Convergence and Optimality of Policy Gradient Methods in Weakly Smooth Settings

Matthew Shunshi Zhang[1], **Murat A. Erdogdu**[1, 2]. **Animesh Garg**[1]

[1] Department of Computer Science, University of Toronto and Vector Institute
[2] Department of Statistical Sciences, University of Toronto
matthew.zhang@mail.utoronto.ca, erdogdu@cs.toronto.edu, garg@cs.toronto.edu

Abstract

Policy gradient methods have been frequently applied to problems in control and reinforcement learning with great success, yet existing convergence analysis still relies on non-intuitive, impractical and often opaque conditions. In particular, existing rates are achieved in limited settings, under strict smoothness and bounded conditions. In this work, we establish explicit convergence rates of policy gradient methods without relying on these conditions, instead extending the convergence regime to weakly smooth policy classes with L_2 integrable gradient. We provide intuitive examples to illustrate the insight behind these new conditions. We also characterize the sufficiency conditions for the ergodicity of near-linear MDPs, which represent an important class of problems. Notably, our analysis also shows that fast convergence rates are achievable for both the standard policy gradient and the natural policy gradient algorithms under these assumptions. Lastly we provide conditions and analysis for optimality of the converged policies.

Introduction

Modern Reinforcement Learning (RL) has solved challenges in diverse fields such as finance, healthcare, and robotics (Deng et al. 2016; Yu, Liu, and Nemati 2019; Kober, Bagnell, and Peters 2013). Nonetheless, the theory behind these methods remains poorly understood, with convergence and optimality results being limited to narrow classes of problems. Classical approaches to RL theory focus on tabular problems where discrete techniques can be applied (see (Agarwal et al. 2020b; Sidford et al. 2018)). However, most practical problems exist in continuous, high-dimensional domains (Doya 2000), and may even be infinite-dimensional or non-compact.

Theoretical results in continuous domains do not effectively characterize algorithms for policy gradient. In contrast, value-based estimators have obtained strong results in some regimes such as linear MDPs, both in on- and off-line settings (Cai et al. 2019; Yang and Wang 2019). Nonetheless, direct policy estimators possess numerous advantages, in that they are (theoretically) insensitive to perturbations in the problem parameters, and are smoother to estimate. However, bounds for direct parameterizations of the policy have been less successful. They either restrict the cardinality or size of the space (Agarwal et al. 2020b), or apply strong assumptions on the policy and MDP (Liu et al. 2019; Xu, Wang, and Liang 2020; Liu et al. 2020). This conflicts with practical results, where convergence often occurs without boundedness or smoothness preconditions on the function approximator.

Consequently, in this paper, we analyse two key questions: (i) how can we *relax existing conditions on MDPs* while retaining guarantees for fast convergence, (ii) how can *optimality of the value function be obtained* in these contexts. Arguably, the convergence of gradient algorithms needs to rely on some constraints of the function class. Prior work has relied on assumptions of (a) MDP ergodicity, (b) policy smoothness and (c) absolute boundedness of the gradient. However, these conditions are overly restrictive and exclude many useful function approximators.

Summary of Contributions. We make significant contributions with respect to each of these assumptions in a stochastic setting: (a) ergodicity is assumed in the general case, but we show ergodicity for a class of smooth linear MDPs using Markov Chain theory; (b) strong smoothness is relaxed to weak smoothness (Hölder conditions) of the policy and its gradient; (c) absolute boundedness is relaxed to L_2 integrability under regular measures. While this is an important theoretical development, it also expands the scope of practical convergence results. We include many practical examples of MDPs and policies that satisfy our criteria, with applications to exploration and safety in reinforcement learning. To the best of our knowledge, ours is the first study to consider this setting, and to show explicit ergodicity results for continuous-state MDPs.

Under these assumptions, we find that with the optimal learning rate, the gradient decreases as $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^{\frac{1+\beta_0}{\beta_0}}\right] \leq \epsilon$ for both the standard and natural policy gradient with $T \times B = O\left(\epsilon^{-\frac{3+\beta_0}{2}}(1-\gamma)^{-\frac{1}{\beta_0}}\right)$, where β_0 is the weak smoothness parameter defined explicitly in Assumption 1. We also show (Theorem 2) that the converged policy for the natural policy gradient is optimal up to a policy-dependent factor:

$$J(\pi^*) - \frac{1}{T}\sum_{t\leq T}\mathbb{E}\left[J(\theta_t)\right] \leq \epsilon + \mathcal{O}\left(\frac{E_\Pi}{(1-\gamma)^{1.5}}\right), \qquad (1)$$

with rate $T \times B = O\left((1-\gamma)^{(2\beta_0+11)/2\beta_0}\epsilon^{(\beta_0+3)/\beta_0}\right)$, for

Table 1: Comparison of Optimality Rates (Theorem 2) in Different Settings. β_0 is the weak smoothness parameter.

Work	Setting	Rate (PG)
(Agarwal et al. 2020b)	$\lvert\mathcal{S}\rvert < \infty$, Exact PG	$\epsilon^2(1-\gamma)^3$
(Liu et al. 2020)	Smooth ∇, L_∞	$\epsilon^4(1-\gamma)^2$
Ours	Weak Smooth ∇, L_2	$\epsilon^{\frac{\beta_0+3}{\beta_0}}(1-\gamma)^{\frac{\beta_0+4}{\beta_0}}$

a batch size B, where E_Π can be tuned by choosing an appropriately regular policy class. E_Π is formally defined in the statement of Theorem 1. Under a strong additional assumption, standard policy gradient is also asymptotically optimal: $J(\pi^*) - \frac{1}{T}\sum_{t\leq T}\mathbb{E}\left[J(\theta_t)\right] \leq \epsilon$, with order $T \times B = O\left((1-\gamma)^{(\beta_0+4)/\beta_0}\epsilon^{(\beta_0+3)/\beta_0}\right)$. In the strictly smooth limit these rates have previously been discovered (Agarwal et al. 2020a; Xu, Wang, and Liang 2020; Zou, Xu, and Liang 2019), although our results hold for a wider range of functions and MDPs.

