Convergence and Optimality of Policy Gradient Methods in Weakly Smooth Settings - Extended Abstract

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

1. Introduction

Modern Reinforcement Learning (RL) has solved challenges in diverse fields such as finance, healthcare, and robotics (Deng et al., 2016; Yu et al., 2019; Kober et al., 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.

Theoretical results in continuous domains do not effectively characterize practical algorithms. Recently, some papers have proposed the linear class of Markov Decision Processes (MDPs), where the transition kernel and reward function take the form of inner products of arbitrary feature transforms. In particular, value-based estimators have obtained strong results in this context, both in on- and off-line settings (Cai et al., 2019; Yang & 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 are smoother to estimate. 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 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 relax all three of these assumptions significantly: (a) ergodicity is proved by relying on near-linearity of the problem dynamics; (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. In addition, our conditions are significantly easier to verify through numerical simulation as they are direct constraints on either the policy or the MDP. 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.

1.1. Policy Class

Let $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ denote an MDP, where P is linear in the sense that $P(\cdot|s,a) = \langle \phi(s,a), \boldsymbol{\mu}(\cdot) \rangle$ for some feature transformation $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. 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 function $\nu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, as: $\pi_\nu(a|s) = \frac{\exp(\nu(s,a))}{\int_{\mathcal{A}} \exp(\nu(s,a))}$. We require that ν is differentiably parameterized by a variable $\theta \in \Theta$ (where $\Theta \subseteq \mathbb{R}^N$).The gradient can be written as $\nabla J(\theta) = \mathbb{E}_{(s,a)\sim d^\rho_\pi}[Q_{\pi_\theta}(s,a)\psi_\theta(s,a)]$ due to the value function having an expectation of zero. Let us denote the score function as $\psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s)$. In this work, we consider all softmax functions that satisfy the following smoothness properties:

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

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

[1]Department of Computer Science at the University of Toronto, Toronto, Canada [2]Vector Institute, Toronto, Canada. Correspondence to: Matthew Shunshi Zhang <matthew.zhang@mail.utoronto.ca>.

Algorithm 1 Policy Gradient for Hölder Smooth Objectives

Initial parameter θ_0
for Step $t = 1, \dots, T-1$ **do**
  for $i = 1, \dots B$ **do**
    Sample $s_{t,i}, a_{t,i} \sim d_{\pi_t}^{\rho_0}$, $r_{t,i} \sim R(\cdot|s_{t,i}, a_{t,i})$
  end for
  Choose h_t according to a specified rule.
  $\theta_t \leftarrow \theta_{t-1} + \frac{1}{B}\sum_{n=1}^{B} h_t r_{t,i} \psi(s_{t,i}, a_{t,i})$
end for
Return θ_T

$$\int_{\mathcal{A}} \|\nabla \nu_\theta(s,a) - \nabla \nu_{\theta+\eta}(s,a)\| \, \pi_\theta(a|s) \leq C_{\nu,2} \|\eta\|^{\beta_2}, \quad (2)$$

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.

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:*

$$\int \|\psi_\theta(s,a)\|_2^2 \, d_\theta^\rho(s,a) \leq \psi_\infty \quad (3)$$

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

Finally, under weak assumptions on the policy and MDP, we prove the following , 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.

Proposition 1 *(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.

1.2. 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

$$\theta_t = \theta_{t-1} + h_t \nabla_\theta J(\theta_{t-1}), \quad (4)$$

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

$$\theta_t = \theta_{t-1} + h_t K^\dagger(\theta) \nabla_\theta J(\theta_{t-1}), \quad (5)$$

Algorithm 2 Natural Policy Gradient for Hölder Smooth Objectives

Initial parameter θ_0, initial matrix K_0, stability parameter $\xi > 0$
for Step $t = 1, \dots, T-1$ **do**
  for $i = 1, \dots B$ **do**
    Sample $s_{t,i}, a_{t,i} \sim d_{\pi_t}^{\rho_0}$, $r_{t,i} \sim R(\cdot|s_{t,i}, a_{t,i})$
  end for
  Choose h_t according to a specified rule.
  $K_t \leftarrow \frac{1}{B}\sum_{i=1}^{B} \psi(s_{t,i}, a_{t,i}) \psi^\top(s_{t,i}, a_{t,i})$
  $\theta_t \leftarrow \theta_{t-1} + \frac{1}{B}\sum_{i=1}^{B} h_t (K_t + \xi I)^{-1} r_{t,i} \psi(s_{t,i}, a_{t,i})$
end for
Return θ_T

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 estimators:

$$\widehat{\nabla_\theta J(\theta_t)} = r_t \psi_\theta(s_t, a_t), \quad (6)$$

$$\widehat{K(\theta_t)^\dagger} = \left(\psi_\theta(s_t,a_t)\psi_\theta^\top(s_t,a_t) + \xi I\right)^{-1}, \quad (7)$$

where $\xi > 0$ is a hyperparameter that guarantees the estimator is numerically stable. The first estimator is unbiased while ξ controls the bias of the second term; this bias vanishes as $\xi \to 0$.

