Convergence and Optimality of Policy Gradient Methods in Weakly Smooth Settings

Matthew S. Zhang[1,3], **Murat A. Erdogdu**[1,2,3], **Animesh Garg**[1,3]

[1] Department of Computer Science at the University of Toronto,
[2] Department of Statistical Sciences at the University of Toronto,
[3] Vector Institute for Artificial Intelligence
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 regularity conditions. In this work, we establish explicit convergence rates of policy gradient methods, 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. Notably, our analysis also shows that convergence rates are achievable for both the standard policy gradient and the natural policy gradient algorithms under these assumptions. Lastly we provide performance guarantees for 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 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 practical algorithms. 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). In contrast to value-based methods, direct policy estimators possess numerous advantages, in that they are theoretically insensitive to perturbations in the problem parameters, and typically satisfy better smoothness assumptions. Nonetheless, 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 convergence, (ii) how can we bound *the performance of the policies* under these conditions. Arguably, the convergence of gradient algorithms needs to rely on some constraints of the function class. Prior work has relied on assumptions of (a) smoothness and (b) absolute boundedness of the gradient. However, these conditions are overly restrictive and exclude many useful function approximators.

Summary of Contributions. We make contributions with respect to each of these assumptions: (a) strong smoothness is relaxed to weak smoothness (Hölder conditions); (b) absolute boundedness of the gradient is relaxed to L_2 integrability under the visitation distribution. 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.

Under these assumptions, we find (Corollary 1) that with an appropriate learning rate, the gradients satisfy $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_{t-1})\|^2\right] \leq \epsilon$ for both the standard and natural policy gradient for large enough T, B, where T is the number of iterations and B is the batch size. We also show (Corollary 2) that natural policy gradient satisfies the following bound:

$$\min_{t=0,1,\dots T-1} J(\theta_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon + \mathcal{O}\left(\frac{\sqrt{E_\Pi}}{1-\gamma}\right), \qquad (1)$$

for large enough T, B, where E_Π can be tuned by choosing an appropriately regular policy class and θ_* is the maximizer of J. E_Π is formally defined in §5. Under a strong additional assumption, standard policy gradient also satisfies $\min_{t=0,1\dots T-1} J(\theta_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon$ for large enough T, B. In the strictly smooth limit these results 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 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. By abuse of notation, we use the same notation for a measure and its density, unless otherwise specified. Let $\|z\| = \|z\|_2$ the 2-norm for vectors, $\|z\|_{op}$ the operator norm for matrices, and $\|p-q\|_{TV} \triangleq \int |p(x) - q(x)|\, dx$ the total variation distance for signed measures. 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: For a given state $s \in \mathcal{S}$, we denote a stochastic policy with $\pi(\cdot|s)$, which is a probability distribution over $\mathcal{A}$.

Trajectories: To generate trajectories, we start from an initial state distribution ρ, and then at each time $t \in \mathbb{N}$, we sample an action from the policy: $a_t \sim \pi(\cdot|s_t)$. Subsequently a state and reward are queried as $s_{t+1} \sim P(\cdot|a_t, s_t), r_t \sim R(\cdot|a_t, s_t)$, and the process continues indefinitely. Consequently $\pi, \rho, \mathcal{M}$ together parameterize a probability distribution on the set of trajectories. Letting ρ be fixed, we write this as $\{(s_t, a_t), t = 0, 1, 2, \ldots\} \sim$ MDP following π.

Value Functions: 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 following the policy π. If $|r_t| \leq \alpha$ almost surely for all t, both functions are bounded by $[-\alpha/(1-\gamma), \alpha/(1-\gamma)]$ almost surely as well. 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 probability densities through the following: $d_\pi^{s'}(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p_t(s, a | s_0 = s')$ where $p_t(s, a|s_0 = s')$ is the conditional probability density for s, a being sampled at time t from the MDP following π, given the initial state $s_0 = s'$. We overload notation and write $d_\pi^\rho(s,a) = \int d_\pi^{s'}(s,a)\rho(s')\, ds'$. This defines a probability density function on $\mathcal{S} \times \mathcal{A}$. We also write $H_\theta^\rho(s) = \int d_\theta^\rho(s,a)\, da$ for the state-component of the visitation distribution.

Reinforcement Learning: A reinforcement learning agent is one which produces a sequence of policies π_t based on the queries from the MDP, and seeks to iteratively maximize the value function: $J(\pi) = \int V_\pi(s)\rho(s)\, ds$. The existence of an optimum in the space of stochastic functions has been shown as a classical result (Bellman 1954).

