action pair (s,a) from the distribution $\rho_\theta(\cdot,\cdot)$ and (ii) obtain an unbiased estimate of the Q-function $Q_{\pi_\theta}(s,a)$, or the advantage function $A_{\pi_\theta}(s,a)$.

Both requirements can be satisfied by using a random horizon T that follows a certain geometric distribution in the sampling process. In particular, to ensure that condition (i) is satisfied, we use the last sample (s_T, a_T) of a finite sample trajectory $(s_0, a_0, s_1, \ldots, s_T, a_T)$ to be the sample at which $Q_{\pi_\theta}(\cdot,\cdot)$ and $\nabla \log \pi_\theta(\cdot \,|\, \cdot)$ are evaluated, where the horizon $T \sim \text{Geom}(1-\gamma)$. It can be shown that $(s_T, a_T) \sim \rho_\theta(\cdot,\cdot)$. Moreover, given (s_T, a_T), we perform Monte Carlo rollouts for another horizon $T' \sim$

[2]Here we use $\int$ to represent both summation over finite sets and integral over continuous spaces.

Geom$(1-\gamma^{1/2})$ independent of T, and estimate the Q-function value $Q_{\pi_\theta}(s,a)$ by collecting the $\gamma^{1/2}$-discounted rewards along the trajectory:

$$\hat{Q}_{\pi_\theta}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t,a_t) \,\big|\, s_0 = s, a_0 = a. \tag{3.5}$$

Algorithm 3.1 summarizes the subroutine of estimating the Q-function.

REMARK 3.2. *Thanks to the random horizon, the aforementioned sampling process yields an unbiased estimate of the Q-function in the* discounted infinite-horizon *setting, using the Monte Carlo rollouts of* finite horizons*; see Theorem* 4.3. *While in practice, usually fixed-length finite-horizon rollouts are used to approximate the infinite-horizon Q-function, e.g., in the REINFORCE* [49] *and G(PO)MDP* [4] *algorithms, which causes bias in the Q-function estimate, and hence the policy gradient estimate. Note that the proposed sampling technique improves the one in* [38] *that uses Geom*$(1-\gamma)$ *(instead of Geom*$(1-\gamma^{1/2})$*) to generate the rollout horizon* T'*. In particular, the proposed Q-function estimate is almost surely bounded due to the* $\gamma^{1/2}$*-discount factor in* (3.5)*, which then leads to almost sure boundedness of the stochastic PG, an important assumption required in the convergence analysis to approximate second-order stationary points next.*

Motivated by (3.3), we propose the following stochastic estimate $\hat{\nabla} J(\theta)$ of $\nabla J(\theta)$:

$$\hat{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot \hat{Q}_{\pi_\theta}(s_T,a_T) \cdot \nabla \log[\pi_\theta(a_T \,|\, s_T)]. \tag{3.6}$$

In addition, we can also estimate the policy gradient using advantage functions as in (3.4), where the advantage function is estimated by either the difference between the value function and the action-value function, or the temporal difference (TD) error. In particular, we propose the following two stochastic policy gradients:

$$\check{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot [\hat{Q}_{\pi_\theta}(s_T,a_T) - \hat{V}_{\pi_\theta}(s_T)] \cdot \nabla \log[\pi_\theta(a_T \,|\, s_T)], \tag{3.7}$$

$$\widetilde{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot [R(s_T,a_T) + \gamma \hat{V}_{\pi_\theta}(s'_T) - \hat{V}_{\pi_\theta}(s_T)] \cdot \nabla \log[\pi_\theta(a_T \,|\, s_T)], \tag{3.8}$$

where $\hat{V}_{\pi_\theta}(s)$ is an unbiased estimate of the value function $V_{\pi_\theta}(s)$, and s'_T is the next state given state s_T and a_T. The process of estimating $\hat{V}_{\pi_\theta}(s)$ employs the same idea as the EstQ algorithm, where $\hat{V}_{\pi_\theta}(s)$ is obtained by collecting the $\gamma^{1/2}$-discounted rewards along the trajectory starting from $s_0 = s$ (instead of a state-action pair (s,a)), following $a_t \sim \pi_\theta(\cdot \,|\, s_t)$, and of length $T' \sim$ Geom$(1-\gamma^{1/2})$, i.e., $\hat{V}_{\pi_\theta}(s) = \sum_{t=0}^{T'} \gamma^{1/2} \cdot R(s_t,a_t) \,|\, s_0 = s$. We refer to this subroutine as EstV and Algorithm 3.3. The benefit of using (3.7)–(3.8) is that the baseline incurs smaller variances [23], compared to (3.6).

Random-horizon PG method. Now we shift our focus to how to use the aforementioned stochastic gradient estimates. Let k be the iteration index and θ_k be the associated estimate for θ. We then propose the following update:

$$\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla} J(\theta_k) = \theta_k + \frac{\alpha_k}{1-\gamma} \cdot \hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \cdot \nabla \log[\pi_{\theta_k}(a_{T_{k+1}} \,|\, s_{T_{k+1}})], \tag{3.9}$$

Algorithm 3.2 RPG: Random-horizon Policy Gradient Algorithm

Input: s_0 and θ_0, initialize $k \leftarrow 0$.
Repeat:
Draw T_{k+1} from the geometric distribution Geom$(1-\gamma)$.
Draw $a_0 \sim \pi_{\theta_k}(\cdot \,|\, s_0)$
for all $t = 0, \ldots, T_{k+1} - 1$ **do**
Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \,|\, s_t, a_t)$ and action $a_{t+1} \sim \pi_{\theta_k}(\cdot \,|\, s_{t+1})$.
end for
Obtain an estimate of $Q_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$ by Algorithm 3.1, i.e.,

$$\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \leftarrow \mathbf{EstQ}(s_{T_{k+1}}, a_{T_{k+1}}, \theta_k).$$

Perform policy gradient update

$$\theta_{k+1} \leftarrow \theta_k + \tfrac{\alpha_k}{1-\gamma} \cdot \hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \cdot \nabla \log[\pi_{\theta_k}(a_{T_{k+1}} \,|\, s_{T_{k+1}})].$$

Update the iteration counter $k \leftarrow k+1$.
Until Convergence

Algorithm 3.3 EstV: Unbiasedly Estimating State-Value Function

Input: s and θ. Initialize $\hat{V} \leftarrow 0$, $s_0 \leftarrow s$, and draw $a_0 \sim \pi_\theta(\cdot \,|\, s_0)$.
Draw T from the geometric distribution Geom$(1-\gamma)$.
for all $t = 0, \ldots, T-1$ **do**
Collect the instantaneous reward $R(s_t, a_t)$ and add to value $\hat{V}$,

$$\hat{V} \leftarrow \hat{V} + \gamma^{t/2} \cdot R(s_t, a_t).$$

Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \,|\, s_t, a_t)$ and action $a_{t+1} \sim \pi(\cdot \,|\, s_{t+1})$.
end for
Collect $R(s_T, a_T)$ by $\hat{V} \leftarrow \hat{V} + \gamma^{T/2} \cdot R(s_T, a_T)$.
return $\hat{V}$.

where α_k is the stepsize that is either diminishing or constant, and $\{T_k\}$ are drawn independent and identically distributed (i.i.d.) from Geom$(1-\gamma)$. The estimate $\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$ is obtained from the EstQ algorithm, i.e., the update (3.5). We refer to the algorithm as the *random-horizon policy gradient* (RPG) algorithm, as summarized in Algorithm 3.2. Note that the estimate of $\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}})$, i.e., Algorithm 3.1, is conducted in the *inner loop* of the stochastic policy gradient update.

We note that under conditions stated in the next section, it is possible to establish that (3.9) converges to a stationary-point policy almost surely, as well as derive finite-iteration complexity, from an optimization perspective. In addition, by modifications of the stepsize used in (3.9) to be introduced next, the RPG algorithm can achieve bonafide local maxima, as many existing PG algorithms claim to obtain.

Modified stepsize rules for RPG. The modified RPG (MRPG) algorithms are built upon the RPG algorithm previously discussed. These modifications can yield escape from saddle points under certain conditions and hence the convergence to approximate local extrema. To decrease the variance of PG estimates, we employ the stochastic policy gradients $\check{\nabla} J(\theta)$ and $\widetilde{\nabla} J(\theta)$ with baselines as defined in (3.7) and (3.8), respectively. Recall that the evaluations of both $\check{\nabla} J(\theta)$ and $\widetilde{\nabla} J(\theta)$ need to estimate the state-value function $\hat{V}_{\pi_\theta}(s)$, which can be achieved using the EstV

Algorithm 3.4 EvalPG: Calculating the Three Types of Stochastic Policy Gradients

Input: s, a, θ, and the gradient type $\diamondsuit$.
if gradient type $\diamondsuit = \hat{}$ **then**
Obtain an estimate $\hat{Q}_{\pi_\theta}(s,a) \leftarrow \mathbf{EstQ}(s,a,\theta)$.
Calculate $\hat{\nabla} J(\theta)$, i.e., let

$$g_\theta \leftarrow \tfrac{1}{1-\gamma} \cdot \hat{Q}_{\pi_\theta}(s,a) \cdot \nabla \log \pi_\theta(a \,|\, s).$$

else if gradient type $\diamondsuit = \check{}$ **then**
Obtain estimates $\hat{Q}_{\pi_\theta}(s,a) \leftarrow \mathbf{EstQ}(s,a,\theta)$ and $\hat{V}_{\pi_\theta}(s) \leftarrow \mathbf{EstV}(s,\theta)$.
Calculate $\check{\nabla} J(\theta)$, i.e., let

$$g_\theta \leftarrow \tfrac{1}{1-\gamma} \cdot [\hat{Q}_{\pi_\theta}(s,a) - \hat{V}_{\pi_\theta}(s)] \cdot \nabla \log \pi_\theta(a \,|\, s).$$

else if gradient type $\diamondsuit = \tilde{}$ **then**
Simulate the next state: $s' \sim \mathbb{P}(\cdot \,|\, s,a)$.
Obtain estimates $\hat{V}_{\pi_\theta}(s) \leftarrow \mathbf{EstV}(s,\theta)$ and $\hat{V}_{\pi_\theta}(s') \leftarrow \mathbf{EstV}(s',\theta)$.
Calculate $\widetilde{\nabla} J(\theta)$, i.e., let

$$g_\theta \leftarrow \tfrac{1}{1-\gamma} \cdot [R(s,a) + \gamma \cdot \hat{V}_{\pi_\theta}(s') - \hat{V}_{\pi_\theta}(s)] \cdot \nabla \log \pi_\theta(a \,|\, s).$$

end if
return Stochastic policy gradient g_θ

Algorithm 3.5 MRPG: Modified Random-horizon Policy Gradient Algorithm

Input: s_0, θ_0, and the gradient type $\diamondsuit$, initialize $k \leftarrow 0$, return set $\hat{\Theta}^* \leftarrow \emptyset$.
Repeat:
Draw $T_{k+1} \sim \mathrm{Geom}(1-\gamma)$, and draw $a_0 \sim \pi_{\theta_k}(\cdot \,|\, s_0)$.
for all $t = 0, \ldots, T_{k+1} - 1$ **do**
Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \,|\, s_t, a_t)$ and action $a_{t+1} \sim \pi_{\theta_k}(\cdot \,|\, s_{t+1})$.
end for
Calculate the stochastic gradient

$$g_k \leftarrow \mathbf{EvalPG}(s_{T_{k+1}}, a_{T_{k+1}}, \theta_k, \diamondsuit)$$

if $(k \bmod k_{\mathrm{thre}}) = 0$ **then**

$$\hat{\Theta}^* \leftarrow \hat{\Theta}^* \cup \{\theta_k\}, \qquad \theta_{k+1} \leftarrow \theta_k + \beta \cdot g_k$$

else

$$\theta_{k+1} \leftarrow \theta_k + \alpha \cdot g_k$$

end if
Update the iteration counter $k = k+1$.
Until Convergence
return θ uniformly at random from the set $\hat{\Theta}^*$.

subroutine. Due to space limitation, we defer the subroutine for calculating all three types of stochastic policy gradients, named EvalPG$(s,a,\theta,\diamondsuit)$, to Algorithm 4 in [54]. The input arguments of EvalPG refer to the PG estimate of $\nabla J(\theta)$ evaluated at (s,a), with $\diamondsuit$ being $\hat{}$, $\check{}$, and $\tilde{}$ representing the evaluation for $\hat{\nabla} J(\theta)$, $\check{\nabla} J(\theta)$, and $\widetilde{\nabla} J(\theta)$, respectively.