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)** an optimal learning rate $h_t = O(\|\boldsymbol{g}_t\|^{\frac{1-\beta_0}{\beta_0}})$.

1.3. Main Results

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

$$\sum_{t=1}^{T} h_t \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right] \leq J(\theta_0) - J(\theta_*)$$
$$+ \sum_{t=1}^{T} \frac{C}{(\beta_0+1)(1-\gamma)} h_t^{\beta_0+1} \left(\tilde{\sigma} B^{-\frac{\beta_0+1}{2}} + \mathbb{E}\left[\|\boldsymbol{g}_t\|\right]^{\beta_0+1}\right),$$

where $\boldsymbol{g}_t = \nabla J(\theta_t)$, $\beta = \beta_0/(\beta_0+1) \leq 0.5$. $J(\theta_0)$ 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 remaining constants are specified in the Appendix.

Table 1. 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
λ	$O(T^{-1} + (1-\gamma)^{-1}B^{-\frac{\beta_0+1}{2}} + (1-\gamma)^{-\frac{2}{1-\beta_0}})$
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$O((1-\gamma)^{-\frac{2}{1-\beta_0}}T^{-\frac{2\beta_0}{1+\beta_0}}) + (1-\gamma)^{-1}T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}}B^{-\frac{\beta_0+1}{2}})$
λt^q	$O((1-\gamma)^{-\frac{2}{1-\beta_0}}T^{f(q,\beta_0)} + (1-\gamma)^{-1}T^{g(q,\beta_0)}B^{-\frac{\beta_0+1}{2}})$
$O\left(\|\boldsymbol{g}_t\|^{\frac{1-\beta_0}{\beta_0}}\right)$	$O((1-\gamma)^{-\frac{1}{\beta_0}}T^{-1} + B^{-\frac{\beta_0+1}{2}})$

Natural Policy Gradient achieves the following:

$$\sum_{t=1}^{T} h_t \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right] \le (\psi_\infty + \xi)^2 (J(\theta_0) - J(\theta_*)) + \sum_{t=1}^{T} \frac{L(\psi_\infty+\xi)^2}{(\beta_0+1)\xi^{\beta_0+1}} h_t^{\beta_0+1}\left(\tilde{\sigma}B^{-\frac{\beta_0+1}{2}} + \mathbb{E}\left[\|\boldsymbol{g}_t\|^{\beta_0+1}\right]\right).$$

Remarks: As the norm in Assumption 2 strengthens to $\|\cdot\|_q, q \to \infty$, 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 $r + \frac{1}{q} = 1$. The case $q = \infty, \beta_0 = 1$ was previously discovered by numerous works, see e.g. (Agarwal et al., 2020b; Xu et al., 2020).

Corollary 1 *(Rates under various step-size schemes) Table 1 encapsulates the orders of growth of $\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2$ for each of the learning rates examined in our paper. Note that for the optimal learning rate, we must instead bound $\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^{\frac{1+\beta_0}{\beta_0}}$.*

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

Assumption 3 *(Global Convergence Requirements for Policy Gradient) Let $\theta_1, \theta_2 \in \Theta$ be any two parameterizations for the exponential class ν (recall that $\pi_\theta = \exp\nu_\theta$). Then, we assume that ν is dominated, i.e. that the following holds for all a, s:*

$$|\nu_{\theta_1}(a|s) - \nu_{\theta_2}(a|s)| \le \log\left(\left\|\nabla_\theta \nu_{\theta_2}(s,a) - \mathbb{E}_{a\sim\pi_{\theta_2}(s)}[\nabla_\theta \nu_{\theta_2}(s,\cdot)]\right\|_2\right).$$

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 achieves the following convergence rate:*

$$J(\pi_*) - \mathbb{E}\left[J(\theta_t)\right] \le \sqrt{\frac{C_{21}}{(1-\gamma)^3}}\left(O(B^{-1/2}) + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_\infty+1}} + O(\|\boldsymbol{g}_t\|)\right).$$

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 $C_{21} = \sqrt{(\psi_\infty+1)\left\|\frac{Dd_*}{\rho}\right\|_\infty}$ measures the irregularity of the initial distribution. If, additionally, Assumption 3 is added, then the standard Policy Gradient achieves the following convergence rate:

$$J(\pi^*) - \mathbb{E}[J(\theta_t)] \le \frac{1}{1-\gamma} C_{21}\|\boldsymbol{g}_t\|. \tag{8}$$

We note that 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)).