Algorithms

Policy Class

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 $\theta \in \Theta \subseteq \mathbb{R}^N$, such that $\pi_{\nu_\theta}(a|s) = \frac{\exp(\nu_\theta(s,a))}{\int_\mathcal{A} \exp(\nu_\theta(s,a))\, da}$ where $\nu_\theta : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$. We require that the integral $\int_\mathcal{A} \exp(\nu_\theta(s,\cdot))\, da < \infty$ 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 $\pi_\theta \triangleq \pi_{\nu_\theta}$, $J(\theta) \triangleq J(\pi_\theta)$ and use θ instead of π_θ in subscripts where there is no confusion. Let us denote the score function as $\psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s)$. Then the gradient can be written as $\nabla J(\theta) = \mathbb{E}_{s,a\sim d_\theta^\rho}[Q_\theta(s,a)\psi_\theta(s,a)]$. 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 and possibly uncountable. Typically some form of well-chosen 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:*

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

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

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, a. Consequently we define $\beta_0 = \min(\beta_1/4, \beta_2)$ as the dominant order of smoothness.

It will also be useful in our analysis to define $\beta_{max} = \max(\beta_1/4, \beta_2)$. 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 score function.

Remarks: $\beta_1 < 2, \beta_2 < 1$ are weakly smooth cases. This is a weaker assumption than traditional assumptions on Lipschitz smoothness; particularly, it allows for slow tail decay. It is also possible to relax this assumption to local conditions (i.e. only holding when $\|\eta\| \leq C$), while having Lipschitz conditions at large scales.

We introduce an additional assumption on the second moment of the score function:

Assumption 2. *(Boundedness of Moments) Assume that the score function is absolutely bounded in L_2 across all policies i.e. that the following holds for all θ*

$$\int_\mathcal{S}\int_\mathcal{A} \|\psi_\theta(s,a)\|^2 d_\theta^\rho(s,a) da\, ds \leq \psi_\infty, \tag{4}$$

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

Remarks: Higher order integrability assumptions are possible. In fact, if $\|\psi_\theta\| \leq \sqrt{\psi_\infty}$ holds d_θ^ρ-almost surely, we recover the standard bounded gradient assumption found in other works (Xu, Wang, and Liang 2020; Liu et al. 2020).

Finally, we require the following standard assumption (see e.g. Xu, Wang, and Liang (2020); Zou, Xu, and Liang (2019)) which we use to show smoothness of the objective function.

Algorithm 1: Policy Gradient for Hölder Smooth Objectives

```
Initial parameter $\theta_0$.
for Step $t = 1, \dots, T$ do
    for $i = 1, \dots B$ do
        Let $j \sim$ Geom$(1-\gamma)$, $h \sim$ Geom$(1-\gamma^{1/2})$, $\tau = j + h$.
        Sample $(s_0, a_0, \dots s_\tau, a_\tau) \sim$ MDP following $\pi_{\theta_{t-1}}$.
        $s_{t,i} \leftarrow s_j, a_{t,i} \leftarrow a_j$.
        $v_{t,i} \leftarrow \sum_{u=j}^{\tau} \gamma^{(u-j)/2} r_u$, $r_u \sim R(\cdot|s_u, a_u)$.
    end for
    Choose $h_t$ specified in our learning rates section.
    $\theta_t \leftarrow \theta_{t-1} + \frac{h_t}{B} \sum_{i=1}^{B} v_{t,i} \psi_{\theta_{t-1}}(s_{t,i}, a_{t,i})$.
end for
Return $\theta_T$
```