The remainder of the paper is structured as follows: in §2 we cover the mathematical formulation of MDPs; in §3 we introduce the policy gradient algorithm as well as our assumptions. In §4, we list several candidate policies that satisfy our assumptions, and demonstrate their utility in a variety of contexts. §5 then states our main convergence and optimality results; §6 summarizes works related to optimization and RL theory.

Background

Markov Decision Processes

Let a state-space be denoted by $\mathcal{S}$, and an action-space by $\mathcal{A}$. Let a transition measure $P(\cdot|s,a)$ and a reward measure $R(\cdot|s,a)$ be probability measures on $\mathcal{S}$ and $\mathbb{R}$ respectively, both conditioned on variables $(s,a) \in \mathcal{S}\times\mathcal{A}$. A Markov Decision Process $\mathcal{M}$ is formally defined as a tuple $(\mathcal{S},\mathcal{A},P,R,\gamma)$, where $\gamma \in [0,1)$ is the discount factor. Unless otherwise specified, let $\|\cdot\| = \|\cdot\|_2$ the 2-norm for vectors, and $\|\cdot\|_{op}$ the operator norm for matrices. Hereafter we assume that the absolute magnitude of the rewards are bounded, i.e. $R(\cdot|s,a)$ only has support on $[-\alpha,\alpha]$ for some $\alpha \geq 0$, and all s,a.

Policies: We denote a stochastic policy $\pi : \mathcal{S} \to \Delta_\mathcal{A}$ where $\Delta_\mathcal{A}$ is the set of probability measures on $\mathcal{A}$. By abuse of notation, we also allow $\pi(\cdot|s)$ to denote the marginal density of the measure $\pi(s)$.

Trajectories: To generate trajectories, we start from an initial state distribution ρ_0, and then at each time $t \in \mathbb{N}$, we sample an action from the policy: $a_t \sim \pi(s_t)$. Subsequently a state and reward are queried $s_{t+1} \sim P(\cdot|a_t,s_t), r_t \sim R(\cdot|a_t,s_t)$, and the process continues.

Distributions: π, ρ_0 parameterize a probability distribution on the set of trajectories. Taking ρ_0 as fixed, we denote the distribution as $\{(s_t,a_t), t = 0,1,2,\ldots\} \sim \mathbb{P}_{\pi,\rho_0}$.

Value Functions: Consequently we can define the value function as: $V_\pi(s) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s]$, and the Q-function as: $Q_\pi(s,a) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s, a_0 = a]$. Note that both expectations are taken over trajectories. If $|r_t| \leq \alpha$, both functions are bounded by $[-\alpha/(1-\gamma), \alpha/(1-\gamma)]$. We can also define the advantage function $A_\pi(s,a) \triangleq Q_\pi(s,a) - V_\pi(s)$.

Discounted Visitation: It will be useful to define the sum of time-discounted visitation probabilities through the following: $d_\pi^\rho(s,a) = (1-\gamma)\sum_{t=1}^\infty \gamma^t \mathbb{P}(s_t = s, a_t = a | s_0 \sim \rho)$. This is bounded in $[0,1]$, and is a probability density for $\mathcal{S}\times\mathcal{A}$.

Reinforcement Learning A reinforcement learning agent is one which produces a sequence of policies π_t based on the queried states s_t, r_t, which seeks to iteratively maximize the value function: $J(\pi) = \mathbb{E}_{s\sim\rho_0}[V_\pi(s)]$, when π is constrained to some family of policies Π. The existence of an optimum in the space of stochastic functions has been shown as a classical result (Bellman 1954).

Our Proposed Method

In this work, we limit our discussion to exponential policy classes which are continuously differentiable. In particular, we denote the distribution of an exponential policy, parameterized by a variable $\theta \in \Theta \subset \mathbb{R}^N$, such that $\pi_{\nu_\theta}(a|s) = \frac{\exp(\nu_\theta(s,a))}{\int_\mathcal{A}\exp(\nu_\theta(s,a))}$, $\nu_\theta : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$. We require that the integral $\int_\mathcal{A}\exp(\nu_\theta(s,\cdot)) < \infty$ is finite for all $\theta \in \Theta, s \in \mathcal{S}$, and that $\nu_\theta(s,a)$ is differentiable in θ for all s,a. Let us define $\pi_\theta \triangleq \pi_{\nu_\theta}$ and $J(\theta) \triangleq J(\pi_\theta)$ as shorthands. Then the gradient can be written as $\nabla J(\theta) = \mathbb{E}_{(s,a)\sim d_\pi^\rho}[Q_{\pi_\theta}(s,a)\psi_\theta(s,a)]$ due to the value function having an expectation of zero. Let us denote the score function as $\psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s)$. While successful tabular approaches rely on explicit computation of each softmax probability, this is not feasible for most MDPs where the action space is infinite. Typically some form of function class is required to address this issue. In this work, we consider all softmax functions that satisfy the following smoothness properties:

Assumption 1. *(Smoothness of Policy Class) Consider policies* $\pi_\theta \propto \exp(\nu_\theta)$. *We require that* π *obeys the following two smoothness conditions:*

$$\int_\mathcal{A} \pi_\theta(a|s)\log\frac{\pi_\theta(a|s)}{\pi_{\theta+\eta}(a|s)}da \leq C_{\nu,1}\|\eta\|^{\beta_1}, \tag{2}$$

$$\int_\mathcal{A} \|\psi_\theta(s,a) - \psi_{\theta+\eta}(s,a)\|\,\pi_\theta(a|s) \leq C_{\nu,2}\|\eta\|^{\beta_2}; \tag{3}$$

where the constants $C_{\nu,1}, C_{\nu,2} \geq 0$, $\beta_1 \in [1,2], \beta_2 \in (0,1]$ *are valid for all* θ, s. *Consequently we define* $\beta_0 = \min(\beta_1/4, \beta_2)$ *as the primary order of smoothness.*

We note that (2) is a Hölder condition on the Kullback–Leibler (KL) divergence of the policies, while (3) is a Hölder requirement on the gradient.

