

Convergence and Optimality of Policy Gradient Methods in Weakly Smooth Settings

Matthew S. Zhang^{*} Murat A. Erdogdu[†] Animesh Garg[‡]

November 2, 2021

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.

1 Introduction

Modern Reinforcement Learning (RL) has solved challenges in diverse fields such as finance, healthcare, and robotics [1]–[3]. 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 [4], [5]). However, most practical problems exist in continuous, high-dimensional domains [6], and may even be infinite-dimensional or non-compact.

Theoretical results in continuous domains do not effectively characterize practical algorithms. While value-based estimators have obtained strong results in some regimes such as linear MDPs, both in on- and off-line settings [7], [8]. 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 [4], or apply strong assumptions on the policy and MDP [9]–[11]. This conflicts with practical results, where convergence often occurs without boundedness or smoothness preconditions on the function approximator.

^{*}Department of Computer Science at the University of Toronto, and Vector Institute, matthew.zhang@mail.utoronto.ca

[†]Department of Computer Science and Department of Statistical Sciences at the University of Toronto, and Vector Institute, erdogdu@cs.toronto.edu

[‡]Department of Computer Science at the University of Toronto, and Vector Institute, garg@cs.toronto.edu

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: (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:

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

with rate $T \times B = O\left((1-\gamma)^{(2\beta_0+11)/2\beta_0} \epsilon^{(\beta_0+3)/\beta_0}\right)$, 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: $J(\pi^*) - \frac{1}{T} \sum_{t \leq T} \mathbb{E}[J(\theta_t)] \leq \epsilon$, 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 [10], [12], [13], 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.

2 Background

2.1 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 $\pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$ 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 indefinitely.

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

Value Functions: We can define the value function as: $V_{\pi}(s) \triangleq \mathbb{E}_{\mathbb{P}_{\pi, \rho_0}}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s]$, and the Q-function as: $Q_{\pi}(s, a) \triangleq \mathbb{E}_{\mathbb{P}_{\pi, \rho_0}}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a]$. 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 [14].

3 Our Proposed Method

3.1 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 $\theta \in \Theta \subset \mathbb{R}^N$, 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 $\int_{\mathcal{A}} \exp(\nu_{\theta}(s, \cdot)) \leq \infty$ 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 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:

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

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

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:

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

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. [10], [13]) 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 $s \in \mathcal{S}$:

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

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 $P(s'|s, a) = \langle \mu(s'), \phi(s, a) \rangle$ for all s, a and some function $\mu(\cdot) : \mathcal{S} \mapsto \mathbb{R}^d$ and mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$. Under some weak assumptions on μ (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.

3.2 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

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

where $h_t \in \mathbb{R}$ is an adaptive step size. Alternatively, natural policy gradient (NPG), first introduced by [15], is a parameter invariant method that applies the following update

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

Algorithm 1 Policy Gradient for Hölder Smooth Objectives

```
1: Initial parameter  $\theta_0$ 
2: for Step  $t = 1, \dots, T - 1$  do
3:   for  $i = 1, \dots, B$  do
4:     Let  $j \sim \text{Geom}(1 - \gamma)$ ,  $h \sim \text{Geom}(1 - \gamma)$ ,  $\tau = j + h$ 
5:     Sample  $(s_{t,i,0}, a_{t,i,0}, \dots, s_{t,i,\tau}, a_{t,i,\tau})$ .
6:      $s_{t,i} \leftarrow s_{t,i,j}$ ,  $a_{t,i} \leftarrow a_{t,i,j}$ 
7:      $g_{t,i} \leftarrow \sum_{u=j}^{\tau} r_{t,i,u}$ ,  $r_{t,i,u} \sim R(s_{t,i,u}, a_{t,i,u})$ 
8:   end for
9:   Choose  $h_t$  specified in our learning rates section
10:   $\theta_t \leftarrow \theta_{t-1} + \frac{h_t}{B} \sum_{i=1}^B g_{t,i} \psi_{\theta_t}(s_{t,i}, a_{t,i})$ 
11: end for
12: Return  $\theta_T$ 
```

Algorithm 2 Natural Policy Gradient for Hölder Smooth Objectives

```
1: Initial parameter  $\theta_0$ , initial matrix  $K_0$ , stability parameter  $\xi > 0$ 
2: for Step  $t = 1, \dots, T - 1$  do
3:   for  $i = 1, \dots, B$  do
4:     Let  $j \sim \text{Geom}(1 - \gamma)$ ,  $h \sim \text{Geom}(1 - \gamma)$ ,  $\tau = j + h$ 
5:     Sample  $(s_{t,i,0}, a_{t,i,0}, \dots, s_{t,i,\tau}, a_{t,i,\tau})$ .
6:      $s_{t,i} \leftarrow s_{t,i,j}$ ,  $a_{t,i} \leftarrow a_{t,i,j}$ 
7:      $g_{t,i} \leftarrow \sum_{u=j}^{\tau} r_{t,i,u}$ ,  $r_{t,i,u} \sim R(s_{t,i,u}, a_{t,i,u})$ 
8:   end for
9:   Choose  $h_t$  specified in our learning rates section
10:   $K_t \leftarrow \frac{1}{B} \sum_{i=1}^B \psi(s_{t,i}, a_{t,i}) \psi^\top(s_{t,i}, a_{t,i}) + \xi I$ 
11:   $\theta_t \leftarrow \theta_{t-1} + K_t^{-1} \frac{h_t}{B} \sum_{i=1}^B g_{t,i} \psi_{\theta_t}(s_{t,i}, a_{t,i})$ 
12: end for
13: Return  $\theta_T$ 
```

where $K(\theta) = \mathbb{E}_{s,a \sim d_\theta^\rho} [\psi_\theta(s, a) \psi_\theta(s, a)^\top]$. 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:

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

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

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 [4].