Algorithm 2: Natural Policy Gradient for Hölder Smooth Objectives

```
Initial parameter $\theta_0$, stability parameter $\xi \in (0, 1]$.
for Step $t = 1, \dots, T$ do
    for $i = 1, \dots B$ do
        Let $j \sim$ Geom$(1-\gamma)$, $h \sim$ Geom$(1-\gamma^{1/2})$, $\tau = j + h$.
        Sample $(s_0, a_0, \dots s_\tau, a_\tau) \sim$ MDP following $\pi_{\theta_{t-1}}$.
        $s_{t,i} \leftarrow s_j, a_{t,i} \leftarrow a_j$.
        $v_{t,i} \leftarrow \sum_{u=j}^{\tau} \gamma^{(u-j)/2} r_u$, $r_u \sim R(\cdot|s_u, a_u)$.
    end for
    for $i = 1, \dots B$ do
        Let $j \sim$ Geom$(1-\gamma)$.
        Sample $(s'_0, a'_0, \dots s'_j, a'_j) \sim$ MDP following $\pi_{\theta_{t-1}}$.
        $s'_{t,i} \leftarrow s'_j, a'_{t,i} \leftarrow a'_j$.
    end for
    Choose $h_t$ specified in our learning rates section.
    $K_t \leftarrow \frac{1}{B} \sum_{i=1}^{B} \psi_{\theta_{t-1}}(s'_{t,i}, a'_{t,i}) \psi_{\theta_{t-1}}^\top(s'_{t,i}, a'_{t,i})$.
    $M_t \leftarrow (K_t + \xi I)^{-1}$
    $\theta_t \leftarrow \theta_{t-1} + \frac{h_t}{B} \sum_{i=1}^{B} M_t v_{t,i} \psi_{\theta_{t-1}}(s_{t,i}, a_{t,i})$.
end for
Return $\theta_T$
```

Assumption 3. *(Ergodicity) We have for all states $s_0 \in \mathcal{S}$:*

$$\|\mathbb{P}_\theta^n(\cdot|s_0) - \rho_*(\cdot)\|_{TV} \leq C_0 \delta^n,$$

where $\mathbb{P}_\theta^n$ 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_0, θ.

Policy Gradient

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), applies the following update

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

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 advantage of this method is that the optimization landscape becomes well-behaved.

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 for ∇J and K:

$$\widehat{\nabla J(\theta_{t-1})} = \frac{1}{B} \sum_{i=1}^{B} v_{t,i} \psi_{\theta_{t-1}}(s_{t,i}, a_{t,i}), \tag{7}$$

$$K_t = \frac{1}{B} \sum_{i=1}^{B} \psi_{\theta_{t-1}}(s_{t,i}, a_{t,i}) \psi_{\theta_{t-1}}^\top(s_{t,i}, a_{t,i}), \tag{8}$$

and we use $(K_t + \xi I)^{-1}$ to approximate the inverse of the Fisher matrix, where $\xi \in (0, 1]$ is a parameter that guarantees the estimator is numerically stable, $v_{t,i}$ is an unbiased estimator for $Q_{\theta_{t-1}}(s, a)$ given in Algorithms 1-2 wherein we follow Zhang et al. (2020b, Algorithm 1), and to sample from the occupancy measure d_θ^ρ, we sample trajectories following Agarwal et al. (2020b, Algorithm 1). This procedure is summarized in Algorithms 1-2.

Learning Rates

In the sequel, we consider the following learning rates: **(i)** constant $h_t = \lambda$, **(ii)** dependent on the total number of steps $h_t = \lambda T^{\frac{\beta_0 - 1}{\beta_0 + 1}}$, **(iii)** decaying $h_t = \lambda t^{-q}, q \in [0, 1)$. We also state our Theorems 1-2 more generally for any step size sequence h_t.

Applications

We note two prominant applications of our assumptions: (i) an application of 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:*

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

See Figure 1(a) for a visualization of the smoothness of this policy.

This distribution is covered by our assumptions; 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.