Remarks: $\beta_1 < 2, \beta_2 < 1$ are weakly smooth cases. This is much more permissive than traditional assumptions on Lipschitz smoothness; particularly, it allows for slow tail decay (of order $\|x\|^{\beta_2}$) but fast local growth. We can alternatively phrase the KL requirement in terms of the Total Variation distance between the two distributions ($\int |\pi_\theta - \pi_{\theta+\eta}|$) using the Pinsker inequality. It is also possible to relax this assumption to local conditions (i.e. only holding when $\|\eta\| \leq R$), with an additional error term.

We introduce an additional assumption on the variances of the gradient:

Assumption 2. *(Boundedness of Gradient Moments) Assume that the score function is absolutely bounded in L_2 across all policies, with respect to its own generated state-action distribution, i.e. that the following holds:*

$$\int \|\psi_\theta(s,a)\|_2^2 \, d_\theta^\rho(s,a) \leq \psi_\infty, \tag{4}$$

for any θ in our parameter space, where $\psi_\infty < \infty$ is a constant independent of θ.

Remarks: The second condition is weak and simply guarantees the expected gradient is bounded; for the first condition, we can strengthen the norm to higher orders ($L_q, q > 2$) to smoothly approach the rates achieved with absolute boundedness.

Finally, we require the following standard assumption (see e.g. (Xu, Wang, and Liang 2020; Zou, Xu, and Liang 2019)) which is sufficient to show smoothness of the objective function. It is also of independent interest, since it can be used to show the convergence of samplers of states and actions.

Assumption 3. *(Ergodicity) We have for all states $s \in \mathcal{S}$:*

$$\left\| \mathbb{P}_{\pi_\theta}^n(\cdot | s_0 = s) - \rho_*(\cdot) \right\| \leq C_0 \delta^n,$$

where $\mathbb{P}_{\pi_\theta}^n$ is the n-step state transition kernel following π_θ, ρ_* is the invariant state distribution, $C_0 \geq 0, \delta < 1$ are constants independent of s, θ.

We show this property explicitly for a subclass of linear MDPs, which serves as an additional contribution for our paper.

Proposition 1. *Suppose that the MDP is linear, such that the dynamics can be written as $P(s'|s,a) = \langle \boldsymbol{\mu}(s'), \phi(s,a) \rangle$ for all s, a and some function $\boldsymbol{\mu}(\cdot) : \mathcal{S} \mapsto \mathbb{R}^d$ and mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$. Under some weak assumptions on $\boldsymbol{\mu}$ (see Appendix Section A), Assumption 3 holds.*

While the class of linear and near-linear MDPs is known to be rather limited, this result still represents an important contribution as the first known proof of ergodicity in continuous MDPs. The proof in the appendix also serves as a blueprint for proofs of ergodicity in more general cases.

Given these assumptions on the policy class, we can apply direct policy ascent on the space of parameters in order to get the gradient update

$$\theta_t = \theta_{t-1} + h_t \nabla_\theta J(\theta_{t-1}), \tag{5}$$

where $h_t \in \mathbb{R}$ is an adaptive step size. Alternatively, natural policy gradient (NPG), first introduced by (Kakade 2001), is a parameter invariant method that applies the following update

$$\theta_t = \theta_{t-1} + h_t K^\dagger(\theta) \nabla_\theta J(\theta_{t-1}), \tag{6}$$

where $K(\theta) = \mathbb{E}_{s,a\sim d_\theta^\rho}\left[\psi_\theta(s,a)\psi_\theta(s,a)^\top\right]$. Here $(\cdot)^\dagger$ is the matrix pseudo-inverse. The advantages of this method are that the optimization landscape becomes nearly convex, as we see in our analysis. Since the true loss function and Fisher information matrix are not available to us, we estimate each of them through sampling. In particular, we use the following minibatch estimators:

$$\widehat{\nabla_\theta J(\theta_t)} = \frac{1}{B}\sum_{i=1}^{B} r_{t,i}\psi_\theta(s_{t,i}, a_{t,i}), \tag{7}$$

$$\widehat{K(\theta_t)}^\dagger = \left(\frac{1}{B}\sum_{i=1}^{B} \psi_\theta(s_{t,i}, a_{t,i})\psi_\theta^\top(s_{t,i}, a_{t,i}) + \xi I\right)^{-1}, \tag{8}$$

where $\xi > 0$ is a hyperparameter that guarantees the estimator is numerically stable. To sample these without bias from the occupancy measure d_π^ρ, we sample trajectories with length $\frac{1}{1-\sqrt{\gamma}}$ following Algorithm 1 from (Agarwal et al. 2020b). This is equivalent to the geometric sampling found in Algorithms 1 and 2.

Algorithm 1: Policy Gradient for Hölder Smooth Objectives

```
Initial parameter $\theta_0$
for Step $t = 1, \ldots, T-1$ do
    for $i = 1, \ldots B$ do
        Let $j \sim \mathrm{Geom}(1-\gamma), h \sim \mathrm{Geom}(1-\gamma), \tau = j+h$
        Sample $(s_{t,i,0}, a_{t,i,0}, \ldots s_{t,i,\tau}, a_{t,i,\tau})$.
        $s_{t,i} \leftarrow s_{t,i,j}, a_{t,i} \leftarrow a_{t,i,j}$
        $g_{t,i} \leftarrow \sum_{u=j}^{\tau} r_{t,i,u}, r_{t,i,u} \sim R(s_{t,i,u}, a_{t,i,u})$
    end for
    Choose $h_t$ specified in our learning rates section
    $\theta_t \leftarrow \theta_{t-1} + \frac{h_t}{B}\sum_{i=1}^{B} g_{t,i}\psi_{\theta_t}(s_{t,i}, a_{t,i})$
end for
Return $\theta_T$
```