In particular, the proposed MRPG algorithm, i.e., outlined as Algorithm 3.5, converges to approximate second-order stationary points. Note that SOSPs coincide with local maxima under certain conditions on the problem structure. α and β in Algorithm 3.5 are the constant stepsizes with $\beta > \alpha > 0$. The idea of MRPG is to *periodically enlarge* the *constant* stepsize of the update once every k_{thre} steps. Larger stepsizes β can amplify the variance along the eigenvector corresponding to the largest eigenvalue of the Hessian, which provides a direction for the update to escape the saddle points. This idea was first introduced in [16] for general stochastic gradient methods, whose proofs were identified to be flawed by us. As part of the contribution, we improve the proofs therein to establish the convergence of our MRPG algorithm, as will be shown next.

4. Main results. We now provide main results for the convergence of modified policy gradient algorithms RPG and MRPG. We start with the definition of (approximate) second-order stationary points [35].[3]

DEFINITION 4.1. *An (ϵ_g, ϵ_h)-approximate second-order stationary point θ satisfies*

$$\|\nabla J(\theta)\| \le \epsilon_g, \qquad \lambda_{\max}[\nabla^2 J(\theta)] \le \epsilon_h.$$

If $\epsilon_g = \epsilon_h = 0$, the point θ is a second-order stationary point.

As is commonly known, the gradient is null and the Hessian is negative semidefinite at a local maximum. The definition is established as a relaxation of these two conditions. Moreover, with the further assumption that all saddle points are strict (i.e., for any saddle point θ, $\lambda_{\max}[\nabla^2 J(\theta)] > 0$) [22, 25], all SOSPs ($\epsilon_g = \epsilon_h = 0$) are local maxima. In this case, converging to (approximate) SOSPs is equivalent to converging to (approximate) local minima.

As a useful result, we first establish the Lipschitz continuity of the policy gradient $\nabla J(\theta)$, under Assumption 3.1. The proof, which by necessity is long and tedious and thus has been deferred to Appendix A.1 in [54], follows directly from the definition of Lipschitz continuity, and relies on the Lipschitz continuity and boundedness of the score function $\nabla \log \pi_\theta(a \,|\, s)$.

LEMMA 4.2 (Lipschitz-continuity of policy gradient). *Under Assumption* 3.1, *the policy gradient $\nabla J(\theta)$ is Lipschitz continuous with some $L > 0$, i.e., for any $\theta^1, \theta^2 \in \mathbb{R}^d$*

$$\|\nabla J(\theta^1) - \nabla J(\theta^2)\| \le L \cdot \|\theta^1 - \theta^2\|,$$

where the Lipschitz constant $L := U_R \cdot L_\Theta \cdot (1-\gamma)^{-2} + (1+\gamma) \cdot U_R \cdot B_\Theta^2 \cdot (1-\gamma)^{-3}$.

We then establish that all the stochastic policy gradients $\hat{\nabla} J(\theta), \check{\nabla} J(\theta)$, and $\widetilde{\nabla} J(\theta)$ are unbiased estimates of $\nabla J(\theta)$ [cf. (3.3) and (3.4)]. We can also establish the boundedness of the stochastic gradient estimates, which are used in the ensuing analysis.

THEOREM 4.3 (properties of stochastic policy gradients). *For any θ, $\hat{\nabla} J(\theta), \check{\nabla} J(\theta)$, and $\widetilde{\nabla} J(\theta)$ obtained from* (3.6), (3.7), *and* (3.8), *respectively, are all unbiased estimates of $\nabla J(\theta)$ in* (3.3), *i.e., for any θ*

$$\mathbb{E}[\hat{\nabla} J(\theta) \,|\, \theta] = \mathbb{E}[\check{\nabla} J(\theta) \,|\, \theta] = \mathbb{E}[\widetilde{\nabla} J(\theta) \,|\, \theta] = \nabla J(\theta),$$

[3]Note that Definition 4.1 is based on the maximization problem we consider here, which is slightly different from the definition for minimization problems where $\lambda_{\max}[\nabla^2 J(\theta)] \le \epsilon_h$ is replaced by $\lambda_{\min}[\nabla^2 J(\theta)] \ge -\epsilon_h$.

where the expectation is with respect to the random horizon T', the trajectory along $(s_0, a_0, s_1, \ldots, s_{T'}, a_{T'})$, and the random sample (s_T, a_T). Moreover, the norm of the policy gradient $\nabla J(\theta)$ is bounded, and its stochastic estimates $\hat{\nabla} J(\theta), \check{\nabla} J(\theta), \widetilde{\nabla} J(\theta)$ are all almost surely (a.s.) bounded, i.e., $\|\nabla J(\theta)\| \le B_\Theta U_R (1-\gamma)^{-2}$, and $\|\hat{\nabla} J(\theta)\| \le \hat{\ell}$, $\|\check{\nabla} J(\theta)\| \le \check{\ell}$, $\|\widetilde{\nabla} J(\theta)\| \le \widetilde{\ell}$ a.s., for some constants $\hat{\ell}, \check{\ell}, \widetilde{\ell} > 0$, whose values are given in (4.9)–(4.11).

Proof. We start by showing unbiasedness of the Q-estimate; i.e., for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \mathbb{R}^d$, $\mathbb{E}[\hat{Q}_{\pi_\theta}(s,a) \,|\, \theta, s, a] = Q_{\pi_\theta}(s,a)$. By definition of $\hat{Q}_{\pi_\theta}(s,a)$, we have

$$\mathbb{E}[\hat{Q}_{\pi_\theta}(s,a) \,|\, \theta, s, a] = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \mathbb{1}_{T' \ge t \ge 0} \cdot \gamma^{t/2} \cdot R(s_t, a_t) \,\bigg|\, \theta, s_0 = s, a_0 = a\bigg], \tag{4.1}$$

where we have replaced T' by ∞ by use of the indicator function $\mathbb{I}$. Now we show that the inner expectation over T' and summation in (4.1) can be interchanged. In fact, by the boundedness of the reward from Assumption 3.1, for any $N > 0$, we have

$$\mathbb{E}_{T'}\bigg(\bigg|\sum_{t=0}^{N} \mathbb{1}_{0 \le t \le T'} \cdot \gamma^{t/2} \cdot R_t\bigg|\bigg) \le U_R \cdot \mathbb{E}_{T'}\bigg(\sum_{t=0}^{N} \mathbb{1}_{0 \le t \le T'} \cdot \gamma^{t/2}\bigg). \tag{4.2}$$

Note that on the right-hand side (RHS) of (4.2), the random variable in the expectation is monotonically increasing and the limit as $N \to \infty$ exists. Thus, by the monotone convergence theorem [52], we can interchange the limit with the integral, i.e., the sum and inner expectation in (4.1), as follows:

$$\mathbb{E}\bigg\{\bigg[\sum_{t=0}^{\infty} \mathbb{1}_{T' \ge t \ge 0} \cdot \gamma^{t/2} \cdot R(s_t, a_t) \,\bigg|\, \theta, s_0 = s, a_0 = a\bigg]\bigg\} = \sum_{t=0}^{\infty} \mathbb{E}\Big[\gamma^t \cdot R(s_t, a_t) \,|\, \theta, s_0 = s, a_0 = a\Big], \tag{4.3}$$

where we have used both the facts that (i) T' is independent of the system evolution $(s_{1:T'}, a_{1:T'})$; (ii) $T' \sim \mathrm{Geom}(1-\gamma^{1/2})$ and thus $\mathbb{E}_{T'}(\mathbb{1}_{T' \ge t \ge 0}) = \mathbb{P}(T' \ge t \ge 0) = \gamma^{t/2}$. Furthermore, since $|\sum_{t=0}^{N} \gamma^t R(s_t, a_t)| \le \sum_{t=0}^{N} \gamma^t U_R$, and $\lim_{N \to \infty} \mathbb{E}(\sum_{t=0}^{N} \gamma^t U_R)$ exists, by dominated convergence theorem [2], the RHS of (4.3) can be written as

$$\sum_{t=0}^{\infty} \mathbb{E}\Big[\gamma^t \cdot R(s_t, a_t) \,|\, \theta, s_0 = s, a_0 = a\Big] = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \,|\, \theta, s_0 = s, a_0 = a\bigg] = Q_{\pi_\theta}(s,a),$$

which proves the unbiasedness of $\hat{Q}_{\pi_\theta}(s,a)$. Similar arguments apply to $\hat{V}_{\pi_\theta}(s)$, showing that it is an unbiased estimate of $V_{\pi_\theta}(s)$, i.e., for any $s \in \mathcal{S}$ and $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\bigg[\sum_{t=0}^{\infty} \mathbb{1}_{T' \ge t \ge 0} \cdot \gamma^{t/2} \cdot R(s_t, a_t) \,\bigg|\, \theta, s_0 = s\bigg] = \mathbb{E}[\hat{V}_{\pi_\theta}(s) \,|\, \theta, s] = V_{\pi_\theta}(s),$$

where the expectation is taken along the trajectory as well as with respect to the random horizon $T' \sim \mathrm{Geom}(1-\gamma^{t/2})$. Thus, if $s' \sim \mathbb{P}(\cdot \,|\, s, a)$ and $a' \sim \pi_\theta(\cdot \,|\, s')$, we have

$$\mathbb{E}[\hat{Q}_{\pi_\theta}(s,a) - \hat{V}_{\pi_\theta}(s) \,|\, \theta, s, a] = \mathbb{E}[R(s,a) + \gamma \hat{V}_{\pi_\theta}(s') - \hat{V}_{\pi_\theta}(s) \,|\, \theta, s, a] = A_{\pi_\theta}(s,a). \tag{4.4}$$

That is, $\hat{Q}_{\pi_\theta}(s,a) - \hat{V}_{\pi_\theta}(s)$ and $R(s,a) + \gamma \hat{V}_{\pi_\theta}(s') - \hat{V}_{\pi_\theta}(s)$ are both unbiased estimates of the advantage function $A_{\pi_\theta}(s,a)$.