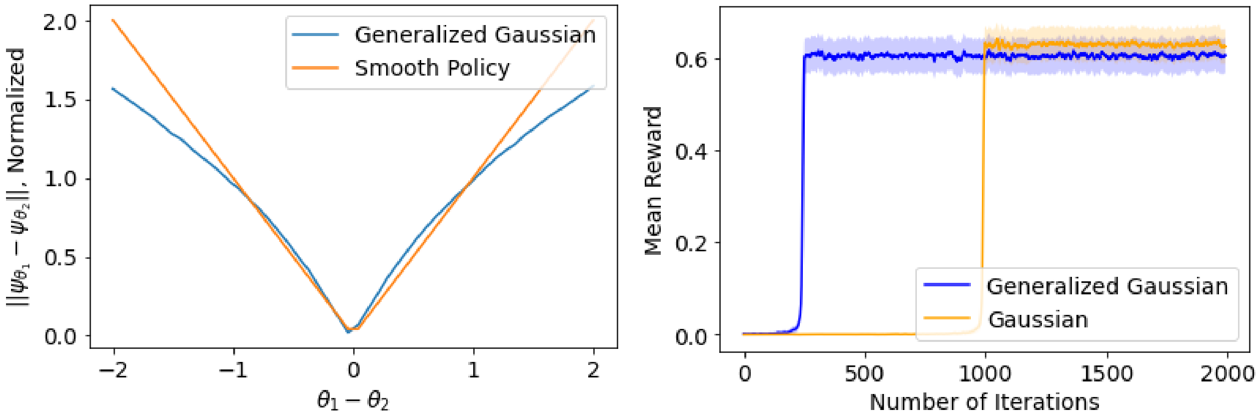

Figure 1: (a) Tail Growth: Comparing the growth of ψ_θ in one-dimension for a hypothetical policy class with 1-Lipschitz gradient, versus the Generalized Gaussian (Example 1) with $\alpha = 0.1$, for the $[0,0]$ state in the MountainCar environment. The gradients are normalized for ease of comparison. (b) Exploration Performance: Comparing the performance of Generalized Gaussian and the standard Gaussian policy, with $\alpha = 0.7$, for the reward function found in Equation (10), $|\theta^* - \theta| = 3.3$. The Generalized Gaussian significantly outperforms during the exploration phase. The result is similar for both PG and NPG.

Indeed, let us consider the following single-state exploration problem with the following (deterministic) reward

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

with policies $\nu_\theta(a) = -|a-\theta|^\kappa$ 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

$$\pi_{\kappa,\theta_0}(a_t \in [\theta^*-1,\theta^*+1]) \\ = \frac{1}{2\Gamma(\kappa+1/\kappa)} \int_{\theta^*-1}^{\theta^*+1} \exp(-|a-\theta_0|^\kappa) da,$$

where π_{κ,θ_0} is the policy measure under the generalized Gaussian with exponent κ and parameter θ_0. If $\mathcal{U} = [\theta^*-1, \theta^*+1]$

$$\pi_{\kappa,\theta_0}(a_t \in \mathcal{U}) - \pi_{2,\theta_0}(a_t \in \mathcal{U}) \\ \ge \frac{1}{2\Gamma(\kappa+1/\kappa)} \int_{\theta^*-1}^{\theta^*+1} \exp(-|a-\theta_0|^\kappa) \\ - \exp(-|a-\theta_0|^2 + \log 2) da,$$

which is > 0 by simply comparing the terms in the exponents, when $\kappa \ll 2$ and $|\theta^* - \theta_0| \gg 0$. 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. *(p-Harmonic minimizers) It is known (Coscia and Mingione 1999; Lindqvist 2017) that local minimizers ν to the p-Harmonic functional, for $p(x) : \mathbb{R}^d \mapsto \mathbb{R}$*

$$\mathcal{F}(\nu) \triangleq \int_{\mathbb{R}^d} \|\nabla \nu\|^{p(x)} \, dx, \qquad (11)$$

are weakly smooth of some order $L(p) < 1$ when $p(x) > 1$.

One can also restrict the integration above to a compact set. Consequently, these can serve as interesting potential functions. Note that we can add any potential with bounded and Lipschitz gradient to such functions while preserving Hölder regularity. Weak smoothness has also been shown for many other elliptic families of PDEs (Høeg and Lindqvist 2020; Sciunzi 2014), which may also motivate some candidate policies.

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

Example 3. *(Safe Policies) Consider the following potential for $\theta \in [-1,1]$, $\|\phi^*\| \le 1$:*

$$\nu_\theta(s,a) = -\theta \log \|\phi(s,a) - \phi^*\|. \qquad (12)$$

Under uniform dynamics and a uniform distribution of $\phi(s,a)$ on a ball of radius 1 around the origin, this family satisfies Assumption 2, but not the standard assumption of absolute boundedness $\sup_{s,a} \|\psi_\theta(s,a)\|_\infty < \infty$. This policy explicits avoids the state-action region around ϕ^*; this can arise practically when considering safety or instability constraints in RL.

Main Results

In the sequel, define the quantity E_Π as $\max_{\theta\in\Theta} \mathbb{E}_{s,a\sim d_\theta^\rho}\left[\left\|\psi_\theta(s,a)^\top K(\theta)^\dagger \nabla J(\theta) - A_{\pi_\theta}(s,a)\right\|^2\right]$ and the quantity $D_\infty = \sup_{\theta_1,\theta_2} \left\|\frac{d_{\theta_1}^\rho}{d_{\theta_2}^\rho}\right\|_\infty + 1$. For brevity, we will let $\Sigma = \frac{\sigma}{(1-\gamma)\sqrt{B}}$, where $\sigma = 3\alpha\sqrt{\psi_\infty}$ controls the variance of the gradient and B is the batch size.