Algorithm 2: Natural Policy Gradient for Hölder Smooth Objectives

```
Initial parameter $\theta_0$, initial matrix $K_0$, stability parameter
  $\xi > 0$
for Step $t = 1, \ldots, T-1$ do
    for $i = 1, \ldots B$ do
        Let $j \sim \mathrm{Geom}(1-\gamma), h \sim \mathrm{Geom}(1-\gamma), \tau = j+h$
        Sample $(s_{t,i,0}, a_{t,i,0}, \ldots s_{t,i,\tau}, a_{t,i,\tau})$.
        $s_{t,i} \leftarrow s_{t,i,j}, a_{t,i} \leftarrow a_{t,i,j}$
        $g_{t,i} \leftarrow \sum_{u=j}^{\tau} r_{t,i,u}, r_{t,i,u} \sim R(s_{t,i,u}, a_{t,i,u})$
    end for
    Choose $h_t$ specified in our learning rates section
    $K_t \leftarrow \frac{1}{B}\sum_{i=1}^{B} \psi(s_{t,i}, a_{t,i})\psi^\top(s_{t,i}, a_{t,i}) + \xi I$
    $\theta_t \leftarrow \theta_{t-1} + K_t^{-1}\frac{h_t}{B}\sum_{i=1}^{B} g_{t,i}\psi_{\theta_t}(s_{t,i}, a_{t,i})$
end for
Return $\theta_T$
```

Learning Rates

In the sequel, we consider the following learning rates: **(i)** constant $h_t = \lambda$, **(ii)** dependent on the number of steps $h_t = \lambda T^{-\frac{\beta_0 - 1}{\beta_0 + 1}}$, **(iii)** decaying $h_t = \lambda t^q, (q \in (-1, 0])$, **(iv)** a

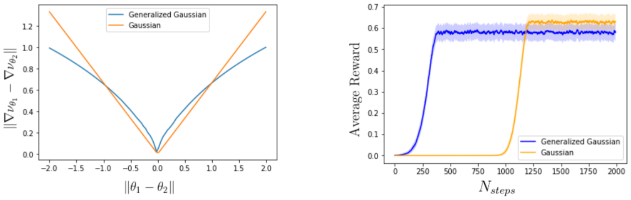

Figure 1: (a) **Tail Growth**: Comparing the growth of ψ_θ in one-dimension for Gaussian policies versus the Generalized Gaussian (Example 1) with $\alpha = 0.5$, for the $[0, 0]$ state in the MountainCar environment. (b) **Exploration Performance**: Comparing the performance of Generalized Gaussian and the standard Gaussian policy, with $\alpha = 0.5$, for the reward function found in Equation (10), $|\theta^* - \theta| = 3.9$. The Generalized Gaussian significantly outperforms during the exploration phase. The result is similar for both PG and NPG.

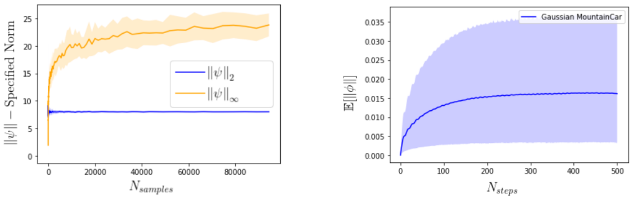

Figure 2: (a) **Gradient Norm Growth**: Comparing the growth of Example 3 using the L_2 norm described by Assumption 2, versus $\max_n \|\psi(s_n, a_n)\|$ with growing number of samples. While our criterion is stable, the $\max$ diverges logarithmically. (b) **Ergodicity of the Test Function**: Convergence in expectation of the test function $\zeta(s, a) = \|\boldsymbol{\phi}(s, a)\|$ for Gaussian policies on the MountainCar environment, using the average over 10000 trajectories, with confidence intervals of the resulting distribution shaded in blue. This large variance impedes practical verification of ergodicity.

learning rate of theoretical interest $h_t = O(\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}})$, which is dependent on the true gradient norm $\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}}$. The last step-size configuration can be estimated using sample data with an additional error contribution; see our Appendices.

Applications

We note two prominant applications of our assumptions: (i) Assumption 1 to exploration has been explicitly shown in (Chou, Maturana, and Scherer 2017), (ii) Assumption 2 has been shown to apply to Safe RL via the work of (Papini, Pirotta, and Restelli 2019). Some additional examples will serve to illustrate these points below.

For ease of demonstration, we consider policies and environments which independently satisfy Assumptions 1-2 and Assumption 3 respectively, so long as the other component is sufficiently regular. The following policies illustrate why we might value weak smoothness:

Example 1. *(Generalized Gaussian Policy) If we choose the parameter $\kappa \in (1, 2]$, we can choose the generalized Gaussian distribution to parameterize our policy:*

$$\nu(a|s, \theta) = -\left|\langle \boldsymbol{\phi}(s, a), \theta\rangle\right|^{\kappa}. \tag{9}$$

Figure 1(a) visualizes this policy's smoothness.

This distribution is covered by our framework; in contrast, previous works only permitted the strictly Gaussian distribution, where $\kappa = 2$. In particular, the tails of this distribution decay much more slowly than the tails of the Gaussian distribution, which has applications to exploration-based strategies. Indeed, let us consider the following single-state exploration problem with the following reward

$$r(a_t) = \left(1 - (a_t - \theta^*)^2\right) \mathbb{1}_{|a_t - \theta^*| \leq 1}, \tag{10}$$

with policies $\nu(a|\theta) = -|a - \theta|^{\kappa}$ for $\kappa = 2$ (a Gaussian policy) and $\kappa \in (1, 2]$ (a generalized Gaussian). $\theta^* \in \mathbb{R}$ is an unknown target. If θ^* is far from our initial parameter, the agent will receive no gradient information so long as it does not sample actions from the region of interest $[\theta^* - 1, \theta^* + 1]$. For a policy with exponent κ, this occurs with probability

$$\begin{aligned} &\mathbb{P}_\kappa(a_t \in [\theta^* - 1, \theta^* + 1]) \\ &\quad = \frac{1}{2\Gamma(\kappa + 1/\kappa)} \int_{\theta^* - 1}^{\theta^* + 1} \exp(-|a - \theta_0|^{\kappa}) da. \end{aligned}$$