Now we are ready to show the unbiasedness of $\hat{\nabla}J(\theta)$, $\check{\nabla}J(\theta)$, and $\widetilde{\nabla}J(\theta)$. First, for $\hat{\nabla}J(\theta)$, we have from (4.3) that

$$\begin{aligned}\mathbb{E}[\hat{\nabla}J(\theta)\,|\,\theta] &= \mathbb{E}_{T,(s_T,a_T)}\big\{\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}[\hat{\nabla}J(\theta)\,|\,\theta, s_T=s, a_T=a]\,\big|\,\theta\big\}\\ &= \mathbb{E}_{T,(s_T,a_T)}\bigg\{\frac{1}{1-\gamma}\cdot Q_{\pi_\theta}(s_T,a_T)\cdot\nabla\log[\pi_\theta(a_T\,|\,s_T)]\,\bigg|\,\theta\bigg\}.\end{aligned} \tag{4.5}$$

Using the identity function $\mathbb{1}_{t=T}$, (4.5) can be further written as

$$\mathbb{E}[\hat{\nabla}J(\theta)\,|\,\theta] = \frac{1}{1-\gamma}\cdot\mathbb{E}_{T,(s_T,a_T)}\bigg\{\sum_{t=0}^{\infty}\mathbb{1}_{t=T}\cdot Q_{\pi_\theta}(s_t,a_t)\cdot\nabla\log[\pi_\theta(a_t\,|\,s_t)]\,\bigg|\,\theta\bigg\}. \tag{4.6}$$

Note that by Assumption 3.1, $\|\hat{\nabla}J(\theta)\|$ is directly bounded by $(1-\gamma)^{-2}\cdot U_R\cdot B_\Theta$, since there is only one nonzero term in the summation in (4.6). Thus, by the dominated convergence theorem, we can interchange the summation and expectation in (4.6) as

$$\begin{aligned}\mathbb{E}[\hat{\nabla}J(\theta)\,|\,\theta] &= \sum_{t=0}^{\infty}\frac{P(t=T)}{1-\gamma}\cdot\mathbb{E}\bigg\{Q_{\pi_\theta}(s_t,a_t)\cdot\nabla\log[\pi_\theta(a_t\,|\,s_t)]\,\bigg|\,\theta\bigg\}\\ &= \sum_{t=0}^{\infty}\gamma^t\cdot\int_{s\in\mathcal{S},a\in\mathcal{A}} p(s_t=s,a_t=a\,|\,s_0,\pi_\theta)\cdot Q_{\pi_\theta}(s,a)\cdot\nabla\log[\pi_\theta(a\,|\,s)]dsda,\end{aligned} \tag{4.7}$$

where in (4.7) we use $T\sim\text{Geom}(1-\gamma)$ and thus $P(t=T)=(1-\gamma)\gamma^t$, and define $p(s_t=s,a_t=a\,|\,s_0,\pi_\theta)=p(s_t=s\,|\,s_0,\pi_\theta)\cdot\pi_\theta(a_t=a\,|\,s_t)$, with $p(s_t=s\,|\,s_0,\pi_\theta)$ being the probability of state $s_t=s$ given initial state s_0 and policy π_θ. By the dominated convergence theorem, we can further rewrite (4.7) by interchanging the summation and the integral, which yields the policy gradient expression given in (3.3). This shows the unbiasedness of $\hat{\nabla}J(\theta)$.

For $\check{\nabla}J(\theta)$, we have the following identity similar to (4.5):

$$\begin{aligned}\mathbb{E}[\check{\nabla}J(\theta)\,|\,\theta] &= \mathbb{E}_{T,(s_T,a_T)}\big\{\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}[\check{\nabla}J(\theta)\,|\,\theta, s_T=s, a_T=a]\,\big|\,\theta\big\}\\ &= \mathbb{E}_{T,(s_T,a_T)}\bigg\{\frac{1}{1-\gamma}\cdot A_{\pi_\theta}(s_T,a_T)\cdot\nabla\log[\pi_\theta(a_T\,|\,s_T)]\,\bigg|\,\theta\bigg\}.\end{aligned} \tag{4.8}$$

By definition of $A_{\pi_\theta}(s_T,a_T)$ and an argument similar to (4.7), (4.8) further implies

$$\mathbb{E}[\check{\nabla}J(\theta)\,|\,\theta] = \int_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{t=0}^{\infty}\gamma^t\cdot p(s_t=s\,|\,s_0,\pi_\theta)\cdot[Q_{\pi_\theta}(s,a)-V_{\pi_\theta}(s)]\cdot\nabla\pi_\theta(a\,|\,s)dsda,$$

which coincides with the policy gradient given in (3.4). Similar arguments also hold for the stochastic policy gradient $\widetilde{\nabla}J(\theta)$, since (4.8) can also be obtained from $\mathbb{E}[\widetilde{\nabla}J(\theta)\,|\,\theta]$. This proves the unbiasedness of $\check{\nabla}J(\theta)$ and $\widetilde{\nabla}J(\theta)$.

Now we establish almost sure boundedness of the stochastic policy gradients $\hat{\nabla}J(\theta)$, $\check{\nabla}J(\theta)$, and $\widetilde{\nabla}J(\theta)$. In particular, by definition of $\hat{\nabla}J(\theta)$ in (3.6),

$$\|\hat{\nabla}J(\theta)\| \le \frac{B_\Theta}{1-\gamma}\sum_{t=0}^{T'}\gamma^{t/2}\cdot U_R \le \frac{B_\Theta}{1-\gamma}\sum_{t=0}^{\infty}\gamma^{t/2}\cdot U_R = \frac{B_\Theta U_R}{(1-\gamma)(1-\gamma^{1/2})} =: \hat{\ell}, \tag{4.9}$$

where we have used Assumption 3.1, namely that $|R(s,a)| \le U_R$ and $\|\nabla \log \pi_\theta(a \,|\, s)\| \le B_\Theta$ for any s, a and θ. Similarly, we arrive at the following bounds:

(4.10)
$$\|\check{\nabla} J(\theta)\| \le \frac{2B_\Theta}{1-\gamma} \sum_{t=0}^{\infty} \gamma^{t/2} \cdot U_R = \frac{2B_\Theta U_R}{(1-\gamma)(1-\gamma^{1/2})} =: \check{\ell},$$

(4.11)
$$\|\widetilde{\nabla} J(\theta)\| \le \frac{B_\Theta}{1-\gamma}\left[1 + \left(\gamma + 1\right)\left(\sum_{t=0}^{\infty} \gamma^{t/2}\right)\right] \cdot U_R \le \frac{(2+\gamma-\gamma^{1/2}) B_\Theta U_R}{(1-\gamma)(1-\gamma^{1/2})} =: \widetilde{\ell},$$

which completes the proof. ∎

We now state the almost sure convergence of the RPG algorithm to *stationary-point* policies. Note that this result has been established in our published paper [53] and is included here (without proof) for completeness.

THEOREM 4.4 (asymptotic convergence of RPG Algorithm, Theorem 2 in [53]). *Let $\{\theta_k\}_{k\ge 0}$ be the sequence of parameters of the policy π_{θ_k} given by the RPG algorithm in (3.9). Then under Assumption 3.1, with stepsize $\{\alpha_k\}_{k\ge 0}$ satisfying $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we have $\lim_{k\to\infty} \theta_k \in \Theta^*$, where Θ^* is the set of stationary points of $J(\theta)$.*

The proof of Theorem 4.4 (see section IV in [53]) is built upon the supermartingale convergence argument [43], following the standard convergence analysis in stochastic optimization. Note that it differs from the asymptotic convergence analysis of actor-critic algorithms that is based on ODE methods [11]. Such an optimization perspective can be leveraged thanks to the unbiasedness of the stochastic policy gradients by Theorem 4.3. Besides asymptotic convergence, this perspective also facilitates the finite-iteration complexity analysis of PG methods—see Theorem 3 in [53] (also section 4 of [54]) for details—which is challenging to obtain using ODE methods.

We now shift gears to establish finite-iteration convergence for the MRPG algorithm, i.e., Algorithm 3.5. To this end, we first introduce the following condition, built upon Assumption 3.1, which is required in what follows.

ASSUMPTION 4.5. *The MDP and the parameterized policy π_θ satisfy the following:*

(i) *The reward $R(s,a)$ is either positive or negative for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Thus, $|R(s,a)| \in [L_R, U_R]$ with some $L_R > 0$.*

(ii) *The score function $\nabla \log \pi_\theta$ exists, and its norm is bounded by $\|\nabla \log \pi_\theta\| \le B_\Theta$ for any θ. Also, the Jacobian of $\nabla \log \pi_\theta$ has bounded norm and is Lipschitz continuous; i.e., there exist $\rho_\Theta > 0$ and $L_\Theta < \infty$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\nabla^2 \log \pi_{\theta^1}(a \,|\, s) - \nabla^2 \log \pi_{\theta^2}(a \,|\, s)\| \le \rho_\Theta \cdot \|\theta^1 - \theta^2\| \quad \text{for any} \quad \theta^1, \theta^2$$
$$\|\nabla^2 \log \pi_\theta(a \,|\, s)\| \le L_\Theta \quad \text{for any} \quad \theta.$$

(iii) *The integral of the Fisher information matrix induced by $\pi_\theta(\cdot \,|\, s)$ is positive-definite uniformly for any $\theta \in \mathbb{R}^d$; i.e., there exists a constant $L_I > 0$ such that*

$$\int_{s\in\mathcal{S}, a\in\mathcal{A}} \rho_\theta(s,a) \cdot \nabla \log \pi_\theta(\cdot \,|\, s) \cdot [\nabla \log \pi_\theta(\cdot \,|\, s)]^\top d a d s \succeq L_I \boldsymbol{I} \quad \text{for all} \quad \theta \in \mathbb{R}^d.$$

Note that Assumption 4.5 implies Assumption 3.1 in general and is not stringent. First, the strict positivity (or negativity) of the reward function in Assumption 4.5(i) can be easily satisfied by adding an offset to the original nonnegative and upper-bounded reward, while it is known that adding an offset does not affect the optimal policy of the original MDP (see Lemma 5.3 in [54] for a formal argument).

The positivity (or negativity) of the rewards ensures that the absolute value of the Q-function is also lower-bounded by $L_R/(1-\gamma)$. This enables the convergence of the MRPG algorithm to SOSP, as will be specified shortly. In other words, such a *reshaping* of the rewards yields better convergence results. To the best of our knowledge, our result is the first theoretical study regarding the benefit of reward-reshaping on the convergence property of PG methods, despite its popularity in practice.

Second, we note that such positivity of $|Q_{\pi_\theta}|$ will cause a relatively large variance in the original RPG update (3.9). This makes the RPG with a baseline function, i.e., the use of $\check{\nabla} J(\theta)$ and $\widetilde{\nabla} J(\theta)$, necessary in variance reduction.

Condition (ii) in Assumption 4.5 can also be satisfied easily by commonly used policies. For example, for a Gaussian policy, $\nabla^2 \log \pi_\theta(a \,|\, s)$ reduces to the matrix $\phi(s)\phi(s)^\top/\sigma^2$, which is a constant function of θ and thus satisfies (ii). Such a condition is used to show the Lipschitz continuity of the Hessian matrix of $J(\theta)$, which is standard in establishing the convergence to SOSP in nonconvex optimization literature [16, 22, 25, 51]. Formally, the Lipschitz continuity of the Hessian is substantiated in the following lemma. The proof, which follows by direct calculation of the Lipschitz constant, is tedious and thus deferred to Appendix A.6 in [54].

LEMMA 4.6. *The Hessian matrix $\mathcal{H}(\theta)$ of the objective function $J(\theta)$ is Lipschitz continuous, i.e., with some constant $\rho > 0$, $\|\mathcal{H}(\theta^1) - \mathcal{H}(\theta^2)\| \le \rho \cdot \|\theta^1 - \theta^2\|$, for any $\theta^1, \theta^2 \in \mathbb{R}^d$. The value of the Lipschitz constant ρ is given in* (A.68) *in Appendix* A.6 *in* [54].

Condition (iii) in Assumption 4.5 holds for many *regular* policy parameterizations and has been assumed in prior works on the natural policy gradient [27] and actor-critic algorithms [10].

We now show that $\hat{\nabla} J(\theta)$, $\check{\nabla} J(\theta)$, and $\widetilde{\nabla} J(\theta)$ all satisfy the so-called correlated negative curvature (CNC) condition [16], which is crucial in the ensuing analysis.