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

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 2 for policy gradient so that the following holds:*

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

For natural policy gradient, the rates are outlined in Table 3 so that the following holds:

$$\min_{t=1,\dots T} J(\pi_*) - \mathbb{E}\left[J(\theta_t)\right] \le \epsilon + \sqrt{\frac{C_{21}}{(1-\gamma)^3}}\frac{\sqrt{E_\Pi}}{\sqrt{\psi_\infty+1}}$$

The exception is with the constant learning rate, which contains an additional bias term of order $\lambda^{\frac{\beta_0+1}{2(1-\beta_0)}}(1-\gamma)^{\frac{-1}{1-\beta_0}-\frac{3}{2}}$.

1.4. Applications

For ease of demonstration, we consider policies and environments which independently satisfy Assumptions 1-2 and ergodicity 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:*

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

Table 2. 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 only cumbersome results.

h_t	T^{-1}	B^{-1}
λ	$\epsilon^2(1-\gamma)^2$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{6}{1+\beta_0}}$
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(2-\beta_0)(\beta_0+1)}{(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{4\beta_0-2}{\beta_0+1}}$
$O\left(\|\boldsymbol{g}_t\|^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{\beta_0+2}{\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{2}{\beta_0}}$

Table 3. Optimality results of various learning rate schemes, for NPG. We only track the primary dependence in ϵ, γ.

h_t	T^{-1}	B^{-1}
λ	$\epsilon^2(1-\gamma)^3$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{8}{1+\beta_0}}$
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5-3\beta_0)(\beta_0+1)}{2(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{6\beta_0-2}{\beta_0+1}}$
$O\left(\|\boldsymbol{g}_t\|^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5+2\beta_0)}{2\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{3}{\beta_0}}$

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

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

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

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 $[\theta^* - 1, \theta^* + 1]$.

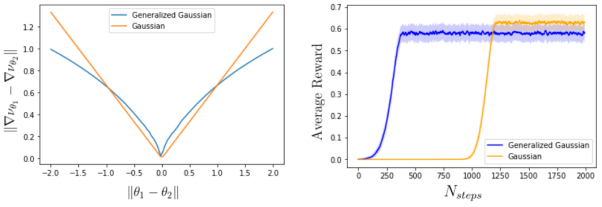

Figure 1. (a) **Tail Growth**: Comparing the growth of ψ_θ in one-dimension for Gaussian policies versus the Generalized Gaussian (Example 1) with $\alpha = 0.5$, for the $[0, 0]$ state in the MountainCar environment. (b) **Exploration Performance**: Comparing the performance of Generalized Gaussian 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.

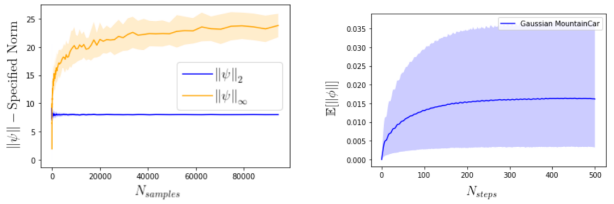

Figure 2. (a) **Gradient Norm Growth**: Comparing the growth of Example 3 using the L_2 norm described by Assumption 2, versus $\max_n \|\psi(s_n, a_n)\|$ with growing number of samples. While our criterion is stable, the max diverges logarithmically. (b) **Ergodicity of the Test Function**: Convergence 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.

For a policy with exponent κ, this occurs with probability

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

Assuming that $|\theta^* - \theta_0| \gg 0$, then

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

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). 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*

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

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 while preserving Hölder regularity. Weak smoothness has also been shown for many other elliptic families of PDEs (Høeg & Lindqvist, 2020; Sciunzi, 2014), which may also serve as 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 permitted under Assumption **??**, consider the following.

Example 4 *(Simplex MDPs) 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 $[\mu_1(s), \mu_2(s) \ldots]$ satisfying Assumption **??** forms a valid linear MDP. For example, $\boldsymbol{\mu}$ can be Gaussian in each component.*

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. 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.

1.5. Related Work

Optimization and Stochastic Approximation

We primarily refer to work on stochastic approximation, which began with the work by authors (Polyak & Juditsky, 1992; Kushner & 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 & Szepesvari, 2018; Jain et al., 2018).

The theory for optimizing weakly smooth rather than Lipschitz functionals was primarily developed in the following works (Devolder et al., 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 et al., 2013; Rudolf 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 et al., 1999; Konda & 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 & 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 & 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. (Xu et al., 2020; Kumar et al., 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 & 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.

1.6. 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 numerous lines of future research. Our assumptions are also practically applicable, as we demonstrate through several examples.

References

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

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

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

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

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

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

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

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

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

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020.

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

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

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