Theorem 1. *Under Assumptions 1-3, **Policy Gradient** and **Natural Policy Gradient** achieves the following convergence:*

$$\begin{aligned}&\sum_{t=1}^{T} h_t \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \leq C_{k,1}\left(J_* - J(\theta_0)\right)\\&\quad + \frac{C_{k,2}}{1-\gamma}\sum_{t=1}^{T}\Bigg(h_t^{\frac{\beta_1}{4}+1}\left(\mathbb{E}\left[\|\nabla J(\theta_t)\|^{\frac{\beta_1}{4}+1}\right] + \Sigma^{\frac{\beta_1}{4}+1}\right)\\&\quad + h_t^{\beta_2+1}\Big(\mathbb{E}\left[\|\nabla J(\theta_t)\|^{\beta_2+1}\right] + \Sigma^{\beta_2+1}\Big)\Bigg),\end{aligned}$$

where $C_{k,1}, C_{k,2}$ depend on whether policy gradient or natural policy gradient is considered, and are defined in the Appendix. They do not depend on ϵ, γ for either algorithm. $J(\theta_0)$ is the initial performance and $J_* = \sup_{\theta\in\Theta} J(\theta)$, which is finite due to the boundedness of the reward. The remaining constants are specified in the Appendices.

Remarks: If we replace Assumption 2 with an almost sure bound on $\|\psi_\theta\|$, the exponent $\beta_1/4+1$ becomes instead $\beta_1/2+1$, which recovers previous results.

With respect to the ergodicity mixing rate δ, $C_{k,2}$ scales as $1/(1-\delta)$ for either algorithm, which is analogous to other works with ergodicity (Xu, Wang, and Liang 2020, Proposition 1).

Corollary 1. *(Rates under various step-size schemes) Table 1 encapsulates the orders of growth of $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right]$ for each of the learning rates examined in our paper, for the choice of λ sufficiently small and $B \gtrsim \frac{\sigma^2}{(1-\gamma)^2}$.*

For our subsequent results, policy gradient requires another opaque assumption:

Assumption 4. *(Requirements for Policy Gradient) Assume that there is a $\theta_* \in \Theta$ where J attains its maximum. Furthermore, let $\theta \in \Theta$ be any parameter. Then, we assume that J is m-dominated for any $m > 0$, i.e. that the following holds*

$$J(\theta_*) - J(\theta) \leq \frac{m}{1-\gamma}\left\langle \theta_* - \theta, \nabla J(\theta)\right\rangle.$$

Assume Diam$(\Theta) \triangleq \sup_{\theta_1,\theta_2\in\Theta}\|\theta_1 - \theta_2\| < \infty$ as well.

See (Bhandari and Russo 2019, Lemma 3(a)) for analogous conditions, which are often violated in practice.

Theorem 2. *Let $\theta_* = \arg\max_{\theta\in\Theta} J(\theta)$. Under Assumptions 1-3, **Natural Policy Gradient** is bounded with the following for $t \in 1\ldots T$:*

$$\begin{aligned}&J(\theta_*) - \mathbb{E}\left[J(\theta_{t-1})\right]\\&\leq \frac{C_{NPG,2}}{1-\gamma}h_t^{\beta_1-1}\left(\left(\xi^{-1}\Sigma\right)^{\beta_1} + \mathbb{E}\left[\|\nabla J(\theta_{t-1})\|^{\beta_1}\right]\right)\\&\quad + \frac{C_{NPG,3}}{1-\gamma}h_t^{\beta_2}\Big(\left(\xi^{-1}\Sigma\right)^{\beta_2+1} + \mathbb{E}\left[\|\nabla J(\theta_{t-1})\|^{\beta_2+1}\right]\Big)\\&\quad + \frac{C_{NPG,4}}{1-\gamma}\left(\xi^{-1}\Sigma + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_\infty}} + \frac{1}{\xi}\mathbb{E}\left[\|\nabla J(\theta_{t-1})\|\right]\right).\end{aligned}$$

Here, E_Π is a policy dependent parameter, σ is the variance from Theorem 1, and ξ is the stability constant found in Algorithm 2.