LEMMA 4.7. *Under Assumption* 4.5*, all three stochastic policy gradients $\hat{\nabla} J(\theta)$, $\check{\nabla} J(\theta)$, and $\widetilde{\nabla} J(\theta)$ satisfy the CNC condition; i.e., letting v_θ be the unit-norm eigenvector corresponding to the maximum eigenvalue of the Hessian matrix $\mathcal{H}(\theta)$, there exist constants $\hat{\eta}, \check{\eta}, \widetilde{\eta} > 0$ such that for any $\theta \in \mathbb{R}^d$, $\mathbb{E}\big\{[\mathrm{v}_\theta^\top \hat{\nabla} J(\theta)]^2 \,\big|\, \theta\big\} \ge \hat{\eta}$, $\mathbb{E}\big\{[\mathrm{v}_\theta^\top \check{\nabla} J(\theta)]^2 \,\big|\, \theta\big\} \ge \check{\eta}$, and $\mathbb{E}\big\{[\mathrm{v}_\theta^\top \widetilde{\nabla} J(\theta)]^2 \,\big|\, \theta\big\} \ge \widetilde{\eta}$.*

Proof. We start with $\mathbb{E}\big\{[\mathrm{v}_\theta^\top \hat{\nabla} J(\theta)]^2 \,\big|\, \theta\big\}$. By definition, for any $\mathrm{v} \in \mathbb{R}^d$ and $\|\mathrm{v}\| = 1$,

$$\begin{aligned}\mathbb{E}\big\{[\mathrm{v}^\top \hat{\nabla} J(\theta)]^2 \,\big|\, \theta\big\} &= \mathbb{E}\big\{\big[\hat{Q}_{\pi_\theta}(s_T, a_T) \cdot \mathrm{v}^\top \nabla \log \pi_\theta(a_T \,|\, s_T)\big]^2 \,\big|\, \theta\big\} \\ &= \mathbb{E}_{T,(s_T,a_T)}\big\{\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}^2_{\pi_\theta}(s_T, a_T) \,\big|\, \theta, s_T, a_T\big] \cdot \big[\mathrm{v}^\top \nabla \log \pi_\theta(a_T \,|\, s_T)\big]^2 \,\big|\, \theta\big\}.\end{aligned} \tag{4.12}$$

We write $\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}[\hat{Q}^2_{\pi_\theta}(s_T, a_T) \,|\, \theta, s_T, a_T]$ as $\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}[\hat{Q}^2_{\pi_\theta}(s_T, a_T)]$ for notational simplicity. Note that the conditional expectation is taken over the sequence $(s_{1:T'}, a_{1:T'})$ and the random variable T', given θ and s_T, a_T. Then since the reward

satisfies $|R| > L_R > 0$, we can uniformly lower-bound $\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}[\hat{Q}^2_{\pi_\theta}(s_T,a_T)]$ as

$$\begin{aligned}\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}^2_{\pi_\theta}(s_T,a_T)\big] &= \mathbb{E}_{T'}\bigg(\mathbb{E}_{(s_{1:T'},a_{1:T'})}\Big\{\Big[\sum_{t=0}^{T'}\gamma^{t/2}\cdot R(s_t,a_t)\Big]^2\,\Big|\,T'=\tau\Big\}\bigg)\\ &\geq \mathbb{E}_{T'}\bigg(\frac{1-\gamma^{(T'+1)/2}}{1-\gamma^{1/2}}\cdot L_R\bigg)^2 \geq L_R^2\cdot\sum_{\tau=0}^{\infty}\gamma^{\tau/2}(1-\gamma^{1/2}) = L_R^2 > 0,\end{aligned}$$

where the first inequality holds because $R(s,a)$ is either all positive or all negative for any s,a, and the second inequality follows by $[1-\gamma^{(T'+1)/2}]\cdot(1-\gamma^{1/2})^{-1}\geq 1$ for all $T'\geq 0$. Substituting the preceding expression into the first product term on the RHS of (4.12) and pulling out the vector v yield

$$\begin{aligned}\mathbb{E}\big\{[\mathrm{v}^\top\hat{\nabla}J(\theta)]^2\,\big|\,\theta\big\} &\geq L_R^2\cdot\mathrm{v}^\top\cdot\mathbb{E}_{T,(s_T,a_T)}\big\{\nabla\log\pi_\theta(a_T\,|\,s_T)\cdot\nabla\log\pi_\theta(a_T\,|\,s_T)^\top\,\big|\,\theta\big\}\cdot\mathrm{v}\\ &\geq L_R^2\cdot L_I =: \hat{\eta} > 0,\end{aligned}\tag{4.13}$$

where the second inequality follows from condition (iii) in Assumption 4.5. Note that (4.13) holds for any unit-norm vector v and also holds for any eigenvector v_θ corresponding to the maximum eigenvalue of $\mathcal{H}(\theta)$. This verifies the first argument.

Similarly, for $\mathbb{E}\big\{[\mathrm{v}^\top\check{\nabla}J(\theta)]^2\,\big|\,\theta\big\}$, we start with the expected value of the square of the inner product of $\check{\nabla}J(\theta)$ with a unit vector v. By definition of $\check{\nabla}J(\theta)$, we have

$$\begin{aligned}&\mathbb{E}\big\{[\mathrm{v}^\top\check{\nabla}J(\theta)]^2\,\big|\,\theta\big\} = \mathbb{E}\big\{\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\cdot\big[\mathrm{v}^\top\nabla\log\pi_\theta(a_T\,|\,s_T)\big]^2\,\big|\,\theta\big\}\\ &\quad= \mathbb{E}_{T,(s_T,a_T)}\big\{\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\cdot[\mathrm{v}^\top\nabla\log\pi_\theta(a_T\,|\,s_T)]^2\,\big|\,\theta\big\},\end{aligned}\tag{4.14}$$

where for notational simplicity we write $\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big\{\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\,\big|\,\theta, s_T,a_T\big\}$ as $\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2$. Notice that

$$\begin{aligned}&\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\\ &\quad= \big\{\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]\big\}^2 + \mathrm{Var}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]\\ &\quad= \big[Q_{\pi_\theta}(s_T,a_T)-V_{\pi_\theta}(s_T)\big]^2 + \mathrm{Var}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)\big] + \mathrm{Var}\big[\hat{V}_{\pi_\theta}(s_T)\big],\end{aligned}\tag{4.15}$$

where the second equation follows from the fact that $\hat{Q}_{\pi_\theta}(s_T,a_T)$ and $\hat{V}_{\pi_\theta}(s_T)$ are independent and unbiased estimates of $Q_{\pi_\theta}(s_T,a_T)$ and $V_{\pi_\theta}(s_T)$, respectively. Note that the first term in (4.15) may be zero, for example, when π_θ is a degenerate policy such that that $\pi_\theta(a\,|\,s_T)=\mathbb{1}_{a=a_T}$. Hence, a uniform lower bound on the two variance terms in (4.15) needs to be established. By definition, we have

$$\begin{aligned}\mathrm{Var}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)\big] &= \mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-Q_{\pi_\theta}(s_T,a_T)\big]^2\\ &= \mathbb{E}_{T'}\Big(\mathbb{E}_{(s_{1:T'},a_{1:T'})}\big\{\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-Q_{\pi_\theta}(s_T,a_T)\big]^2\,\big|\,T'=\tau\big\}\Big).\end{aligned}\tag{4.16}$$

Given (s_T,a_T), θ, and $T'=\tau$, the conditional expectation in (4.16) is expanded as

$$\begin{aligned}&\mathbb{E}_{(s_{1:T'},a_{1:T'})}\big\{\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-Q_{\pi_\theta}(s_T,a_T)\big]^2\,\big|\,T'=\tau\big\}\\ &\quad= \mathbb{E}_{(s_{1:T'},a_{1:T'})}\Big\{\Big[\sum_{t=0}^{T'}\gamma^{t/2}\cdot R(s_t,a_t)-Q_{\pi_\theta}(s_T,a_T)\Big]^2\,\Big|\,T'=\tau\Big\}.\end{aligned}$$

Now we first focus on the case when $R(s,a)$ is strictly positive, i.e., $R(s,a) \in [L_R, U_R]$. In this case, $Q_{\pi_\theta}(s_T, a_T)$ is a scalar that lies in $[L_R/(1-\gamma), U_R/(1-\gamma)]$. Also, notice that $\mathbb{E}_{(s_{1:T'}, a_{1:T'})}[\sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t)]$ is a strictly increasing function of T' since $R(s,a) \geq L_R > 0$ for any (s,a). Moreover, notice that given (s_T, a_T), $\mathbb{E}_{T', (s_{1:T'}, a_{1:T'})}[\sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t)]$ is an unbiased estimate of $Q_{\pi_\theta}(s_T, a_T)$, and T' follows the geometric distribution over nonnegative support. Thus, there must exist a finite $T_* \geq 0$, such that

(4.17)
$$\mathbb{E}_{(s_{1:T_*}, a_{1:T_*})}\left[\sum_{t=0}^{T_*} \gamma^{t/2} \cdot R(s_t, a_t)\right] < Q_{\pi_\theta}(s_T, a_T) \leq \mathbb{E}_{(s_{1:T_*+1}, a_{1:T_*+1})}\left[\sum_{t=0}^{T_*+1} \gamma^{t/2} \cdot R(s_t, a_t)\right].$$

As a result, we can substitute (4.17) into the RHS of (4.16), yielding

$$\mathrm{Var}\left[\hat{Q}_{\pi_\theta}(s_T, a_T)\right] = \sum_{\tau=0}^{\infty} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \mathbb{E}_{(s_{1:\tau}, a_{1:\tau})}\left[\sum_{t=0}^{\tau} \gamma^{t/2} \cdot R(s_t, a_t) - Q_{\pi_\theta}(s_T, a_T)\right]^2$$

(4.18)
$$\geq \sum_{\tau=0}^{\infty} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left\{\mathbb{E}_{(s_{1:\tau}, a_{1:\tau})}\left[\sum_{t=0}^{\tau} \gamma^{t/2} \cdot R(s_t, a_t) - Q_{\pi_\theta}(s_T, a_T)\right]\right\}^2$$

(4.19)
$$\geq \sum_{\tau=0}^{T_*} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \sum_{t=\tau+1}^{T_*} \gamma^{t/2}\right]^2 + \sum_{\tau=T_*+2}^{\infty} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \sum_{t=T_*+2}^{\tau} \gamma^{t/2}\right]^2,$$

where (4.18) uses $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$, and (4.19) follows by removing the term with $\tau = T_*$ and $\tau = T_*+1$ in the summation in (4.18) that sandwiched $Q_{\pi_\theta}(s_T, a_T)$, and noticing the fact that the term $\mathbb{E}_{(s_{1:\tau}, a_{1:\tau})}[\sum_{t=0}^{\tau} \gamma^{t/2} \cdot R(s_t, a_t)]$ is at least $L_R \cdot \sum_{t=T_*+2}^{\tau} \gamma^{t/2}$ away from $Q_{\pi_\theta}(s_T, a_T)$ when $\tau \geq T_* + 2$, and at least $L_R \cdot \sum_{t=\tau+1}^{T_*} \gamma^{t/2}$ away from[4] $Q_{\pi_\theta}(s_T, a_T)$ when $\tau \leq T_*$. Furthermore, multiplying the first term in (4.19) by $\gamma^{3/2}$ yields

$$\begin{aligned}
\mathrm{Var}\left[\hat{Q}_{\pi_\theta}(s_T, a_T)\right] &\geq \gamma^{3/2} \cdot \sum_{\tau=0}^{T_*} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \frac{\gamma^{(\tau+1)/2} - \gamma^{(T_*+1)/2}}{1-\gamma^{1/2}}\right]^2 \\
&\quad + \sum_{\tau=T_*+2}^{\infty} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \frac{\gamma^{(\tau+1)/2} - \gamma^{(T_*+2)/2}}{1-\gamma^{1/2}}\right]^2, \qquad (4.20) \\
&= \gamma^{3/2} \cdot \sum_{\tau=0}^{T_*} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \frac{\gamma^{(\tau+1)/2} - \gamma^{(T_*+1)/2}}{1-\gamma^{1/2}}\right]^2, \\
&\quad + \gamma^{3/2} \cdot \sum_{\tau=T_*+1}^{\infty} \gamma^{\tau/2}(1-\gamma^{1/2}) \cdot \left[L_R \cdot \frac{\gamma^{(\tau+1)/2} - \gamma^{(T_*+1)/2}}{1-\gamma^{1/2}}\right]^2, \qquad (4.21)
\end{aligned}$$

where (4.20) follows by $\gamma^{3/2} < 1$, and (4.21) is obtained by changing the starting point of the summation of the second term to $T_* + 1$, and then pulling out $\gamma^{1/2}$ from