3.3 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)** an optimal learning

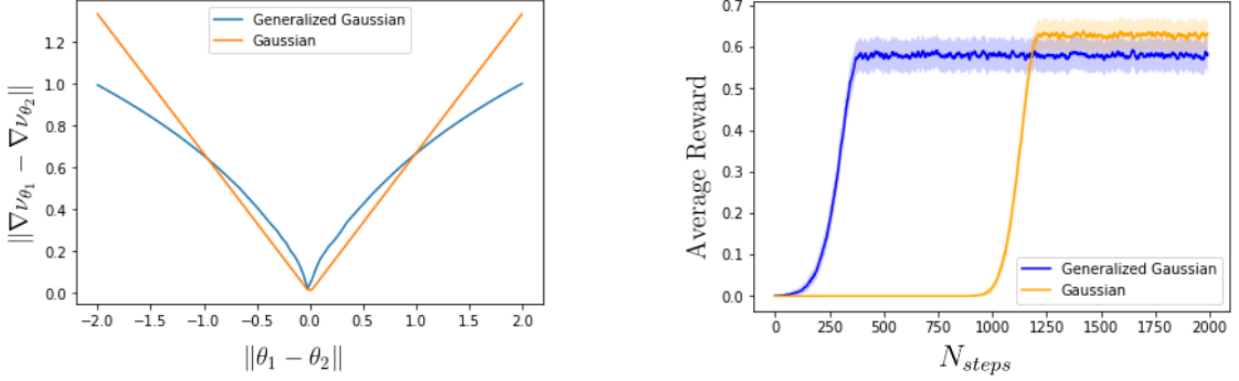


Figure 1: (a) **Tail Growth**: Comparing the growth of ψ_θ in one-dimension for Gaussian policies versus the Generalized Gaussian (Example 1) with $\alpha = 0.5$, for the $[0, 0]$ state in the MountainCar environment. (b) **Exploration Performance**: Comparing the performance of Generalized Gaussian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

rate $h_t = O(\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}})$, 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, which we defer to the appendices.

4 Applications

We note two prominent applications of our assumptions: (i) Assumption 1 to exploration has been explicitly shown in [16], (ii) Assumption 2 has been shown to apply to Safe RL via the work of [17]. 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:

$$\nu(a|s, \theta) = -|\langle \phi(s, a), \theta \rangle|^\kappa. \quad (9)$$

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

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

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

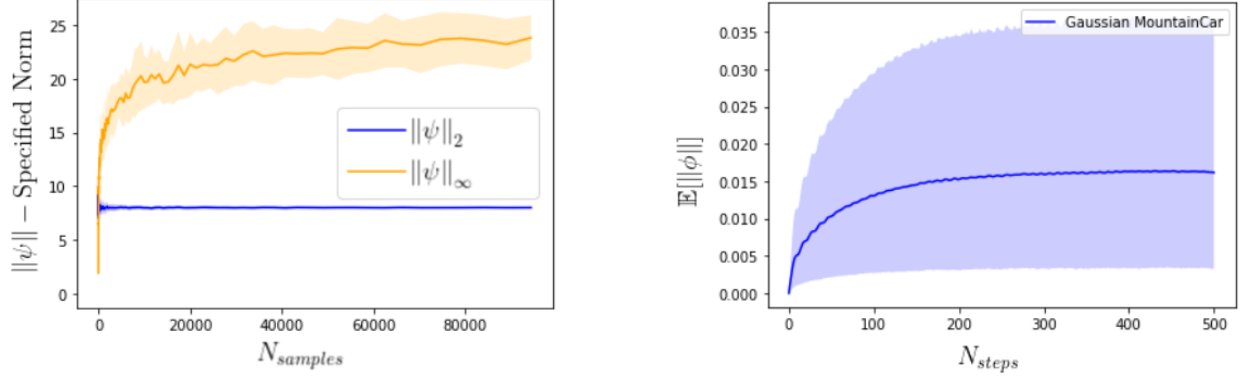


Figure 2: (a) **Gradient Norm Growth:** Comparing the growth of Example 3 using the L_2 norm described by Assumption 2, versus $\max_n \|\psi(s_n, a_n)\|$ with growing number of samples. While our criterion is stable, the max diverges logarithmically. (b) **Ergodicity of the Test Function:** Convergence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

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

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

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

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

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 [18] that solutions to the p -Laplacian

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

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 [19], [20], 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 $\theta \in [-1, 1]$, $\phi^* \in \mathbb{R}^d$:

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

Under uniform dynamics and a uniform distribution of $\phi(s, a)$ 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 explicitly avoids the state-action region around ϕ^* ; 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 $\{\sum_{i=1}^d \phi_i(s, a) = 1, \phi_i \geq 0\}$, then any vector of probability measures $[\mu_1(s), \mu_2(s) \dots]$ where each $\mu_i(s) \geq c_i$ is lower bounded forms a valid linear MDP. For example, μ 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.

5 Main Results

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

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

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 remaining constants are specified in the Appendix.

Natural Policy Gradient achieves the following convergence:

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

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 $\beta_0 \rightarrow 1$.

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
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$O((1-\gamma)^{-\frac{2}{1-\beta_0}}T^{-\frac{2\beta_0}{1+\beta_0}}) + O((1-\gamma)^{-1}T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}}B^{-\frac{\beta_0+1}{2}})$	
λt^q	$O((1-\gamma)^{-\frac{2}{1-\beta_0}}T^{f(q,\beta_0)}) + O((1-\gamma)^{-1}T^{g(q,\beta_0)}B^{-\frac{\beta_0+1}{2}})$	
$O\left(\ \nabla J(\theta_t)\ ^{\frac{1-\beta_0}{\beta_0}}\right)$	$O((1-\gamma)^{-\frac{1}{\beta_0}}T^{-1}) + O(B^{-\frac{\beta_0+1}{2}})$	Not practical

Remarks: As the norm in Assumption 2 strengthens to $\|\cdot\|_p, p \rightarrow \infty$, 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. [4], [10].

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 [10].

Corollary 1. (Rates under various step-size schemes) Table 1 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 $\rightarrow 2$ as $\beta_0 \rightarrow 1$.

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 :

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

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

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

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

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 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

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

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

$C_{NPG,2}, C_{NPG,3}, C_{PG,2}$ 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 [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 $t = 1 \dots T$, 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:

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

For **Natural Policy Gradient**, the rates are outlined in Table 3 so that the following holds:

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

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 $\beta \rightarrow 1$.

Remark: Although there is asymptotic bias, the term E_Π can be minimized with the appropriate choice of policy class.