Assuming that $|\theta^* - \theta_0| \gg 0$, then if $\mathcal{U} = [\theta^* - 1, \theta^* + 1]$

$$\begin{aligned} &\mathbb{P}_\kappa(a_t \in \mathcal{U}) - \mathbb{P}_2(a_t \in \mathcal{U}) \\ &\geq \frac{1}{2\Gamma(\kappa + 1/\kappa)} \int_{\theta^* - 1}^{\theta^* + 1} \exp(-|a - \theta_0|^{\kappa}) \\ &\qquad - \exp(-|a - \theta_0|^2 + \log 2) da \geq 0, \end{aligned}$$

by simply comparing the terms in the exponents. This difference in probability can improve sample efficiency by many orders of magnitude. The empirical performance of the two policies is found in Figure 1(b), with a large improvement in number of samples needed to discover the correct action. This example can be easily generalized to more complex bandits/MDPs.

Another example shows the richness of the weakly smooth assumption:

Example 2. *(Solutions to p-Laplacian) It is well known (Lindqvist 2017) that solutions to the p-Laplacian*

$$\Delta_p \nu(\theta) \triangleq \nabla \cdot \left(\|\nabla \nu\|^{p-2} \nabla \nu\right) = 0, \tag{11}$$

where $\nabla\cdot$ is the divergence operator, are weakly smooth of order p when $p \in (0, 1]$.

These arise naturally as minimizers of divergence integrals, and thus serve as a useful class of potentials for practical agents; note that we can add any bounded Lipschitz potential to such functions. Weak smoothness has also been shown for many other elliptic families of PDEs (Høeg and Lindqvist 2020; Sciunzi 2014), which can also be candidate policies.

To illustrate the distinction of Assumption 2 from standard $\|\cdot\|_\infty$ bounds, consider the policy class:

Example 3. *(Safe Policies) Consider the following potential for $\theta \in [-1, 1], \boldsymbol{\phi}^* \in \mathbb{R}^d$:*

$$\nu_\theta(s, a) = -\theta \log \|\boldsymbol{\phi}(s, a) - \boldsymbol{\phi}^*\| . \tag{12}$$

Under uniform dynamics and a uniform distribution of $\boldsymbol{\phi}(s, a)$ on $\mathbb{R}^d$, this family satisfies Assumption 2, but not the standard assumption of absolute boundedness $\sup_{s,a} \|\psi_\theta(s, a)\|_\infty < \infty$ (see Figure 2(a)). This policy explicits avoids the state-action region around $\boldsymbol{\phi}^*$; this can arise practically when considering safety or instability constraints in RL.

For some examples of MDPs which exhibit ergodicity, consider the following.

Example 4. *(Simplex MDPs) Suppose the policy has full support on $\mathcal{A}$ for all states. If the feature space is a subset of a d-dimensional simplex $\{\sum_{i=1}^d \boldsymbol{\phi}_i(s, a) = 1, \boldsymbol{\phi}_i \geq 0\}$, then any vector of probability measures $[\boldsymbol{\mu}_1(s), \boldsymbol{\mu}_2(s) \ldots]$ where each $\boldsymbol{\mu}_i(s) \geq c_i$ is lower bounded forms a valid linear MDP. For example, $\boldsymbol{\mu}$ can be uniform in each component. This MDP falls under the assumptions for our Proposition 1.*

Example 5. *(MountainCar) The MountainCar environment, with sufficiently growing slope, empirically obeys ergodicity for regular policy classes such as the generalized Gaussian policy. We can experimentally verify this by computing the geometric convergence of test functions $\mathbb{E}_{s_t, a_t}[\zeta(s_t, a_t)]$, which can be found in Figure 2(b). Note that even for a simple example, this quantity has large variance.*

Note that environments with discontinuous dynamics or unbounded states typically fail ergodicity, but can be preserved if the policy class is finely constrained.

Main Results

Theorem 1. *(Local Convergence) Under Assumptions 1, 2, **Policy Gradient** achieves the following convergence:*

$$\sum_{t=1}^{T} h_t \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq J(\theta_0) - J(\theta_*)$$
$$+ \sum_{t=1}^{T} \frac{C_{PG}}{(1-\gamma)} h_t^{\beta_0+1} \left(\left(\frac{\sigma}{\sqrt{B}}\right)^{\beta_0+1} + \mathbb{E}\left[\|\nabla J(\theta_t)\||\right]^{\beta_0+1} \right),$$

where $\nabla J(\theta_t) = \nabla J(\theta_t)$, and $\sigma = 2\alpha\sqrt{\psi_\infty}$ is the variance of the gradient. $J(\theta_0)$ is our initial performance and J_* is an upper bound on J (which exists due to the boundedness of the reward). B is the batch size and the the remaining constants are specified in the Appendix.

***Natural Policy Gradient** achieves the following convergence:*

$$\sum_{t=1}^{T} h_t \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq (J(\theta_0) - J(\theta_*))$$
$$+ \sum_{t=1}^{T} \frac{C_{NPG}}{(1-\gamma)} h_t^{\beta_0+1} \left(\left(\frac{\sigma}{\sqrt{B}}\right)^{\beta_0+1} + \mathbb{E}\left[\|\nabla J(\theta_t)\|^{\beta_0+1}\right] \right).$$

Remarks: As the norm in Assumption 2 strengthens to $\|\cdot\|_p, p \to \infty$, we can instead take $\beta_0 = \min(\frac{\beta_1}{2}, \beta_2)$ which recovers previous rates. In general the coefficient on β_1 is $r/2$, where $r + \frac{1}{p} = 1$. The case $q = \infty, \beta_0 = 1$ was previously discovered by numerous works, see e.g. (Agarwal et al. 2020b; Xu, Wang, and Liang 2020).