[4]Note that we define $\sum_{t=\tau+1}^{T_*} \gamma^{t/2} = 0$ if $\tau + 1 < T_*$.

the square bracket. Equation (4.21) can then be further bounded as

$$\begin{aligned}\operatorname{Var}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)\big] &\geq \gamma^{3/2}\cdot L_R^2\cdot \mathbb{E}_{T'}\left[\frac{\gamma^{(T'+1)/2}-\gamma^{(T_*+1)/2}}{1-\gamma^{1/2}}\right]^2\\ &\geq \gamma^{3/2}\cdot L_R^2\cdot \operatorname{Var}\left[\frac{\gamma^{(T'+1)/2}-\gamma^{(T_*+1)/2}}{1-\gamma^{1/2}}\right]=\gamma^{3/2}\cdot L_R^2\cdot \operatorname{Var}\left[\frac{\gamma^{(T'+1)/2}}{1-\gamma^{1/2}}\right],\end{aligned}\tag{4.22}$$

where the first inequality follows since the RHS of (4.21) is an expectation over T', the second inequality follows from $\mathbb{E}(X^2)\geq \operatorname{Var}(X)$, and the last equation is due to the fact that T_* is deterministic. Note that $\operatorname{Var}[\gamma^{(T'+1)/2}]$ can be uniformly bounded as

$$\begin{aligned}\operatorname{Var}[\gamma^{(T'+1)/2}] &= \mathbb{E}[\gamma^{(T'+1)/2}]^2-\big\{\mathbb{E}[\gamma^{(T'+1)/2}]\big\}^2=\frac{\gamma(1-\gamma^{1/2})}{1-\gamma^{3/2}}-\left[\frac{\gamma^{1/2}(1-\gamma^{1/2})}{1-\gamma}\right]^2\\ &= \frac{\gamma^{3/2}\cdot(1-\gamma^{1/2})^3}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}>0.\end{aligned}\tag{4.23}$$

Combining (4.22) and (4.23), we obtain

$$\operatorname{Var}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)\big]\geq \frac{\gamma^{3/2}\cdot L_R^2}{(1-\gamma^{1/2})^2}\cdot\frac{\gamma^{3/2}\cdot(1-\gamma^{1/2})^3}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}=\frac{L_R^2\cdot\gamma^3\cdot(1-\gamma^{1/2})}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}.\tag{4.24}$$

By the same arguments as above, we can also obtain that

$$\operatorname{Var}\big[\hat{V}_{\pi_\theta}(s_T)\big]\geq \frac{L_R^2\cdot\gamma^3\cdot(1-\gamma^{1/2})}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}.\tag{4.25}$$

Substituting (4.24) and (4.25) into (4.15), we arrive at

$$\mathbb{E}_{T',(s_{1:T'},a_{1:T'})}\big[\hat{Q}_{\pi_\theta}(s_T,a_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\geq\frac{2L_R^2\cdot\gamma^3\cdot(1-\gamma^{1/2})}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}.\tag{4.26}$$

Finally, by combining (4.26) and (4.14), we conclude that

$$\mathbb{E}\big\{[\mathrm{v}^\top\check{\nabla}J(\theta)]^2\,\big|\,\theta\big\}\geq\frac{2L_R^2\cdot\gamma^3\cdot(1-\gamma^{1/2})}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}\cdot L_I=:\check{\eta}>0.$$

The proof for the case when $R(s,a)\in[-U_R,-L_R]$ is similar to the one above, with only some minor modifications due to sign flipping. For example, $\mathbb{E}_{(s_{1:T'},a_{1:T'})}[\sum_{t=0}^{T'}\gamma^{t/2}\cdot R(s_t,a_t)]$ now becomes a strictly decreasing function of T' since $R(s,a)\leq -L_R<0$. The remaining arguments are similar and so are omitted here to avoid repetition.

Regarding $\mathbb{E}\big\{[\mathrm{v}^\top\widetilde{\nabla}J(\theta)]^2\,\big|\,\theta\big\}\geq\widetilde{\eta}$, by definition we have

$$\begin{aligned}\mathbb{E}\big\{[\mathrm{v}^\top\widetilde{\nabla}J(\theta)]^2\,\big|\,\theta\big\} &= \mathbb{E}\big\{\big[R(s_T,a_T)+\gamma\cdot\hat{V}_{\pi_\theta}(s_T')-\hat{V}_{\pi_\theta}(s_T)\big]^2\cdot\big[\mathrm{v}^\top\nabla\log\pi_\theta(a_T\,|\,s_T)\big]^2\,\big|\,\theta\big\}\\ &= \mathbb{E}_{T,(s_T,a_T)}\big\{\mathbb{E}_{s_T',T',T'',(s_{1:T'},a_{1:T'}),(s_{1:T''},a_{1:T''})}\big[R(s_T,a_T)+\gamma\cdot\hat{V}_{\pi_\theta}(s_T')-\hat{V}_{\pi_\theta}(s_T)\big]^2\\ &\quad\cdot[\mathrm{v}^\top\nabla\log[\pi_\theta(a_T\,|\,s_T)]^2\,\big|\,\theta\big\},\end{aligned}\tag{4.27}$$

Table 1: List of parameter values used in the convergence analysis.

Param.	Value	Order	Constraint	Const.
β	$c_1\epsilon^2/(2\ell^2 L)$	$\mathcal{O}(\epsilon^2)$	$\le \epsilon^2/(2\ell^2 L)$	$c_1 = 1$
β	"	"	$\le [J_{\text{thre}}\delta/(2L\ell^2)]^{1/2}$	"
β	"	$\mathcal{O}(\epsilon)$	$\le \eta\lambda^2/(24L\ell^3\rho)$	"
J_{thre}	$c_2\eta\epsilon^4/(2\ell^2 L)$	$\mathcal{O}(\epsilon^4)$	$\le \beta\epsilon^2/2$	$c_2 = c_1/2$
J_{thre}	"	"	$\le \eta\beta\lambda^2/(48\ell\rho)$	"
α	$c_1\epsilon^2/(2\ell^2 L\sqrt{k_{\text{thre}}})$	$\mathcal{O}(\epsilon^{9/2})$	$\le \beta/\sqrt{k_{\text{thre}}}$	
α	"	"	$\le c'\eta\beta\lambda^3/(24L\ell^3\rho)$	"
k_{thre}	$c_4\frac{\log[L\ell_g/(\eta\beta\alpha\sqrt{\rho\epsilon})]}{\alpha(\rho\epsilon)^{1/2}}$	$\Omega(\epsilon^{-5}\log(1/\epsilon))$	$\ge c\frac{\log[L\ell_g/(\eta\beta\alpha\lambda)]}{\alpha\lambda}$	$c_4 = c$
K	$c_5\frac{[J^*-J(\theta_0)]k_{\text{thre}}}{\delta J_{\text{thre}}}$	$\Omega(\epsilon^{-9}\log(1/\epsilon))$	$\ge 2\frac{[J^*-J(\theta_0)]k_{\text{thre}}}{\delta J_{\text{thre}}}$	$c_5 = 2$

where we use T' and T'' to denote the random horizons used in calculating $\hat{V}_{\pi_\theta}(s'_T)$ and $\hat{V}_{\pi_\theta}(s_T)$, respectively, and recall that $s'_T \sim \mathbb{P}(\cdot \,|\, s_T, a_T)$. Given (s_T, a_T), we have

$$\begin{aligned}
&\mathbb{E}_{s'_T,T',T'',(s_{1:T'},a_{1:T'}),(s_{1:T''},a_{1:T''})}\big[R(s_T,a_T)+\gamma\cdot\hat{V}_{\pi_\theta}(s'_T)-\hat{V}_{\pi_\theta}(s_T)\big]^2\\
&\quad=\big\{\mathbb{E}_{s'_T,T',T'',(s_{1:T'},a_{1:T'}),(s_{1:T''},a_{1:T''})}\big[R(s_T,a_T)+\gamma\cdot\hat{V}_{\pi_\theta}(s'_T)-\hat{V}_{\pi_\theta}(s_T)\big]\big\}^2\\
&\qquad+\gamma^2\cdot\mathrm{Var}[\hat{V}_{\pi_\theta}(s'_T)]+\mathrm{Var}[\hat{V}_{\pi_\theta}(s_T)]\\
&\quad=\big[Q_{\pi_\theta}(s_T,a_T)-V_{\pi_\theta}(s_T)\big]^2+\gamma^2\cdot\mathrm{Var}[\hat{V}_{\pi_\theta}(s'_T)]+\mathrm{Var}[\hat{V}_{\pi_\theta}(s_T)], \qquad (4.28)
\end{aligned}$$

where (4.28) is due to the independence and unbiasedness of the estimates $\hat{V}_{\pi_\theta}(s'_T)$ and $\hat{V}_{\pi_\theta}(s_T)$. Then, by (4.24), we can lower-bound (4.28) and thus (4.27) as

$$\mathbb{E}\big\{[\mathrm{v}^\top\widetilde{\nabla}J(\theta)]^2\,\big|\,\theta\big\}\ge\frac{(1+\gamma^2)\cdot L_R^2\cdot\gamma^3\cdot(1-\gamma^{1/2})}{(1-\gamma^{3/2})\cdot(1-\gamma)^2}\cdot L_I=:\widetilde{\eta}>0,$$

which completes the proof. ∎

The CNC condition basically guarantees that the perturbation caused by the stochastic gradient has variance along the direction with positive curvature, i.e., the escaping direction of the objective [16]. Such an escaping direction is dictated by the eigenvectors associated with the maximum eigenvalue of the Hessian matrix $\mathcal{H}(\theta)$.

Now we are ready to lay out the improved convergence guarantees of the MRPG algorithm, i.e., Algorithm 3.5, in the following theorem. The values of the parameters used in the analysis are specified in Table 1.[5] Unlike the recent progress on escaping saddle points for nonconvex optimization [22, 25, 26], our analysis under the CNC condition does not require adding artificial isotropic noises to the update. This is especially necessary for RL, compared to stochastic gradient descent for standard nonconvex optimization, since (i) in RL the noise results from the sampling along the trajectory of the MDP, which does not necessarily satisfy the isotropic property in general; and (ii) the noise of policy gradients is notoriously known to be large, and thus adding artificial noise may further degrade the performance of PG algorithms.

THEOREM 4.8. *Under Assumption* 4.5*, Algorithm* 3.5 *returns an* $(\epsilon, \sqrt{\rho\epsilon})$-

[5]The lower bound of some large enough constant c in the table is defined in the detailed proof in [54].

approximate SOSP with probability at least $(1-\delta)$ *after*

$$\mathcal{O}\Big(\Big(\frac{\rho^{3/2}L\epsilon^{-9}}{\delta\eta}\Big)\log\Big(\frac{\ell_g L}{\epsilon\eta\rho}\Big)\Big) \tag{4.29}$$

steps, where $\delta\in(0,1)$, $\ell_g^2 := 2\ell^2 + 2B_\Theta^2 U_R^2/(1-\gamma)^4$, B_Θ, U_R *are as defined in Assumption* 4.5, ρ *is the Lipschitz constant of the Hessian matrix in Lemma* 4.6, *and* ℓ *and* η *take the values of* $\hat{\ell},\check{\ell},\widetilde{\ell}$ *in Theorem* 4.3 *and* $\hat{\eta},\check{\eta},\widetilde{\eta}$ *in Lemma* 4.7, *when the stochastic policy gradients* $\hat{\nabla}J(\theta),\check{\nabla}J(\theta)$, *and* $\widetilde{\nabla}J(\theta)$ *are used, respectively.*

Proof. Our analysis is separated into three steps that characterize the convergence properties of the iterates in three different regimes, depending on the magnitude of the gradient and the curvature of the Hessian. Note that Algorithm 3.5 returns the iterates that have indices k with $k \bmod k_{\text{thre}} = 0$, i.e., the iterates belong to the set $\hat{\Theta}^*$. For notational convenience, we index the iterates in $\hat{\Theta}^*$ by m, i.e., let $\widetilde{\theta}_m = \theta_{m\cdot k_{\text{thre}}}$ for all $m = 0,1,\ldots,\lfloor K/k_{\text{thre}}\rfloor$. Now we consider the three regimes of the iterates $\{\widetilde{\theta}_m\}_{m\geq 0}$. For notational convenience, we use g_θ to unify the notation of the three stochastic policy gradients $\hat{\nabla}J(\theta)$, $\check{\nabla}J(\theta)$, and $\widetilde{\nabla}J(\theta)$.