6 Related Work

6.1 Optimization and Stochastic Approximation

We primarily refer to work on stochastic approximation, which began with the work by authors [21], [22], who established basic conditions for convergence for linear approximation procedures, with

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}	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

rates being obtained under strong assumptions. Tighter bounds have recently been achieved through improved analysis and techniques, both in asymptotic and non-asymptotic contexts [23]–[25].

The theory for optimizing weakly smooth rather than Lipschitz functionals was primarily developed in the following works [26]–[28], 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 [29] which yields perturbation bounds on the state distribution, and subsequent improvements in the assumptions and condition numbers [30], [31].

6.2 Reinforcement Learning

The general formulation of reinforcement learning can be attributed to Bellman’s formulation of Markov Decision processes [14]. Gradient-based approaches were proposed to solve direct policy parameterizations [32]; developments in this classical setting include [33]–[35]. 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 [15], which borrowed from similar work in standard optimization [36], yielded improved convergence with respect to policy condition numbers. In particular, strong convergence holds for domains such as the linear quadratic regulator [37], [38] and other linearized problems.

Even so, lower bounds for general problems can be quite pessimistic, especially when the conditions are ill-specified [39]. This debate has attracted renewed focus in recent years, with an on-going discussion on the quality of representation and its effect on learnability [40], [41]. Nonetheless, real world problems are either continuous or well-approximated by continuous algorithms, with smooth state-space. [4], [12] 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 [42]–[45]. [10], [46] 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 [9], [47], [48]; 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 [11], 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 $\rightarrow O(\epsilon^{-4}), \beta_0 \rightarrow 1$, this is because of numerous differences between our formulations. [11] 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 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.

7 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.

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 [11]. 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 [27]. 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.

References

- [1] Yue Deng, Feng Bao, Youyong Kong, et al. “Deep direct reinforcement learning for financial signal representation and trading”. In: *IEEE transactions on neural networks and learning systems* 28.3 (2016), pp. 653–664.
- [2] Chao Yu, Jiming Liu, and Shamim Nemati. “Reinforcement learning in healthcare: A survey”. In: *arXiv preprint arXiv:1908.08796* (2019).
- [3] Jens Kober, J Andrew Bagnell, and Jan Peters. “Reinforcement learning in robotics: A survey”. In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1238–1274.
- [4] Alekh Agarwal, Sham M Kakade, Jason D Lee, et al. “Optimality and approximation with policy gradient methods in markov decision processes”. In: *Conference on Learning Theory*. 2020, pp. 64–66.
- [5] Aaron Sidford, Mengdi Wang, Xian Wu, et al. “Near-optimal time and sample complexities for solving discounted Markov decision process with a generative model”. In: *arXiv preprint arXiv:1806.01492* (2018).
- [6] Kenji Doya. “Reinforcement learning in continuous time and space”. In: *Neural computation* 12.1 (2000), pp. 219–245.
- [7] Qi Cai, Zhuoran Yang, Chi Jin, et al. “Provably efficient exploration in policy optimization”. In: *arXiv preprint arXiv:1912.05830* (2019).
- [8] Lin F Yang and Mengdi Wang. “Sample-optimal parametric q-learning using linearly additive features”. In: *arXiv preprint arXiv:1902.04779* (2019).
- [9] Boyi Liu, Qi Cai, Zhuoran Yang, et al. “Neural proximal/trust region policy optimization attains globally optimal policy”. In: *arXiv preprint arXiv:1906.10306* (2019).
- [10] Tengyu Xu, Zhe Wang, and Yingbin Liang. “Improving Sample Complexity Bounds for Actor-Critic Algorithms”. In: *arXiv preprint arXiv:2004.12956* (2020).
- [11] Yanli Liu, Kaiqing Zhang, Tamer Basar, et al. “An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods.” In: *NeurIPS*. 2020.
- [12] Alekh Agarwal, Mikael Henaff, Sham Kakade, et al. “Pc-pg: Policy cover directed exploration for provable policy gradient learning”. In: *arXiv preprint arXiv:2007.08459* (2020).
- [13] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. “Finite-sample analysis for sarsa with linear function approximation”. In: *arXiv preprint arXiv:1902.02234* (2019).
- [14] Richard Bellman. *The theory of dynamic programming*. Tech. rep. Rand corp santa monica ca, 1954.
- [15] Sham M Kakade. “A natural policy gradient”. In: *Advances in neural information processing systems* 14 (2001), pp. 1531–1538.
- [16] Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. “Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution”. In: *International conference on machine learning*. PMLR. 2017, pp. 834–843.
- [17] Matteo Papini, Matteo Pirodda, and Marcello Restelli. “Smoothing policies and safe policy gradients”. In: *arXiv preprint arXiv:1905.03231* (2019).
- [18] Peter Lindqvist. *Notes on the p-Laplace equation*. 161. University of Jyväskylä, 2017.
- [19] Fredrik Arbo Høeg and Peter Lindqvist. “Regularity of solutions of the parabolic normalized p-Laplace equation”. In: *Advances in Nonlinear Analysis* 9.1 (2020), pp. 7–15.