Coefficients C_{PG} and C_{NPG} do not depend on ϵ, γ, and are formally defined in the Appendices. With respect to the ergodicity mixing rate δ, they both scale as $\frac{1}{(1-\delta)^\beta}$, which is analogous to other works with ergodicity (Xu, Wang, and Liang 2020).

Corollary 1. *(Rates under various step-size schemes) Table 2 encapsulates the orders of growth of $\frac{1}{T}\sum_{t=1}^T \|\nabla J(\theta_t)\|^2$ for each of the learning rates examined in our paper. Note that for the optimal learning rate, the bound is instead on the quantity $\frac{1}{T}\sum_{t=1}^T \|\nabla J(\theta_t)\|^{\frac{1+\beta_0}{\beta_0}}$ which $\to 2$ as $\beta_0 \to 1$.*

For global optimality, standard policy gradient requires another assumption in order to demonstrate convergence:

Assumption 4. *(Global Convergence Requirements for Policy Gradient) Let $\theta_1, \theta_2 \in \Theta$ be any two parameterizations for the exponential class ν (recall that $\pi_\theta = C_\theta \exp \nu_\theta$). Then, we assume that ν is dominated, i.e. that the following holds for all a, s:*

$$|\nu_{\theta_1}(a|s) - \nu_{\theta_2}(a|s)| \leq \frac{C_{\theta_1}}{C_{\theta_2}} \log\left(\|\psi_{\theta_2}(s, a)\|\right).$$

Remarks: Thus, we require that the density ν be sub-logarithmic with respect to the gradient $\nabla_\theta \nu(s, a)$. Since ν_θ represents the logits, this equates to a notion of fast growth (outside a local neighbourhood) in θ.

Theorem 2. *(Global Convergence) **Natural Policy Gradient** is bounded with the following rate*

$$J(\pi_*) - \mathbb{E}[J(\theta_t)]$$
$$\leq \frac{C_{NPG,2}}{(1-\gamma)^2} h_t^{\beta_0} \left(\left(\frac{\sigma}{B}\right)^{-\frac{\beta_0+1}{2}} + \mathbb{E}\left[\|\nabla J(\theta_t)\|^{\beta_0+1}\right] \right)$$
$$+ \frac{C_{NPG,3} D_\infty}{(1-\gamma)^{3/2}} \left(\frac{\sigma}{\sqrt{B}} + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_\infty + 1}} + \mathbb{E}\left[\|\nabla J(\theta_t)\||\right] \right).$$

Here, $E_\Pi = \max_{\theta_t} \mathbb{E}_{d^\rho_{\theta_t}}\left[\left\|\psi_{\theta_t}^\top K(\theta_{t-1})^\dagger \nabla J(\theta_t) - A_{\theta_t}\right\|^2\right]$ is a policy dependent parameter that serves to lower bound the optimality of the function class, and $D_\infty = \sqrt{\left\|\frac{Dd_*}{\rho}\right\|_\infty}$ measures the irregularity of the initial distribution.

*If, additionally, Assumption 4 is added, then the standard **Policy Gradient** achieves the following convergence rate:*

$$J(\pi^*) - \mathbb{E}[J(\theta_t)] \leq \frac{1}{1-\gamma} C_{PG,2} \mathbb{E}[\|\nabla J(\theta_t)\||]. \tag{13}$$

$C_{NPG,2}, C_{NPG,3}, C_{PG,2}$ are not dependent on the parameters B, T, γ, and are defined explicitly in the Appendices. We note that for natural policy gradient, there are no additional assumptions apart from the bias term E_Π being finite; this is bounded under weak assumptions (see (Agarwal et al. 2020b)). This is a major advantage of NPG over its vanilla counterpart, which requires a strong additional regularity condition.

For both natural and standard policy gradient, if we take the minimum over $t = 1 \ldots T$, we obtain the rates in the following corollary.

Table 2: Local convergence results of various learning rate schemes, for both policy gradient and natural policy gradient. We only track the primary dependence in T, B, γ. For the decaying learning rate, we define the coefficients $f(q, \beta_0) = \max(\frac{2q\beta_0}{1-\beta_0}, -1)$, $g(q, \beta_0) = \max(q(\beta_0 + 1), -1)$. Note that only the final case generalizes as $\beta_0 \to 1$.

h_t	Order	Considerations
λ	$O(T^{-1}) + O((1-\gamma)^{-1} B^{-\frac{\beta_0+1}{2}}) + O((1-\gamma)^{-\frac{2}{1-\beta_0}})$	Additional Bias Term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$O((1-\gamma)^{-\frac{2}{1-\beta_0}} T^{-\frac{2\beta_0}{1+\beta_0}}) + O((1-\gamma)^{-1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} B^{-\frac{\beta_0+1}{2}})$	
λt^q	$O((1-\gamma)^{-\frac{2}{1-\beta_0}} T^{f(q,\beta_0)}) + O((1-\gamma)^{-1} T^{g(q,\beta_0)} B^{-\frac{\beta_0+1}{2}})$	
$O\left(\Vert\nabla J(\theta_t)\Vert^{\frac{1-\beta_0}{\beta_0}}\right)$	$O((1-\gamma)^{-\frac{1}{\beta_0}} T^{-1}) + O(B^{-\frac{\beta_0+1}{2}})$	Not practical

Corollary 2. *(Rates under various step-size schemes) Under each of the learning rates examined in our paper, we obtain a sample efficiency shown in Table 3 for **Policy Gradient** so that the following holds:*

$$\min_{t=1,\ldots T} J(\pi_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon,$$

*For **Natural Policy Gradient**, the rates are outlined in Table 4 so that the following holds:*

$$\min_{t=1,\ldots T} J(\pi_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon + \frac{C_{NPG,3} D_\infty}{(1-\gamma)^{3/2}} \frac{\sqrt{E_\Pi}}{\sqrt{\psi_\infty + 1}}.$$

The exception is with the constant learning rate $h_t = \lambda$, which contains an additional bias term of order $\lambda^{\frac{\beta_0+1}{2(1-\beta_0)}}(1-\gamma)^{\frac{-1}{1-\beta_0}-\frac{3}{2}}$. For any $\lambda < \frac{1}{1-\gamma}$, this vanishes as $\beta \to 1$.