Regime 1: Large gradient. We first introduce the following lemma, which quantifies the increase of function values when stochastic gradient ascent of a smooth function is adopted.

LEMMA 4.9. *Let* θ_{k+1} *be obtained by one stochastic gradient ascent step at* θ_k, *i.e.,* $\theta_{k+1} = \theta_k + \alpha g_k$, *where* $g_k = g_{\theta_k}$ *is an unbiased stochastic gradient at* θ_k. *Then, for any given* θ_k, *the function value* $J(\theta_{k+1})$ *increases in expectation*[6] *as* $\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \alpha\|\nabla J(\theta_k)\|^2 - L\alpha^2\ell^2/2$.

Proof. By the L-smoothness of $J(\theta)$, we have

$$\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \alpha\nabla J(\theta_k)^\top \mathbb{E}(g_k \,|\, \theta_k) - \frac{L\alpha^2}{2}\|g_k\|^2 = \alpha\|\nabla J(\theta_k)\|^2 - \frac{L\alpha^2}{2}\|g_k\|^2,$$

which completes the proof by using the fact that $\|g_k\|^2 \leq \ell^2$ almost surely. ∎

Therefore, when the norm of the gradient is large at $\widetilde{\theta}_m$, a large increase of $J(\widetilde{\theta})$ from $\widetilde{\theta}_m$ to $\widetilde{\theta}_{m+1}$ is guaranteed, as formally stated below.

LEMMA 4.10. *Suppose the gradient norm at any given* $\widetilde{\theta}_m$ *is large such that* $\|\nabla J(\widetilde{\theta}_m)\| \geq \epsilon$ *for some* $\epsilon > 0$. *Then, the expected value of* $J(\widetilde{\theta}_{m+1})$ *increases as*

$$\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) \geq J_{thre},$$

where the expectation is taken over the sequence from $\theta_{m\cdot k_{thre}+1}$ *to* $\theta_{(m+1)\cdot k_{thre}}$.

Proof. The difference between the expected values of $J(\widetilde{\theta}_{m+1})$ and $J(\widetilde{\theta}_m)$ is

$$\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) = \sum_{p=0}^{k_{\text{thre}}-1} \mathbb{E}\big\{\mathbb{E}[J(\theta_{m\cdot k_{\text{thre}}+p+1})] - J(\theta_{m\cdot k_{\text{thre}}+p}) \,\big|\, \theta_{m\cdot k_{\text{thre}}+p}\big\},$$

where $\mathbb{E}[J(\theta_{m\cdot k_{\text{thre}}}) \,|\, \theta_{m\cdot k_{\text{thre}}}] = J(\theta_{m\cdot k_{\text{thre}}}) = J(\widetilde{\theta}_m)$ for given $\widetilde{\theta}_m$. By Lemma 4.9,

$$\begin{aligned}\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) &\geq \beta\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2 - \frac{L\beta^2\ell^2}{2} + \sum_{p=1}^{k_{\text{thre}}-1} \alpha\mathbb{E}\|\nabla J(\theta_{m\cdot k_{\text{thre}}+p})\|^2 - \frac{k_{\text{thre}}L\alpha^2\ell^2}{2} \\ &\geq \beta\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2 - \frac{L\beta^2\ell^2}{2} - \frac{k_{\text{thre}}L\alpha^2\ell^2}{2} \geq \beta\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2 - L\beta^2\ell^2, \end{aligned} \tag{4.30}$$

[6] Note that the expectation here is taken over the randomness of g_k.

where the last inequality follows from Table 1 and that

$$\beta^2 \geq k_{\text{thre}} \cdot \alpha^2. \tag{4.31}$$

Moreover, by the choice of the large stepsize β, we have

$$\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2 = \|\nabla J(\widetilde{\theta}_m)\|^2 \geq \epsilon^2 \geq 2\ell^2 L\beta, \tag{4.32}$$

which yields a lower bound on the RHS of (4.30) as

$$\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) \geq \beta\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2 - L\beta^2\ell^2 \geq \beta\|\nabla J(\theta_{m\cdot k_{\text{thre}}})\|^2/2 \geq \beta\epsilon^2/2 \geq J_{\text{thre}}. \tag{4.33}$$

The choice of $J_{\text{thre}} \leq \beta\epsilon^2/2$ completes the proof. ∎

Regime 2: Near saddle points. When the iterate reaches the neighborhood of saddle points, the MRPG will use a larger stepsize β to find the positive eigenvalue direction and then use small stepsize α to follow this positive curvature direction. We establish in the following lemma that such an updating strategy also leads to a sufficient increase of function value, provided that the maximum eigenvalue of the Hessian $\mathcal{H}(\widetilde{\theta}_m)$ is large enough. This enables the iterate to escape the saddle points efficiently.

LEMMA 4.11. *Suppose that the Hessian matrix at any given $\widetilde{\theta}_m$ has a large positive eigenvalue such that $\lambda_{\max}[\mathcal{H}(\widetilde{\theta}_m)] \geq \sqrt{\rho\epsilon}$. Then, after k_{thre} steps, we have $\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) \geq J_{thre}$, where the expectation is taken over the sequence from $\theta_{m\cdot k_{thre}+1}$ to $\theta_{(m+1)\cdot k_{thre}}$.*

Proof. The proof is based on the *improve or localize* framework proposed in [26]. The basic idea is as follows: starting from some iterate, if the following iterates of the stochastic gradient update do not improve the objective value to a great degree, then the iterates must not move much from the starting iterate. Our goal here is to show that after k_{thre} steps, the objective value will increase by at least J_{thre}. In particular, the proof proceeds by contradiction: suppose the objective value does not increase by J_{thre} from $\widetilde{\theta}_m$ to $\widetilde{\theta}_{m+1}$, i.e.,

$$\mathbb{E}(J_{k_{\text{thre}}}) - J_0 \leq J_{\text{thre}}; \tag{4.34}$$

then the distance between the two iterates can be upper-bounded by a polynomial function of the number of iterates in between, i.e., k_{thre}, as stated below.

LEMMA 4.12. *Given any θ_0, suppose (4.34) holds for any $0 \leq p \leq k_{thre}$. Then, the expected distance between θ_p and θ_0 can be upper-bounded as*

$$\mathbb{E}\|\theta_p - \theta_0\|^2 \leq \left[4\alpha^2\ell_g^2 + 4\alpha J_{thre} + 2L\alpha(\ell\beta)^2 + 2L\ell^2\alpha^3 k_{thre}\right] \cdot p + 2\beta^2\ell^2, \tag{4.35}$$

where $\ell_g^2 := 2\ell^2 + 2B_\Theta^2 U_R^2 \cdot (1-\gamma)^{-4}$.

Proof. We have obtained from Lemma 4.9 and (4.30) (with $m = 0$) that

$$\begin{aligned}\mathbb{E}(J_{k_{\text{thre}}}) - J_0 &\geq \beta\|\nabla J_0\|^2 - \frac{L\beta^2\ell^2}{2} + \sum_{q=1}^{k_{\text{thre}}-1} \alpha\mathbb{E}\|\nabla J_q\|^2 - \frac{k_{\text{thre}}L\alpha^2\ell^2}{2} \\ &\geq -\frac{L\beta^2\ell^2}{2} + \alpha\sum_{q=0}^{p-1}\mathbb{E}\|\nabla J_q\|^2 - \frac{k_{\text{thre}}L\alpha^2\ell^2}{2},\end{aligned}$$

since $0 \leq \alpha < \beta$ and $0 \leq p \leq k_{\text{thre}}$, where we note that the total expectation is taken along the sequence from θ_1 to $\theta_{k_{\text{thre}}}$, and we write $\|\nabla J_0\|^2 = \mathbb{E}\|\nabla J_0\|^2$ since θ_0 is given and deterministic. Combined with (4.34), we have

$$J_{\text{thre}} \geq \alpha \sum_{q=0}^{p-1} \mathbb{E}\|\nabla J_q\|^2 - \frac{k_{\text{thre}} L \alpha^2 \ell^2}{2} - \frac{L\beta^2\ell^2}{2},$$

which implies that

$$\sum_{q=0}^{p-1} \mathbb{E}\|\nabla J_q\|^2 \leq \frac{J_{\text{thre}}}{\alpha} + \frac{k_{\text{thre}} L \alpha \ell^2}{2} + \frac{L\beta^2\ell^2}{2\alpha}. \tag{4.36}$$

Note that the distance between θ_p and θ_0 can be decomposed as follows:

$$\mathbb{E}\|\theta_p - \theta_0\|^2 = \mathbb{E}\left\|\sum_{q=0}^{p-1} \theta_{q+1} - \theta_q\right\|^2 \leq 2\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} g_q\right\|^2 + 2\beta^2 \mathbb{E}\big\|g_0\big\|^2, \tag{4.37}$$

where the first equality comes from the telescopic property of the summand and the later inequality comes from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

For the first term on the RHS of (4.37), we have

$$\begin{aligned} 2\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} g_q\right\|^2 &\leq 2\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} g_q - \nabla J_q + \nabla J_q\right\|^2 \\ &\leq 4\alpha^2 \mathbb{E}\sum_{q=1}^{p-1} \big\|g_q - \nabla J_q\big\|^2 + 4\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} \nabla J_q\right\|^2, \end{aligned} \tag{4.38}$$

where the inequality follows from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and the fact that $\mathbb{E}[(g_p - \nabla J_p)^\top (g_q - \nabla J_q)] = 0$ for any $p \neq q$, since the stochastic error $g_p - \nabla J_q$ across iterations is independent, and g_p is an unbiased estimate of ∇J_p. Moreover, due to the boundedness of $\|\nabla J_q\|$ and $\|g_q\|$ (cf. Theorem 4.3), we have

$$\mathbb{E}\big\|g_q - \nabla J_q\big\|^2 \leq 2\mathbb{E}\big\|g_q\big\|^2 + 2\mathbb{E}\big\|\nabla J_q\big\|^2 \leq 2\ell^2 + \frac{2B_\Theta^2 U_R^2}{(1-\gamma)^4} =: \ell_g^2.$$

Thus, by the Cauchy–Schwarz inequality and (4.36), we can further upper-bound the RHS of (4.38) as

$$\begin{aligned} 2\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} g_q\right\|^2 &\leq 4\alpha^2 \sum_{q=1}^{p-1} \mathbb{E}\big\|g_q - \nabla J_q\big\|^2 + 4\alpha^2 \mathbb{E}\left\|\sum_{q=1}^{p-1} \nabla J_q\right\|^2 \\ &\leq 4\alpha^2 \cdot (p-1) \cdot \ell_g^2 + 4\alpha^2 \cdot (p-1) \cdot \sum_{q=1}^{p-1} \mathbb{E}\big\|\nabla J_q\big\|^2 \\ &\leq 4\alpha^2 \cdot (p-1) \cdot \ell_g^2 + 4\alpha^2 \cdot (p-1) \cdot \left(\frac{J_{\text{thre}}}{\alpha} + \frac{k_{\text{thre}} L \alpha \ell^2}{2} + \frac{L\beta^2\ell^2}{2\alpha}\right) \\ &\leq 4\alpha^2 \cdot (p-1) \cdot \left(\ell^2 + \frac{J_{\text{thre}}}{\alpha} + \frac{pL\alpha\ell^2}{2} + \frac{L\beta^2\ell^2}{2\alpha}\right), \end{aligned} \tag{4.39}$$

where we recall that the expectation is taken over the random sequence $\{\theta_1, \dots, \theta_{p-1}\}$.