- [20] Berardino Sciunzi. “Regularity and comparison principles for p-Laplace equations with vanishing source term”. In: *Communications in Contemporary Mathematics* 16.06 (2014), p. 1450013.
- [21] Boris T Polyak and Anatoli B Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855.
- [22] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Vol. 35. Springer Science & Business Media, 2003.
- [23] Xi Chen, Jason D Lee, Xin T Tong, et al. “Statistical inference for model parameters in stochastic gradient descent”. In: *arXiv preprint arXiv:1610.08637* (2016).
- [24] Chandrashekar Lakshminarayanan and Csaba Szepesvari. “Linear stochastic approximation: How far does constant step-size and iterate averaging go?” In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1347–1355.
- [25] Prateek Jain, Sham Kakade, Rahul Kidambi, et al. “Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification”. In: *Journal of Machine Learning Research* 18 (2018).
- [26] Olivier Devolder, François Glineur, and Yurii Nesterov. “First-order methods of smooth convex optimization with inexact oracle”. In: *Mathematical Programming* 146.1 (2014), pp. 37–75.
- [27] Yu Nesterov. “Universal gradient methods for convex optimization problems”. In: *Mathematical Programming* 152.1 (2015), pp. 381–404.
- [28] Maryam Yashtini. “On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients”. In: *Optimization letters* 10.6 (2016), pp. 1361–1370.
- [29] A Yu Mitrophanov. “Sensitivity and convergence of uniformly ergodic Markov chains”. In: *Journal of Applied Probability* 42.4 (2005), pp. 1003–1014.
- [30] Déborah Ferré, Loïc Hervé, and James Ledoux. “Regular perturbation of V-geometrically ergodic Markov chains”. In: *Journal of applied probability* 50.1 (2013), pp. 184–194.
- [31] Daniel Rudolf, Nikolaus Schweizer, et al. “Perturbation theory for Markov chains via Wasserstein distance”. In: *Bernoulli* 24.4A (2018), pp. 2610–2639.
- [32] Ronald J Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine learning* 8.3-4 (1992), pp. 229–256.
- [33] Richard S Sutton, Doina Precup, and Satinder Singh. “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning”. In: *Artificial intelligence* 112.1-2 (1999), pp. 181–211.
- [34] Vijay R Konda and John N Tsitsiklis. “Actor-critic algorithms”. In: *Advances in neural information processing systems*. 2000, pp. 1008–1014.
- [35] Sham Machandranath Kakade et al. “On the sample complexity of reinforcement learning”. PhD thesis. 2003.
- [36] Shun-Ichi Amari. “Natural gradient works efficiently in learning”. In: *Neural computation* 10.2 (1998), pp. 251–276.
- [37] Maryam Fazel, Rong Ge, Sham M Kakade, et al. “Global convergence of policy gradient methods for linearized control problems”. In: (2018).
- [38] Stephen Tu and Benjamin Recht. “Least-squares temporal difference learning for the linear quadratic regulator”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5005–5014.

- [39] Richard S Sutton, David A McAllester, Satinder P Singh, et al. “Policy gradient methods for reinforcement learning with function approximation”. In: *Advances in neural information processing systems*. 2000, pp. 1057–1063.
- [40] Simon S Du, Sham M Kakade, Ruosong Wang, et al. “Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?” In: *arXiv preprint arXiv:1910.03016* (2019).
- [41] Benjamin Van Roy and Shi Dong. “Comments on the du-kakade-wang-yang lower bounds”. In: *arXiv preprint arXiv:1911.07910* (2019).
- [42] Jincheng Mei, Yue Gao, Bo Dai, et al. “Leveraging non-uniformity in first-order non-convex optimization”. In: *arXiv preprint arXiv:2105.06072* (2021).
- [43] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, et al. “Variational policy gradient method for reinforcement learning with general utilities”. In: *arXiv preprint arXiv:2007.02151* (2020).
- [44] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, et al. “On the global convergence rates of softmax policy gradient methods”. In: *International Conference on Machine Learning*. PMLR, 2020, pp. 6820–6829.
- [45] Junyu Zhang, Chengzhuo Ni, Zheng Yu, et al. “On the convergence and sample efficiency of variance-reduced policy gradient method”. In: *arXiv preprint arXiv:2102.08607* (2021).
- [46] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. “On the sample complexity of actor-critic method for reinforcement learning with function approximation”. In: *arXiv preprint arXiv:1910.08412* (2019).
- [47] Philip S Thomas and Emma Brunskill. “Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines”. In: (2017).
- [48] Lingxiao Wang, Qi Cai, Zhuoran Yang, et al. “Neural policy gradient methods: Global optimality and rates of convergence”. In: *arXiv preprint arXiv:1909.01150* (2019).
- [49] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

The structure of the appendix is as follows. First we prove Proposition 1, and then we show the weak smoothness of the objective function $J(\theta)$. Then, we prove Theorem 1 and Theorem 2 respectively, as well as their respective corollaries. Lastly, we make some comments on an estimator for the optimal learning rate $h_t = \mathcal{O}(\|\nabla J(\theta_t)\|^{\frac{1-\beta_0}{\beta_0}})$. In the final appendix, we state the details for our numerical experiments.

A Weak Smoothness of Objective

A.1 Linear MDPs

As our first principal contribution, we prove ergodicity for linear and near-linear MDPs. An MDP is linear if there exists a measurable embedding $\phi : S \times A \rightarrow \mathbb{R}^d$, a vector of (signed) measures $\mu : \mathcal{B}(S) \rightarrow \mathbb{R}^d$, and a constant $\alpha > 0$, with the property that:

$$P(\cdot|s, a) = \langle \phi(s, a), \mu(\cdot) \rangle, |R(\cdot|s, a)| \leq \alpha. \quad (14)$$

A near-linear MDP is one where the transition kernel does not differ from a linear one in the L_∞ norm, under the pushforward measure $d\phi$, i.e. $P(\cdot|s, a) = \langle \phi(s, a), \mu(\cdot) \rangle + W_{s,a}(\cdot)$ for a family of signed measures $\{W_z, z \in S \times A\}$ that obey a strong integrability condition (see Assumption 7). Note that unlike the standard linear formulation, we do not assume linearity on the reward. We denote the sequence of sampled rewards as $\{r_t \sim R(\cdot|s_t, a_t)\}$. Without loss of generality, we can assume boundedness on the norm $\|\phi\| \leq 1$.

A.2 Ergodicity of Near-Linear MDPs

First we define a linear MDP as follows,

Recall that the n -step state-transition kernel is denoted by \mathbb{P}_θ^n . We denote \mathbb{P}_θ^1 by \mathbb{P}_θ as a shorthand. For finite state-spaces, ergodicity trivially for most policy classes. On the other hand, for general state spaces this is a strong assumption to make a priori. We decompose these into sufficient conditions on both the MDP and policy class. In the following, we refer to a Markov Chain as a pair $(\mathcal{S}, \mathbb{Q})$ where the first item is the state-space and the second item is the (time-invariant) transition function.