Remark: The bias term E_Π can be minimized with the appropriate choice of policy class.

Related Work

Optimization and Stochastic Approximation

We primarily refer to work on stochastic approximation, which began with the work by authors (Polyak and Juditsky 1992; Kushner and Yin 2003), who established basic conditions for convergence for linear approximation procedures, with rates being obtained under strong assumptions. Tighter bounds have recently been achieved through improved analysis and techniques (Chen et al. 2016; Lakshminarayanan and Szepesvari 2018; Jain et al. 2018).

The theory for optimizing weakly smooth rather than Lipschitz functionals was primarily developed in the following works (Devolder, Glineur, and Nesterov 2014; Nesterov 2015; Yashtini 2016), introducing the definition of weak-smoothness through Hölder conditions, and showing convergence via smoothing or fast decaying learning rates. Lastly, our analysis relies heavily on the theory of ergodicity for MDPs. We build on the works of (Mitrophanov 2005) which yields perturbation bounds on the state distribution, and subsequent improvements in the assumptions and condition numbers (Ferré, Hervé, and Ledoux 2013; Rudolf, Schweizer et al. 2018).

Reinforcement Learning

The general formulation of reinforcement learning can be attributed to Bellman's formulation of Markov Decision processes (Bellman 1954). Gradient-based approaches were proposed to solve direct policy parameterizations (Williams 1992); developments in this classical setting include (Sutton, Precup, and Singh 1999; Konda and Tsitsiklis 2000; Kakade et al. 2003). These works established asymptotically tight bounds for convergence in the tabular setting, while outlining rough conditions for convergence when feature transformations were applied. The introduction of natural gradient techniques (Kakade 2001), which borrowed from similar work in standard optimization (Amari 1998), yielded improved convergence with respect to policy condition numbers. In particular, strong convergence holds for domains such as the linear quadratic regulator (Fazel et al. 2018; Tu and Recht 2018) and other linearized problems.

Even so, lower bounds for general problems can be quite pessimistic, especially when the conditions are ill-specified (Sutton et al. 2000). This debate has attracted renewed focus in recent years, with an on-going discussion on the quality of representation and its effect on learnability (Du et al. 2019; Van Roy and Dong 2019). Nonetheless, real world problems are either continuous or well-approximated by continuous algorithms, with smooth state-space. (Agarwal et al. 2020a,b) provided a convergence and optimality result for both tabular and linear settings, but only when the action space was discrete and the problem was deterministic. Other results in this setting include (Mei et al. 2021; Zhang et al. 2020; Mei et al. 2020; Zhang et al. 2021). (Xu, Wang, and Liang 2020; Kumar, Koppel, and Ribeiro 2019) focus on general settings, but only under generous smoothness and boundedness assumptions. Numerous works have since focused on feature representations in policy learning, particularly through use of neural networks (Thomas and Brunskill 2017; Wang et al. 2019; Liu et al. 2019); these apply similarly strict assumptions on the problem class in order to achieve good rates of convergence.

We would like to comment extensively on the results of (Liu et al. 2020), which obtains highly competitive rates for PG and NPG, of $O(\epsilon^{-4})$ and $O(\epsilon^{-3})$ respectively. While our rate for NPG is worse at $\to O(\epsilon^{-4}), \beta_0 \to 1$, this is because of numerous differences between our formulations. (Liu et al. 2020) rely on more complex sampling and natural gradient

Table 3: Optimality results of various learning rate schemes, for policy gradient. We only track the primary dependence in ϵ, γ. We omit the decaying learning rate since it yields a detailed and rather uninformative rate.

h_t	T^{-1}	B^{-1}	**Considerations**
λ	$\epsilon^2(1-\gamma)^2$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{6}{1+\beta_0}}$	Bias term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(2-\beta_0)(\beta_0+1)}{(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{4\beta_0-2}{\beta_0+1}}$	
$O\left(\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{\beta_0+2}{\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{2}{\beta_0}}$	Not practical

Table 4: Optimality results of various learning rate schemes, for NPG. We only track the primary dependence in ϵ, γ.

h_t	T^{-1}	B^{-1}	**Considerations**
λ	$\epsilon^2(1-\gamma)^3$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{8}{1+\beta_0}}$	Bias term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5-3\beta_0)(\beta_0+1)}{2(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{6\beta_0-2}{\beta_0+1}}$	
$O\left(\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5+2\beta_0)}{2\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{3}{\beta_0}}$	Not practical

procedures, particularly requiring stochastic gradient descent in order to solve for the NPG update vector. It is unclear whether this technique can generalize to the weakly smooth regime. Instead, we analyze a much simpler algorithm that involves direct estimation of the Fisher information matrix, with an additional cost in ϵ, while also handling non-constant learning rates.

Our results are simultaneously valid for continuous settings, while removing many of the strict assumptions found in previous results. In particular, smoothness of the policy class and boundedness of the gradient severely limited the scope of policies, while state-distribution ergodicity was an opaque condition that could not be easily verified. We build upon work in weakly smooth optimization to relax policy assumptions, while showing an ergodicity result explicitly for near-linear MDPs.

Discussion

In this work, we established the convergence guarantees for the policy gradient for weakly smooth and continuous action space settings. To the best of our knowledge, this is the first work to establish the convergence of policy gradient methods under an unbounded gradient without Lipschitz smoothness conditions. We further established the ergodicity of linear MDPs (under generic integrability assumptions), which was previously assumed to hold by prior work. Thus, our work significantly generalizes the scope of existing analysis while opening lines of future research. Our assumptions are also practically applicable, as we demonstrate through several examples.