For the second term on the RHS of (4.37), observe that $\mathbb{E}\|g_0\|^2 \le \ell^2$. Therefore, combined with (4.39), we can upper-bound (4.37) as

$$\begin{aligned}\mathbb{E}(\|\theta_p - \theta_0\|^2) &\le 4\alpha^2 \cdot (p-1) \cdot \ell_g^2 + 4\alpha^2 \cdot (p-1) \cdot \left(\frac{J_{\text{thre}}}{\alpha} + \frac{k_{\text{thre}} L \alpha \ell^2}{2} + \frac{L\beta^2\ell^2}{2\alpha}\right) + 2\beta^2\ell^2 \\ &\le \left[4\alpha^2 \cdot \ell_g^2 + 4\alpha^2 \cdot \left(\frac{J_{\text{thre}}}{\alpha} + \frac{k_{\text{thre}} L \alpha \ell^2}{2} + \frac{L\beta^2\ell^2}{2\alpha}\right)\right] \cdot p + 2\beta^2\ell^2,\end{aligned}$$

which completes the proof. ∎

On the other hand, due to the CNC condition (cf. Lemma 4.7), the distance between $\widetilde{\theta}_m$ and $\widetilde{\theta}_{m+1}$ can be shown to be lower-bounded by an exponential function of k_{thre}. In particular, we note that for any θ close to θ_0, the function value $J(\theta)$ can be approximated by some quadratic function $\mathcal{Q}(\theta)$, i.e.,

$$\mathcal{Q}(\theta) = J_0 + (\theta - \theta_0)^\top \nabla J_0 + \frac{1}{2}(\theta - \theta_0)^\top \mathcal{H}_0 (\theta - \theta_0). \tag{4.40}$$

The benefit of defining $\mathcal{Q}(\theta)$ is that the difference between the gradients of J and $\mathcal{Q}$ can now be bounded as

$$\|\nabla J(\theta) - \nabla \mathcal{Q}(\theta)\| \le \rho \|\theta - \theta_0\|^2/2 \tag{4.41}$$

for ρ-Hessian Lipschitz function J [34]. For convenience, we let $\nabla \mathcal{Q}_p = \nabla \mathcal{Q}(\theta_p)$ for any $p = 0, \dots, k_{\text{thre}} - 1$. Then, we can express the difference between any θ and θ_0 in terms of the difference between the gradients $\nabla \mathcal{Q}_p$ and ∇J_p and thus relate it back to the difference between θ and θ_0 from (4.41). In particular, for any $p \ge 0$, we can decompose $\theta_{p+1} - \theta_0$ as follows:

$$\begin{aligned}\theta_{p+1} - \theta_0 &= \theta_p - \theta_0 + \alpha g_p = \theta_p - \theta_0 + \alpha \nabla \mathcal{Q}_p + \alpha(g_p - \nabla \mathcal{Q}_p + \nabla J_p - \nabla J_p) \\ &= (\mathrm{I} + \alpha \mathcal{H}_0)(\theta_p - \theta_0) + \alpha(\nabla J_p - \nabla \mathcal{Q}_p + g_p - \nabla J_p + \nabla J_0) \\ &= \underbrace{(\mathrm{I} + \alpha \mathcal{H}_0)^p (\theta_1 - \theta_0)}_{\mathrm{u}_p} + \alpha \cdot \Bigg[\underbrace{\sum_{q=1}^{p} (\mathrm{I} + \alpha \mathcal{H}_0)^{p-q} (\nabla J_q - \nabla \mathcal{Q}_q)}_{\boldsymbol{\delta}_p} \\ &\quad + \underbrace{\sum_{q=1}^{p} (\mathrm{I} + \alpha \mathcal{H}_0)^{p-q} \nabla J_0}_{\mathrm{d}_p} + \underbrace{\sum_{q=1}^{p} (\mathrm{I} + \alpha \mathcal{H}_0)^{p-q} (g_q - \nabla J_q)}_{\boldsymbol{\xi}_p} \Bigg],\end{aligned} \tag{4.42}$$

where I is the identity matrix and u_p, $\boldsymbol{\delta}_p$, d_p, and $\boldsymbol{\xi}_p$ are defined as above, and where we recall that $\mathcal{H}_0 = \nabla^2 J(\theta_0)$ denotes the Hessian matrix evaluated at θ_0. The first equation uses the update from θ_p to θ_{p+1}, and the second one adds and subtracts ∇J_q and $\nabla \mathcal{Q}_q$. The third equation uses the definition of $\nabla \mathcal{Q}_p$ from (4.40), and the last one follows by iteratively unrolling the third equation p times. As a result, we can lower-bound the distance $\mathbb{E}\|\theta_{p+1} - \theta_0\|^2$ by

$$\begin{aligned}\mathbb{E}\|\theta_{p+1} - \theta_0\|^2 &\ge \mathbb{E}\|\mathrm{u}_p\|^2 + 2\alpha \mathbb{E}(\mathrm{u}_p^\top \boldsymbol{\delta}_p) + 2\alpha \mathbb{E}(\mathrm{u}_p^\top \mathrm{d}_p) + 2\alpha \mathbb{E}(\mathrm{u}_p^\top \boldsymbol{\xi}_p) \\ &\ge \mathbb{E}\|\mathrm{u}_p\|^2 - 2\alpha \mathbb{E}(\|\mathrm{u}_p\| \|\boldsymbol{\delta}_p\|) + 2\alpha \mathbb{E}(\mathrm{u}_p^\top) \mathrm{d}_p + 2\alpha \mathbb{E}(\mathrm{u}_p^\top \boldsymbol{\xi}_p).\end{aligned} \tag{4.43}$$

Then, one can lower-bound the terms on the RHS of (4.43) as a exponential function of k_{thre}, using the CNC condition, as follows:

$$\mathbb{E}\|\theta_{p+1}-\theta_0\|^2 \geq \eta\beta^2\kappa^{2p} - 2\alpha\cdot\Big[\big(4\ell\alpha^2\cdot\ell_g^2 + 4\ell\alpha J_{\text{thre}} + 2L\ell^3\alpha^3 k_{\text{thre}} + 2L\alpha\beta^2\ell^3\big) \cdot \rho\beta\cdot\frac{\kappa^{2p}}{(\alpha\lambda)^2} + 2\rho\beta^3\ell^3\cdot\frac{\kappa^{2p}}{\alpha\lambda}\Big]$$

$$= \Big(\eta\beta - \frac{8\ell\alpha\ell_g^2\rho}{\lambda^2} - \frac{8\ell J_{\text{thre}}\rho}{\lambda^2} - \frac{4L\ell^3\alpha^2 k_{\text{thre}}\rho}{\lambda^2} - \frac{4L\beta^2\ell^3\rho}{\lambda^2} - \frac{4\beta^2\ell^3\rho}{\lambda}\Big)\cdot\beta\kappa^{2p}. \tag{4.44}$$

Then, by properly choosing the parameters as in Table 1, the terms in the bracket on the RHS of (4.44) can be made positive (see a detailed verification in section A.9 in [54]). As a result, with a large enough k_{thre}, the lower bound in (4.44) will exceed the upper bound in Lemma 4.12, creating a contradiction. This completes the proof.□

Lemma 4.11 is the most crucial part in the proof, which asserts that after k_{thre} steps, the expected function value increases by at least J_{thre}. Together with Lemma 4.10, it is shown that the expected return $\mathbb{E}[J(\widetilde{\theta}_{m+1})]$ is always increasing, as long as the iterate $\widetilde{\theta}_m$ violates the SOSP condition, i.e., $\|\nabla J(\widetilde{\theta}_m)\| \geq \epsilon$ or $\lambda_{\max}[\mathcal{H}(\widetilde{\theta}_m)] \geq \sqrt{\rho\epsilon}$.

Regime 3: Near second-order stationary points. When the iterate converges to the neighborhood of the desired SOSP, both the norm of the gradient and the largest eigenvalue are small. However, due to the variance of the stochastic policy gradient, the function value may still decrease. By Lemma 4.9 and (4.33), we can immediately show that such a decrease is bounded, i.e.,

$$\mathbb{E}[J(\widetilde{\theta}_{m+1})] - J(\widetilde{\theta}_m) \geq -L\beta^2\ell^2 \geq -\delta J_{\text{thre}}/2, \tag{4.45}$$

which is due to the choice of $J_{\text{thre}} \geq 2L(\ell\beta)^2/\delta$ as in Table 1.

Now we combine the arguments above to obtain a probabilistic lower bound on the returned SOSP. Let $\mathcal{E}_m$ be the event that

$$\mathcal{E}_m := \{\|\nabla J(\widetilde{\theta}_m)\| \geq \epsilon \text{ or } \lambda_{\max}[\mathcal{H}(\widetilde{\theta}_m)] \geq \sqrt{\rho\epsilon}\}.$$

By Lemmas 4.10 and 4.11, we have

$$\mathbb{E}[J(\widetilde{\theta}_{m+1}) - J(\widetilde{\theta}_m)\,|\,\mathcal{E}_m] \geq J_{\text{thre}}, \tag{4.46}$$

where the expectation is taken over the randomness of both $\widetilde{\theta}_{m+1}$ and $\widetilde{\theta}_m$ given the event $\mathcal{E}_m$. Namely, after k_{thre} steps, as long as $\widetilde{\theta}_m$ is not an $(\epsilon, \sqrt{\rho\epsilon})$-approximate SOSP, a sufficient increase of $\mathbb{E}[J(\widetilde{\theta}_{m+1})]$ is guaranteed. Otherwise, we can still control the possible decrease of the return using (4.45), which yields

$$\mathbb{E}[J(\widetilde{\theta}_{m+1}) - J(\widetilde{\theta}_m)\,|\,\mathcal{E}_m^c] \geq -\delta J_{\text{thre}}/2, \tag{4.47}$$

where $\mathcal{E}_m^c$ is the complement event of $\mathcal{E}_m$.

Let $\mathcal{P}_m$ denote the probability of the occurrence of the event $\mathcal{E}_m$. Thus, the total expectation $\mathbb{E}[J(\widetilde{\theta}_{m+1}) - J(\widetilde{\theta}_m)]$ can be obtained by combining (4.46) and (4.47) as

$$\mathbb{E}[J(\widetilde{\theta}_{m+1}) - J(\widetilde{\theta}_m)] \geq (1-\mathcal{P}_m)\cdot\Big(-\frac{\delta J_{\text{thre}}}{2}\Big) + \mathcal{P}_m\cdot J_{\text{thre}}. \tag{4.48}$$

Suppose the iterate θ_k runs for K steps starting from θ_0; then there are $M = \lfloor K/k_{\text{thre}}\rfloor$ of $\widetilde{\theta}_m$ for $k = 1,\ldots,K$. Summing up all the M steps of $\{\widetilde{\theta}_m\}_{m=1,\ldots,M}$, we obtain from

(4.48) that $\sum_{m=1}^{M} \mathcal{P}_m/M \le [J^* - J(\theta_0)]/(MJ_{\rm thre}) + \delta/2 \le \delta$, where J^* is the global maximum, and the last inequality follows from the choice of K in Table 1 that satisfies

$$K \ge 2[J^* - J(\theta_0)]k_{\rm thre}/(\delta J_{\rm thre}). \tag{4.49}$$

Therefore, the probability of the event $\mathcal{E}_m^c$ occurring, i.e., the probability of retrieving an $(\epsilon, \sqrt{\rho\epsilon})$ SOSP uniformly over the iterates in $\hat{\Theta}^*$, can be lower-bounded by $1 - \sum_{m=1}^{M} \mathcal{P}_m/M \ge 1 - \delta$. This completes the proof of Theorem 4.8.