Definition 1. Let $\mathcal{B}(S)$ be the Borel algebra on S . Let us first introduce the concept of a small set with respect to the function \mathbb{Q} : define a \mathcal{V} -small set \mathcal{D} as one where the following integral holds:

$$\int_{\mathcal{C}} \mathbb{Q}(s'|s) ds' \geq \delta \mathcal{V}(\mathcal{C}) \quad \forall s \in \mathcal{D}$$

where \mathcal{V} is any measure, $\delta > 0$, for any Borel set $\mathcal{C} \in \mathcal{B}(S)$. If there exists a \mathcal{V} under which a \mathcal{D} is small, then we simply call it a **small set**.

The following assumption is needed to show the existence of a stationary measure on the state-actions.

Assumption 5. Consider the Markov Chain generated by the near-linear MDP, and policy π . We assume the following conditions hold for all θ :

$$\text{Irreducibility:} \quad \forall s_0 \in S, \mathcal{C} \in \mathcal{B}(S), \exists n \geq 1 : \quad \mathbb{P}(s_n \in \mathcal{C}|s_0) > 0 \iff \mathcal{V}(\mathcal{C}) > 0, \quad (15)$$

$$\text{Aperiodicity:} \quad \exists \mathcal{C} \in \mathcal{B}(S) : \quad \mathcal{C} \text{ small,} \quad \mathbb{P}(s_1 \in \mathcal{C}|s_0 \in \mathcal{C}) ds > 0, \quad (16)$$

$$\text{Finite Recurrence Time:} \quad \exists \mathcal{C} \in \mathcal{B}(\mathcal{S}) : \sum_{n=1}^{\infty} n \mathbb{P}(s_n \in \mathcal{C}, s_{n-1} \dots s_1 \notin \mathcal{C} | s_0 \in \mathcal{C}) < \infty, \quad (17)$$

where \mathcal{V} is a non-zero measure.

This is a weak set of assumptions, and it holds for sufficiently regular policy classes and near-linear MDPs. For instance, see the following example for a broad set of valid conditions.

Example 6. (*Conditions for Limiting Measure*) The following conditions guarantee the three properties of Assumption 5, respectively:

$$\begin{aligned} \forall s, \theta : \quad & \inf_{\|v\| \geq 1} \left| \int \pi_{\theta}(s, a) \left(\langle v, \phi(s, a) \rangle - \sup_{s'} |W_{s,a}(s')| \right) da \right| > 0 \\ & \int \int (\langle \mu(s), \phi(s, a) \rangle + W_{s,a}(s)) \pi_{\theta}(a|s) ds da > 0, \\ \forall \epsilon > 0, \exists \mathcal{C}_{\epsilon} \in \mathcal{B}(\mathcal{S}) : \quad & \|\mathcal{C}_{\epsilon}\| < \infty, \int_{\mathcal{C}_{\epsilon}} \|\mu(s)\| < \epsilon. \end{aligned}$$

The first condition essentially requires $\pi_{\theta} \times \phi$ to yield a density with non-zero volume; the second ensures aperiodicity by guaranteeing states transition to themselves with non-zero probability; the last one ensures that the probabilities decay outside sets with finite size.

We also provide some counterexamples to illustrate failure cases of MDPs.

Example 7. Consider the continuous random-walk Markov Chain on \mathbb{R}^m , $m \geq 1$. It is well-known that this fails the recurrence property, and the only invariant measure on the resulting state-space is the uniform measure on \mathbb{R}^m , which is not a probability measure.

Example 8. Consider a linear MDP on \mathbb{R}^2 , comprised of two disjoint sets of states $S_1 = \{s : \mu(s) = [a, 0]\}$ and $S_2 = \{s : \mu(s) = [0, b]\}$. Then suppose S_1 only had actions available for which $\phi(s, a) = [0, c]$, and likewise for S_2 the actions are $[d, 0]$. Then, although there is a unique stationary distribution, it is not the limiting measure for many initial distributions. Indeed, suppose ρ is non-zero only on S_2 , then $\mathbb{P}(s_n \in S_1 | s_0 \sim \rho) = n \bmod 2$ and does not converge; the distribution jumps between S_1 and S_2 depending on whether the time is an even or odd integer.

We introduce two necessary lemmas, which we adapt from prior work without proof.

Lemma 1. (Adapted from [49], Theorem 10.2.1) Let Assumption 5 hold. Then there exists a unique probability measure ρ_* on \mathcal{S} such that $\rho_*(s) = \int \mathbb{Q}(s|s') d\rho_*(s')$, and $\|\mathbb{P}(s_n = s | s_0 \sim \rho) - \rho_*(s)\| \rightarrow 0$ for all initial distributions ρ .

Lemma 2. (Adapted from [49], Theorem 16.0.1) Let $(\mathcal{S}, \mathbb{Q})$ be an aperiodic and irreducible Markov chain. Then if the following condition holds for a small set \mathcal{C} :

$$\int \mathbb{Q}(s'|s) f(s) ds - f(s') \leq -\delta f(s') + \tilde{\omega} 1_{\mathcal{C}}(s') \quad (18)$$

for all $s' \in \mathcal{C}$, Lyapunov function $g : \mathcal{S} \rightarrow [1, \infty)$ and $f = cg$, $\delta > 0$ and $\tilde{\omega} < \infty$, then the following holds

$$\int |\mathbb{Q}(s'|s) - \pi_*(s')| f(s) ds \leq cf(s) \kappa^N.$$

We say the chain satisfies f -uniform ergodicity if it satisfies the above inequality.

Remarks: In particular, if $f = 1$ is constant, then we call this uniform ergodicity.

Let us define the integral $\bar{\phi}_\theta(s) = \int \phi(s, a) \pi_\theta(a|s) da$. We then require additional regularity assumptions on the MDP and feature representation:

Assumption 6. (*Bounded Support of Actions*) Assume that the policies $\pi_\theta(s, a)$ and feature mapping $\phi(s, a)$ have the following property for all θ :

$$\forall s, s' \in \mathcal{S} : \quad \|\mu(s)\|, \|\mu(s')\| \geq (1 - \delta_s) \left\| \int \bar{\phi}_\theta(s) ds \right\|^{-1} \implies \langle \bar{\phi}_\theta(s), \bar{\phi}_\theta(s') \rangle > c,$$

where $c > 0$ is a positive constant, and $\delta_s \in (0, 1]$. This implies that for states with large μ , the policies have sufficiently broad support after pushforward by the map ϕ .