*If, additionally, Assumption 4 is added, then the standard **Policy Gradient** is bounded by the following for $t \in 1\ldots T$:*

$$J(\theta_*) - \mathbb{E}[J(\theta_{t-1})] \leq \frac{m\mathit{Diam}(\Theta)}{1-\gamma}\mathbb{E}\left[\|\nabla J(\theta_{t-1})\|\right], \quad (13)$$

where Diam(Θ) is defined in Assumption 4.

Here, θ_* is the minimizer from Assumption 4, and $C_{NPG,2-4}$ are not dependent on h_t, B, T, γ and are stated explicitly in the appendices. For natural policy gradient, there are no additional assumptions apart from the bias term E_Π being finite; this is bounded under standard assumptions (see Agarwal et al. (2020b, Remark 6.4)). 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.

Corollary 2. *For λ sufficiently small, $B \gtrsim \frac{\sigma^2}{(1-\gamma)^2}$, and the learning rate $h_t = \lambda T^{\frac{\beta_0-1}{\beta_0+1}}$, for **Policy Gradient** the following holds under Assumptions 1-4*

$$\min_{t=0,\ldots T-1} J(\theta_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon.$$

*For **Natural Policy Gradient** the following holds under Assumptions 1-3*

$$\min_{t=0,\ldots T-1} J(\theta_*) - \mathbb{E}\left[J(\theta_t)\right] \leq \epsilon + \frac{\sqrt{D_\infty E_\Pi}}{1-\gamma}.$$

Recall that E_Π is an approximation error, and D_∞ measures the irregularity of the initial distribution. For either algorithm, we need to choose

$$\begin{aligned}T &\gtrsim \epsilon^{-\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{-\frac{2\beta_0^2+3\beta_0+1}{2\beta_0^2}},\\B &\gtrsim \epsilon^{-2}(1-\gamma)^{-\frac{4\beta_0^2+5\beta_0-1}{\beta_0(\beta_0+1)}},\end{aligned}$$

where we only track dependencies on γ, ϵ.

Please refer to the extended version of our paper at https://arxiv.org/abs/2111.00185 for the technical appendices containing proofs of these statements.

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, both in asymptotic and non-asmyptotic contexts (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

h_t	Order	Considerations
λ	$O(\lambda^{-1}T^{-1}) + O((\sqrt{B}(1-\gamma))^{-\beta_0-1}) + \text{Bias}$	Additional Bias
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$O(\lambda^{-1}T^{-\frac{2\beta_0}{1-\beta_0}}) + O(T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}}(\sqrt{B}(1-\gamma))^{-\beta_0-1})$	
λt^{-q}	$\tilde{O}(\lambda^{-1}T^{-f(q,\beta_0)}) + O(T^{-q\beta_0}(\sqrt{B}(1-\gamma))^{-\beta_0-1})$	

Table 1: Results for 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 coefficient $f(q, \beta_0) = \min(\frac{2q\beta_0}{1-\beta_0}, 1-q)$. In each case we require $\lambda^{\beta_0} \lesssim \frac{1-\gamma}{C}$ where C does not depend on γ, ϵ.

(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; Mao and Song 2020).

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. (2020a); 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 $O(\epsilon^{-4}), \beta_0 \to 1$, this is because of numerous differences between our formulations. Liu et al. (2020) rely on more complex sampling and natural gradient 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 limited the scope of policies. We build upon work in weakly smooth optimization to relax these assumptions.

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. Thus, our work significantly generalizes the scope of existing analysis while opening numerous 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, as in other recent work (Liu et al. 2020). A more careful analysis would have more complex dependence on the problem parameters ϕ, ν. It may also be interesting to consider weaker assumptions than 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 the distribution of $J(\theta)$ using functionals such as the KL divergence or Wasserstein metric.

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.

Bhandari, J.; and Russo, D. 2019. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

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.

Coscia, A.; and Mingione, G. 1999. Hölder continuity of the gradient of p (x)-harmonic mappings. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(4): 363–368.

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.; and Mesbahi, M. 2018. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1467–1476. PMLR.

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, Gatsby Computational Neuroscience Unit, University College London.

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*.

Mao, Y.; and Song, Y. 2020. Perturbation theory and uniform ergodicity for discrete-time Markov chains. *arXiv preprint arXiv:2003.06978*.

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. 2020a. 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*.

Zhang, K.; Koppel, A.; Zhu, H.; and Basar, T. 2020b. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612.

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