The proof of Theorem 4.8 originates from but improves the proof in [16] and maps it to the analysis of PG methods in RL.[7] Note that in the proof, we follow the convention of using $(\epsilon, \sqrt{\rho\epsilon})$ as the convergence criterion for approximate SOSPs [25, 35, 59], which reflects the natural relation between the gradient and the Hessian. Theorem 4.8 concludes that it is possible for the policy gradient algorithm to escape the saddle points efficiently and retrieve an SOSP in a polynomial number of steps.[8] Additionally, if all saddle points are *strict* (cf. definition in [22]), the MRPG algorithm will converge to an actual *locally optimal* policy.

To the best of our knowledge, these results are the first convergence to approximately locally optimal policies for *general* RL settings in the literature, which does not conflate convergence to stationarity with local extrema. Due to space limitations, we defer the numerical validation of our theory to section 6 of the companion report [54].

5. Conclusions. Despite their tremendous popularity, PG methods in RL have rarely been investigated in terms of their global convergence, i.e., the limiting as well as the finite-iteration property of policy gradient updates, when the underlying problem is viewed as an optimization routine. Driven by this shortcoming of the existing literature, we have adopted the perspective and tools from nonconvex optimization to clarify and partially overcome some of the challenges of PG methods. In particular, we have developed a series of random-horizon policy gradient algorithms, which generate *unbiased* estimates of the policy gradient for the *infinite-horizon* discounted setting. Under standard assumptions for RL, we have first recovered the convergence to stationary-point policies for such first-order optimization algorithms. Moreover, by virtue of the recent advancements in nonconvex optimization, we have proposed modified RPG (MRPG) algorithms by introducing periodically enlarged stepsizes, which are shown to be able to escape saddle points and converge to actual *locally optimal* policies under additional mild conditions that are satisfied for most RL applications. Specifically, we have provided an intriguing optimization-based explanation as to why one type of reward-reshaping, i.e., shifting the reward by an offset, might be beneficial: it increases the variance of the stochastic policy gradient to some extent, which can be exploited to escape the saddle points and thus improve the convergence (from first- to second-order stationary points), without changing the landscape or optimal policy of the RL problem. Many enhancements are possible for future research directions via the link between policy search and nonconvex optimization, such as rate

[7] Our result mirrors Theorem 2 in [16]. However, we have identified and informed the authors, and have been acknowledged for doing so, that there is a flaw in their proof, which breaks the convergence rate claimed in the original paper. At the time the current draft was prepared, the authors of [16] corrected the arXiv version using an idea similar to what we proposed in the personal communication (as shown here).

[8] Note that the number of steps here in (4.29) corresponds to the notion of *iteration complexity* in the optimization literature, which is not the total *sample complexity* since each step of our algorithm requires two rollouts with random but finite horizon. Thus, the *expected* number of samples, i.e., state-action-reward tuples, equals $1/(1-\gamma) + 1/(1-\gamma^{1/2})$ times the expression in (4.29).

improvements through acceleration, trust region methods, variance reduction through Polyak–Ruppert averaging, and quasi-Newton methods.

REFERENCES

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, *Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes*, preprint, https://arxiv.org/abs/1908.00261v1, 2019.

[2] R. G. Bartle, *The Elements of Integration and Lebesgue Measure*, John Wiley & Sons, New York, 2014.

[3] P. L. Bartlett, J. Baxter, and L. Weaver, *Experiments with infinite-horizon, policy-gradient estimation*, J. Artificial Intelligence Res., 15 (2001), pp. 351–381.

[4] J. Baxter and P. L. Bartlett, *Infinite-horizon policy-gradient estimation*, J. Artificial Intelligence Res., 15 (2001), pp. 319–350.

[5] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol.* I, 3rd ed., Athena Scientific, Belmont, MA, 2005.

[7] J. Bhandari and D. Russo, *Global Optimality Guarantees for Policy Gradient Methods*, preprint, https://arxiv.org/abs/1906.01786, 2019.

[8] S. Bhatnagar, *An actor–critic algorithm with function approximation for discounted cost constrained Markov Decision Processes*, Systems Control Lett., 59 (2010), pp. 760–766.

[9] S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S. Sutton, *Incremental natural actor-critic algorithms*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2008, pp. 105–112.

[10] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, *Natural actor-critic algorithms*, Automatica J. IFAC, 45 (2009), pp. 2471–2482.

[11] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press, Cambridge, UK, 2008.

[12] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, *LQR through the Lens of First Order Methods: Discrete-Time Case*, preprint, https://arxiv.org/abs/1907.08921, 2019.

[13] D. D. Castro and R. Meir, *A convergent online single-time-scale actor-critic algorithm*, J. Mach. Learn. Res., 11 (2010), pp. 367–410.

[14] T. Chen, K. Zhang, G. B. Giannakis, and T. Başar, *Communication-Efficient Distributed Reinforcement Learning*, preprint, https://arxiv.org/abs/1812.03239, 2018.

[15] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, *Risk-constrained reinforcement learning with percentile risk criteria*, J. Mach. Learn. Res., 18 (2017), 167.

[16] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann, *Escaping saddles with stochastic gradients*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 1155–1164.

[17] C. D. Dang and G. Lan, *On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators*, Comput. Optim. Appl., 60 (2015), pp. 277–310.

[18] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2014, pp. 2933–2941.

[19] K. Doya, *Reinforcement learning in continuous time and space*, Neural Comput., 12 (2000), pp. 219–245.

[20] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, *Global convergence of policy gradient methods for the linear quadratic regulator*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 1467–1476.

[21] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, *Counterfactual multi-agent policy gradients*, in Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, 2018, pp. 2794–2982.

[22] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points–online stochastic gradient for tensor decomposition*, in Proceedings of the 28th Conference on Learning Theory, Paris, France, 2015, pp. 797–842.

[23] E. Greensmith, P. L. Bartlett, and J. Baxter, *Variance reduction techniques for gradient estimates in reinforcement learning*, J. Mach. Learn. Res., 5 (2004), pp. 1471–1530.

[24] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, *MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels*, in Proceedings of the Inter-

national Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 2304–2313.
[25] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, in Proceedings of the International Conference on Machine Learning, Sydney, Australia, 2017, pp. 1724–1732.
[26] C. Jin, P. Netrapalli, and M. I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, in Proceedings of the 31st Annual Conference on Learning Theory, Stockholm, Sweden, 2018, pp. 1042–1085.
[27] S. M. Kakade, *A natural policy gradient*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2002, pp. 1531–1538.
[28] V. R. Konda and V. S. Borkar, *Actor-critic–type learning algorithms for Markov decision processes*, SIAM J. Control Optim., 38 (1999), pp. 94–123, https://doi.org/10.1137/S036301299731669X.
[29] V. R. Konda and J. N. Tsitsiklis, *Actor-critic algorithms*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2000, pp. 1008–1014.
[30] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, *Continuous control with deep reinforcement learning*, in Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2016.
[31] B. Liu, Q. Cai, Z. Yang, and Z. Wang, *Neural Proximal/Trust Region Policy Optimization Attains Globally Optimal Policy*, preprint, https://arxiv.org/abs/1906.10306, 2019.
[32] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, *Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments*, preprint, https://arxiv.org/abs/1706.02275, 2017.
[33] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, *Asynchronous methods for deep reinforcement learning*, in Proceedings of the International Conference on Machine Learning, New York, NY, 2016, pp. 1928–1937.
[34] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Boston, MA, 2004.
[35] Y. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[36] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli, *Stochastic variance-reduced policy gradient*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 4026–4035.
[37] M. Papini, M. Pirotta, and M. Restelli, *Adaptive batch size for safe policy gradients*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2017, pp. 3591–3600.
[38] S. Paternain, *Stochastic Control Foundations of Autonomous Behavior*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 2018.
[39] R. Pemantle, *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.
[40] J. Peters and S. Schaal, *Policy gradient methods for robotics*, in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Washingon, DC, 2006, pp. 2219–2225.
[41] M. Pirotta, M. Restelli, and L. Bascetta, *Adaptive step-size for policy gradient methods*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2013, pp. 1394–1402.
[42] M. Pirotta, M. Restelli, and L. Bascetta, *Policy gradient in Lipschitz Markov decision processes*, Mach. Learn., 100 (2015), pp. 255–283.
[43] H. Robbins and D. Siegmund, *A convergence theorem for non-negative almost supermartingales and some applications*, in Herbert Robbins Selected Papers, Springer, New York, 1985, pp. 111–135.
[44] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, *Trust region policy optimization*, in Proceedings of the International Conference on Machine Learning, Lille, France, 2015, pp. 1889–1897.
[45] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Ser. Optim. 9, Mathematical Programming Society, Philadelphia, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718751.
[46] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, *Deterministic policy gradient algorithms*, in Proceedings of the International Conference on Machine Learning, Beijing, China, 2014, pp. 379–387.
[47] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, Cambridge, MA, 2018.
[48] 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, NeurIPS, San Diego, CA, 2000, pp. 1057–1063.

[49] R. J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, Mach. Learn., 8 (1992), pp. 229–256.

[50] S. Wright and J. Nocedal, *Numerical Optimization*, Springer Ser. Oper. Res., Springer, New York, 1999.

[51] P. Xu, F. Roosta-Khorasani, and M. W. Mahoney, *Newton-Type Methods for Non-Convex Optimization under Inexact Hessian Information*, preprint, https://arxiv.org/abs/1708.07164, 2017.

[52] J. Yeh, *Real Analysis: Theory of Measure and Integration*, 2nd ed., World Scientific, Hackensack, NJ, 2006.

[53] K. Zhang, A. Koppel, H. Zhu, and T. Başar, *Convergence and iteration complexity of policy gradient method for infinite-horizon reinforcement learning*, in Proceedings of the IEEE Conference on Decision and Control, IEEE, Washington, DC, 2019, pp. 7415–7422.

[54] K. Zhang, A. Koppel, H. Zhu, and T. Başar, *Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies*, preprint, https://arxiv.org/abs/1906.08383, 2019.

[55] K. Zhang, Z. Yang, and T. Başar, *Networked multi-agent reinforcement learning in continuous spaces*, in Proceedings of the IEEE Conference on Decision and Control, IEEE, Washington, DC, 2018, pp. 2771–2776.

[56] K. Zhang, Z. Yang, and T. Başar, *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, preprint, https://arxiv.org/abs/1911.10635, 2019.

[57] K. Zhang, Z. Yang, and T. Başar, *Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2019, pp. 11602–11614.

[58] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, *Fully decentralized multi-agent reinforcement learning with networked agents*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 5872–5881.

[59] Y. Zhang, P. Liang, and M. Charikar, *A hitting time analysis of stochastic gradient Langevin dynamics*, in Proceedings of the Conference on Learning Theory, Amsterdam, The Netherlands, 2017, pp. 765–775.

[60] Z. Zhou, S. Athey, and S. Wager, *Offline Multi-Action Policy Learning: Generalization and Optimization*, preprint, https://arxiv.org/abs/1810.04778, 2018.

[61] Z. Zhou, M. Bloem, and N. Bambos, *Infinite time horizon maximum causal entropy inverse reinforcement learning*, IEEE Trans. Automat. Control, 63 (2017), pp. 2787–2802.

[62] Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn, *Stochastic mirror descent in variationally coherent optimization problems*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2017, pp. 7040–7049.

[63] Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn, *On the convergence of mirror descent beyond stochastic convex programming*, SIAM J. Optim., 30 (2020), pp. 687–716, https://doi.org/10.1137/17M1134925.

[64] Z. Zhou, P. Mertikopoulos, N. Bambos, P. Glynn, Y. Ye, L.-J. Li, and L. Fei-Fei, *Distributed asynchronous optimization with unbounded delays: How slow can you go?*, in Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018, pp. 5970–5979.