For near-linear MDPs, we require:

Assumption 7. (*Regularity of Near-Linear MDPs*) Assume that the kernel $W_{s,a}$ has the following property:

$$\forall s, a, s' : \quad W_{s,a}(s') \geq -(1 - \delta) \langle \mu(s'), \phi(s, a) \rangle,$$

where $\delta \in (0, 1]$.

This is a weak assumption that holds if W is sufficiently bounded. In particular, any strictly linear MDP satisfying Assumption 5 is sufficient. With this, we can show ergodicity for the Markov Chain on $\phi(s, a)$ generated by π_θ .

Lemma 3. Consider the Markov Chain (B_d, \mathbb{Q}_θ) generated by π_θ and near-linear MDP $\langle \mu, \phi \rangle + W_{s,a}$, satisfying Assumptions 5 and 7. This chain is uniformly ergodic, such that the following holds

$$\|\mathbb{Q}_\theta^N(\cdot | s_0 = s) - \tilde{\rho}_*(\cdot)\|_{TV} \leq c\tilde{\kappa}^N$$

for all n, θ , where $c, \tilde{\kappa} \geq 0$ and $\tilde{\rho}_*$ is the invariant distribution on s , and f is an arbitrary function.

Proof:

Consider the set $\mathcal{C} = \{s : \|\mu\|(s) \geq c\}$. This set is small by Assumption 6, since the transition function is lower bounded by:

$$\int \pi_\theta(a|s) \langle \mu(s'), \phi(s, a) \rangle \geq c \langle \mu(s'), \bar{\phi}_\theta \rangle$$

where we pick $\bar{\phi}_\theta = \bar{\phi}_\theta(s)$ using an arbitrary $s \in \mathcal{C}$. This measure is necessarily non-trivial.

We now show the drift condition for $f = 1$ a constant Lyapunov function, with the small set \mathcal{C} identified earlier. If s' does not lie in \mathcal{C} :

$$\int \pi_\theta(a|s) \langle \mu(s'), \phi(s, a) \rangle ds da \leq \|\mu(s')\| \left\| \int \bar{\phi}_\theta(s) ds \right\| \leq (1 - \delta_s).$$

Thus uniform geometric ergodicity holds. By our regularity assumption on $W_{s,a}$ (7), the same analysis will hold for near-linear MDPs with slight alterations of the constants. This shows ergodicity for some unknown function f . \square

This is a generalized form of Proposition 1 in the main text. Ergodicity holds since the inner product $\langle \mu, \phi \rangle$ is a non-local transformation (so long as the policy is sufficiently regular), with a finite d number of possible directions. Thus we can view a linear MDP as a relaxed form of a finite state MDP, which is always ergodic under the assumptions of irreducibility and aperiodicity.

Now we bound the differences between the two visitation distributions. We first introduce a lemma connecting ergodicity with the state-distribution function.

Lemma 4. (Adapted from [31], Theorem 3.2) Let the Markov Chain $(\mathcal{S}, \mathbb{P})$ be f -ergodic. Then

$$\|\mathbb{P}_{\theta,p}(s_n = \cdot) - \mathbb{P}_{\theta+\eta,p}(s_n = \cdot)\|_{TV} \leq C_1 D(\mathbb{P}_\theta, \mathbb{P}_{\theta+\eta})$$

where $D(\mathbb{Q}, \mathbb{R}) = \sup_s \frac{\|\mathbb{Q}(\cdot|s) - \mathbb{R}(\cdot|s)\|_{TV}}{f(s)}$ is the f -norm distance between two probability kernels, and $C_1 = \frac{\bar{\omega}}{\delta}$.

Lemma 5. Let ρ be the initial probability distribution on $\mathcal{S} \times \mathcal{A}$, satisfying Assumption 7. Then, for any set of parameters $\theta, \theta + \eta$

$$\left\| d_\theta^\rho(s, a) - d_{\theta+\eta}^\rho(s, a) \right\| \leq C_3 \|\eta\|^{\frac{\beta_1}{2}},$$

Proof: We note that, keeping the policy fixed, the state distribution can be written as the following:

$$d_\theta^\rho(s', a) = P_{\theta,\rho}(s') \exp(\nu_\theta(a|s')),$$

where $P_{\theta,\rho}(\cdot)$ is the state-component of the chain's visitation distribution under policy π_θ and initial distribution ρ .

Therefore we can bound the difference of two visitation distributions as:

$$\begin{aligned} \int_{(\mathcal{S} \times \mathcal{A})} |d_\theta^\rho - d_{\theta+\eta}^\rho| &\leq \int |P_{\theta,\rho}(s') \exp(\nu_\theta(a|s')) - P_{\theta+\eta,\rho}(s') \exp(\nu_\theta(a|s'))| \\ &\quad + \int |P_{\theta+\eta,\rho}(s') \exp(\nu_\theta(a|s')) - P_{\theta+\eta,\rho}(s') \exp(\nu_{\theta+\eta}(a|s'))| \\ &\leq \int_a 2\pi_\theta(\cdot|s') \|P_{\theta,\rho} - P_{\theta+\eta,\rho}\| + \int_{s'} 2P_{\theta,\rho}(\cdot) \|\pi_\theta - \pi_{\theta+\eta}\|. \end{aligned}$$

The latter term can be bounded through the total variation or the KL of the policies, i.e. Assumption 1. By Lemma 4, we know that

$$\|\mathbb{P}_{\theta,p}(s_n = \cdot) - \mathbb{P}_{\theta+\eta,p}(s_n = \cdot)\|_{TV} \leq C_1 D(\mathbb{P}_\theta, \mathbb{P}_{\theta+\eta}),$$

where $D(\cdot, \cdot)$ is the operator norm of the transition kernel. Note that this holds for any initial distribution p . We note that under uniform ergodicity of the policy class, the result follows from [10] (ultimately due to an ergodicity result in [29]); this requires a stronger property than we show in this work, since we would then be forced to assume $L(s)$ is finite everywhere.