Nonetheless, there are many important limitations for our analysis. Firstly, it is likely that Assumption 4 can be significantly relaxed, particularly for the near-linear case which currently contains a supremum on $\mathcal{S} \times \mathcal{A}$. A more careful analysis would have more complex dependence on the problem parameters ϕ, ν. Furthermore, we have yet to consider the case where $L(s) = \infty$, which requires a mix of discrete and continuous analysis. It may also be interesting to consider weaker conditions than geometric ergodicity, by adding regularization conditions on the initial distribution of policies. For practical problems, this is often necessary since the smoothness coefficients can be unbounded except in a reasonable starting set. We also believe that weak smoothness can be relaxed further to locally non-smooth problems ($\beta_0 = 0$), by applying smoothing techniques from optimization (Nesterov 2015). In addition, no practical studies on empirical performance have been done when considering the trade-off between smoothness conditions and convergence rates. Finally, we can quantify the convergence of $J(\theta)$ if θ is sampled stochastically, using functionals such as the KL divergence or Wasserstein metric. This would allow us to determine exact confidence bounds in our results.

Acknowledgements

Animesh Garg is supported in part by CIFAR AI Chair and NSERC Discovery Award, and support from LG and Huawei. Murat A. Erdogdu is partially funded by NSERC Grant [2019-06167], Connaught New Researcher Award, CIFAR AI Chairs program, and CIFAR AI Catalyst grant.

References

Agarwal, A.; Henaff, M.; Kakade, S.; and Sun, W. 2020a. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*.

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2020b. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, 64–66.

Amari, S.-I. 1998. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276.

Bellman, R. 1954. The theory of dynamic programming. Technical report, Rand corp santa monica ca.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2019. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Chen, X.; Lee, J. D.; Tong, X. T.; and Zhang, Y. 2016. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*.

Chou, P.-W.; Maturana, D.; and Scherer, S. 2017. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International conference on machine learning*, 834–843. PMLR.

Deng, Y.; Bao, F.; Kong, Y.; Ren, Z.; and Dai, Q. 2016. Deep direct reinforcement learning for financial signal representation and trading. *IEEE transactions on neural networks and learning systems*, 28(3): 653–664.

Devolder, O.; Glineur, F.; and Nesterov, Y. 2014. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1): 37–75.

Doya, K. 2000. Reinforcement learning in continuous time and space. *Neural computation*, 12(1): 219–245.

Du, S. S.; Kakade, S. M.; Wang, R.; and Yang, L. F. 2019. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? *arXiv preprint arXiv:1910.03016*.

Fazel, M.; Ge, R.; Kakade, S. M.; and Mesbahi, M. 2018. Global convergence of policy gradient methods for linearized control problems.

Ferré, D.; Hervé, L.; and Ledoux, J. 2013. Regular perturbation of V-geometrically ergodic Markov chains. *Journal of applied probability*, 50(1): 184–194.

Høeg, F. A.; and Lindqvist, P. 2020. Regularity of solutions of the parabolic normalized p-Laplace equation. *Advances in Nonlinear Analysis*, 9(1): 7–15.

Jain, P.; Kakade, S.; Kidambi, R.; Netrapalli, P.; and Sidford, A. 2018. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18.

Kakade, S. M. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14: 1531–1538.

Kakade, S. M.; et al. 2003. *On the sample complexity of reinforcement learning*. Ph.D. thesis.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274.

Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014.

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

Kushner, H.; and Yin, G. G. 2003. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Lakshminarayanan, C.; and Szepesvari, C. 2018. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, 1347–1355. PMLR.

Lindqvist, P. 2017. *Notes on the p-Laplace equation*. 161. University of Jyväskylä.

Liu, B.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

Liu, Y.; Zhang, K.; Basar, T.; and Yin, W. 2020. An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods. In *NeurIPS*.

Mei, J.; Gao, Y.; Dai, B.; Szepesvari, C.; and Schuurmans, D. 2021. Leveraging non-uniformity in first-order non-convex optimization. *arXiv preprint arXiv:2105.06072*.

Mei, J.; Xiao, C.; Szepesvari, C.; and Schuurmans, D. 2020. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, 6820–6829. PMLR.

Mitrophanov, A. Y. 2005. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4): 1003–1014.

Nesterov, Y. 2015. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1): 381–404.

Papini, M.; Pirotta, M.; and Restelli, M. 2019. Smoothing policies and safe policy gradients. *arXiv preprint arXiv:1905.03231*.

Polyak, B. T.; and Juditsky, A. B. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4): 838–855.

Rudolf, D.; Schweizer, N.; et al. 2018. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A): 2610–2639.

Sciunzi, B. 2014. Regularity and comparison principles for p-Laplace equations with vanishing source term. *Communications in Contemporary Mathematics*, 16(06): 1450013.

Sidford, A.; Wang, M.; Wu, X.; Yang, L. F.; and Ye, Y. 2018. Near-optimal time and sample complexities for solving discounted Markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211.

Thomas, P. S.; and Brunskill, E. 2017. Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines. *arXiv preprint arXiv:1706.06643*.

Tu, S.; and Recht, B. 2018. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, 5005–5014. PMLR.

Van Roy, B.; and Dong, S. 2019. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*.

Wang, L.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

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

Xu, T.; Wang, Z.; and Liang, Y. 2020. Improving Sample Complexity Bounds for Actor-Critic Algorithms. *arXiv preprint arXiv:2004.12956*.

Yang, L. F.; and Wang, M. 2019. Sample-optimal parametric q-learning using linearly additive features. *arXiv preprint arXiv:1902.04779*.

Yashtini, M. 2016. On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. *Optimization letters*, 10(6): 1361–1370.

Yu, C.; Liu, J.; and Nemati, S. 2019. Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*.

Zhang, J.; Koppel, A.; Bedi, A. S.; Szepesvari, C.; and Wang, M. 2020. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*.

Zhang, J.; Ni, C.; Yu, Z.; Szepesvari, C.; and Wang, M. 2021. On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv preprint arXiv:2102.08607*.

Zou, S.; Xu, T.; and Liang, Y. 2019. Finite-sample analysis for sarsa with linear function approximation. *arXiv preprint arXiv:1902.02234*.