Under these assumptions, the norm of the visitation distribution under perturbation can be bounded by the norm of the perturbation kernel:

$$\begin{aligned} \|P_{\theta,\rho} - P_{\theta+\eta,\rho}\| &= \left\| \frac{1}{1-\gamma} \sum_{n=1}^{\infty} \gamma^n [\mathbb{P}_{\theta,\rho}(s_t = \cdot) - \mathbb{P}_{\theta+\eta,\rho}(s_t = \cdot)] \right\| \\ &\leq \frac{1}{1-\gamma} \sum_{n=1}^{\infty} \gamma^n \|\mathbb{P}_{\theta,\rho} - \mathbb{P}_{\theta+\eta,\rho}\|_{TV} \\ &\leq \sup_s \frac{C_1}{f(s)} \|\mathbb{P}_\theta(\cdot|s) - \mathbb{P}_{\theta+\eta}(\cdot|s)\| \\ &\leq \sup_s \frac{C_1}{f(s)} \left(\int_{s'} |\mathbb{P}(s'|s, a) [\pi_\theta(a|s) - \pi_{\theta+\eta}(a|s)] da| ds' da \right) \\ &\stackrel{(i)}{\leq} \sup_s \frac{C_1 \sqrt{2}}{2f(s)} \sqrt{D_{KL}(\pi_\theta, \pi_{\theta+\eta})} \end{aligned}$$

$$\stackrel{(ii)}{\leq} \frac{C_1 \sqrt{2C_{\nu,1}}}{2} \|\eta\|^{\frac{\beta_1}{2}} \leq C_2 \|\eta\|^{\frac{\beta_1}{2}},$$

where (i) follows as $\int_{s'} \mathbb{P}(s'|s, a) = 1$ for any s, a , and (ii) follows from Assumption 1. Finally this suffices to bound the norm of the visitation distribution:

$$\|P_{\theta, \rho}(s) - P_{\theta+\eta, \rho}(s)\| \leq C_2 \|\eta\|^{\frac{\beta_1}{2}}.$$

Returning to our overall bound, we get:

$$\begin{aligned} \int_{S \times \mathcal{A}} |d_{\theta}^{\rho} - d_{\theta+\eta}^{\rho}| &\leq \int_a 2\pi_{\theta}(\cdot|s') \|P_{\theta, \rho} - P_{\theta+\eta, \rho}\| + \int_{s'} 2P_{\theta, \rho}(\cdot) \|\pi_{\theta} - \pi_{\theta+\eta}\| \\ &\leq 2C_2 \|\eta\|^{\frac{\beta_1}{2}} + \sqrt{2C_{\nu,1}} \|\eta\|^{\frac{\beta_1}{2}} \\ &\leq C_3 \|\eta\|^{\frac{\beta_1}{2}}. \end{aligned}$$

This concludes the proof of the Lemma, where we take the constant to be $C_3 = 2C_2 + \sqrt{2C_{\nu,1}}$. \square .

A.3 Q-function Analysis

We show that the assumptions are sufficient to show the following bound.

Lemma 6. *The Q-function is weakly smooth in the $\|\cdot\|$ metric. That is, for $\|\eta\|$ sufficiently small:*

$$\int d_{\theta}^{\rho} |Q_{\theta} - Q_{\theta+\eta}| \leq \frac{C_4}{1-\gamma} \|\eta\|^{\frac{\beta}{2}}.$$

Recall the definition of the Q-function:

$$Q_{\theta}(s, a) = \frac{1}{1-\gamma} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mathbb{P}_{\theta, \rho}^t(s_t, a_t | s_0 = s, a_0 = a).$$

We can write this in the following probabilistic form:

$$Q_{\theta}(s, a) = \int_{S \times \mathcal{A}} r(s_t, a_t) d_{\theta, \delta_{s,a}}(x) d\phi,$$

where d_{θ}^{ρ} is the visitation distribution, and δ is an atomic distribution at (s, a) .

Subsequently the difference reduces to:

$$\int |Q_{\theta} - Q_{\theta+\eta}| \leq \int_{S \times \mathcal{A}} |d_{\theta, \delta_{s,a}} - d_{\theta+\eta, \delta_{s,a}}| \alpha.$$

This then reduces to Lemma 5, under the initial distribution $\delta_{s,a}(x)$ (where δ denotes the singleton distribution). From our earlier bound we obtain the following:

$$\begin{aligned} |Q_{\theta} - Q_{\theta+\eta}|(s', a') &\leq \alpha \left\| d_{\theta}^{\delta_{s,a}} - d_{\theta+\eta}^{\delta_{s,a}} \right\| \\ &\leq \alpha C_3 \|\eta\|^{\frac{\beta_1}{2}}. \end{aligned}$$

We simplify the original integral through the following:

$$\begin{aligned} \int d_{\theta}^{\rho}(s, a) |Q_{\theta} - Q_{\theta+\eta}| &\leq C_3 \|\eta\|^{\frac{\beta_1}{2}} \int_S P_{\theta, \rho}(s) \int_{\mathcal{A}} \pi_{\theta}(a|s) \\ &\leq C_3 \|\eta\|^{\frac{\beta_1}{2}} \int_S P_{\theta, \rho}(s) f(s). \end{aligned}$$

The integral then reduces via Assumption 7:

$$\begin{aligned} \int f(s) P_{\theta, \rho}(s) &\leq \frac{1}{(1-\gamma)} \left[\int f(s) \rho(s) \right] \\ &\leq \frac{C_\rho}{1-\gamma}, \end{aligned}$$

where we use that $P_{\theta, \rho} \leq \frac{\rho}{1-\gamma}$. This completes the proof if we take $C_4 = C_3 \times C_\rho$. \square

A.4 Bounding Score Function

It remains to bound the score function for our models.

$$\begin{aligned} &\int \int d_\theta^\rho(s, a) |\psi_\theta(s, a) - \psi_{\theta+\eta}(s, a)| da ds \\ &\leq \int P_{\theta, \rho}(s) ds \int |\psi_\theta(s, a) - \psi_{\theta+\eta}(s, a)| \pi_\theta(a|s) da \\ &\leq C_{\nu, 2} \|\eta\|^{\beta_2} \int P_{\theta, \rho}(s) ds = C_5 \|\eta\|^{\beta_2}, \end{aligned}$$

where (i) follows from weak smoothness on the score function found in Assumption 1.

A.5 Objective

We conclude by proving the weak smoothness property of the loss function.

Proposition 2. *For any $\theta, \eta \in \mathbb{R}^{d \times d} \times \mathbb{R}^d$, the loss gradient is Hölder at all $s \in \mathcal{S}$: $\|\nabla J(\theta) - \nabla J(\theta + \eta)\| \leq \frac{C}{1-\gamma} \|\eta\|^{\beta_0}$. Here $\beta_0 = \min(\beta_1^{\frac{1}{4}}, \beta_2)$.*

Proof:

$$\begin{aligned} \nabla J(\theta_1) - \nabla J(\theta_2) &= \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) \psi_{\theta_1}(s, a) d_{\theta_1}^\rho(s, a) - \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_2}(s, a) \psi_{\theta_2}(s, a) d_{\theta_2}^\rho(s, a) \\ &\leq \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) \psi_{\theta_1}(s, a) d_{\theta_1}^\rho(s, a) - \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) \psi_{\theta_1}(s, a) d_{\theta_2}^\rho(s, a) \\ &\quad + \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) \psi_{\theta_1}(s, a) d_{\theta_2}^\rho(s, a) - \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_2}(s, a) \psi_{\theta_2}(s, a) d_{\theta_2}^\rho(s, a) \\ &\leq \int_{\mathcal{S} \times \mathcal{A}} [Q_{\theta_1}(s, a) - Q_{\theta_2}(s, a)] \psi_{\theta_2}(s, a) d_{\theta_2}^\rho(s, a) + \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) [\psi_{\theta_1}(s, a) - \psi_{\theta_2}(s, a)] d_{\theta_1}^\rho(s, a) \\ &\quad + \int_{\mathcal{S} \times \mathcal{A}} Q_{\theta_1}(s, a) \psi_{\theta_1}(s, a) [d_{\theta_1}^\rho(s, a) - d_{\theta_2}^\rho(s, a)] \\ &\leq \sqrt{\int \|\psi_{\theta+\eta}^2\| d_{\theta+\eta}^\rho} \sqrt{\int |Q_\theta - Q_{\theta+\eta}|^2 d_{\theta+\eta}^\rho} + \|Q_\theta\|_\infty \int |\psi_\theta - \psi_{\theta+\eta}| d_{\theta+\eta}^\rho \\ &\quad + \|Q_\theta\|_\infty \sqrt{\int \|\psi_\theta\|^2 d\phi} \sqrt{\int |d_\theta^\rho - d_{\theta+\eta}^\rho|^2 d\phi(s, a)} \\ &\leq 2\sqrt{\psi_\infty} \sqrt{\|Q\|_\infty} \sqrt{\int |Q_\theta - Q_{\theta+\eta}| d_{\theta+\eta}^\rho} + \|Q_\theta\|_\infty \int |\psi_\theta - \psi_{\theta+\eta}| d_{\theta+\eta}^\rho \end{aligned}$$

$$+ 2 \|Q\|_\infty \|d\|_\infty \sqrt{\psi_\infty} \left\| d_\theta^\rho - d_{\theta+\eta}^\rho \right\|_{TV}^{\frac{1}{2}},$$

where we use a combination of triangle inequality, Cauchy-Schwarz and boundedness in ∞ -norm of various quantities. We note that as $\|d\|_\infty \leq 1$ and $\|Q\|_\infty \leq \frac{\alpha}{1-\gamma}$, then we can bound the L_2 norms using L_1 and L_∞ norms.

Consequently, it remains to bound each of the terms individually, which have done above. Putting all of this together, we get a final bound of:

$$\begin{aligned} \nabla J(\theta) - \nabla J(\theta + \eta) &\leq 2\sqrt{\psi_\infty \frac{\alpha}{1-\gamma} \frac{C_4}{1-\gamma}} \|\eta\|^{\frac{\beta_1}{4}} + \frac{\alpha}{1-\gamma} C_5 \|\eta\|^{\beta_2} + \frac{2\alpha}{1-\gamma} \sqrt{C_3 \psi_\infty} \|\eta\|^{\frac{\beta_1}{4}} \\ &\leq \frac{C}{1-\gamma} \|\eta\|^{\beta_0}, \end{aligned}$$

where $C = \max(\sqrt{\psi_\infty \alpha C_4}, \alpha C_5, \alpha \sqrt{C_3 \psi_\infty})$ is not dependent on the main variables γ, η .

Remarks: Note that Cauchy-Schwarz is not necessary if we simply assume ∞ -norm boundedness of ψ . Consequently we do not need to take square-roots for the Q and d distributions, which allows us to retain their original smoothness parameters $\frac{\beta_1}{2}, \beta_2$. This fact is noted in the main text. For any intermediate integrability conditions $\psi \in L_p$, we can instead use a Hölder inequality in the integral to arrive at a different value for β_0 .

B Convergence to Local Minima

We begin with the proof for Theorem 1.

B.1 Policy Gradient

We first analyze the vanilla policy gradient. Let $\mathbf{g}_t = \nabla_\theta J(\theta_t)$ be shorthand for the exact gradient at time t . Following Proposition 2, we assume that the target function $J(\theta)$ has Hölder continuous gradient, i.e. $\nabla J(\theta) - \nabla J(\theta + \eta) \leq L \|\eta\|^{\beta_0}$.

Now, analyzing the iterates, we see that $\theta_t - \theta_{t-1} = -h_t \mathbf{g}_t$. Thus we apply mean value theorem to get:

$$\begin{aligned} |J(\theta_t) - J(\theta_{t-1}) + \langle \mathbf{g}_t, h_t \mathbf{g}_t \rangle| &= \left| \int_0^1 \langle \nabla J(\theta_{t-1} + \tau h_t \mathbf{g}_t) - \mathbf{g}_t, h_t \mathbf{g}_t \rangle d\tau \right| \\ &\leq \|h_t \mathbf{g}_t\| \int_0^1 |\nabla J(\theta_{t-1} + \tau h_t \mathbf{g}_t) - \mathbf{g}_t| \\ &\leq \|h_t \mathbf{g}_t\| \left[L \|h_t \mathbf{g}_t\|^{\beta_0} \int_0^1 \tau^{\beta_0} \right] \leq \frac{L}{\beta_0 + 1} \|h_t \mathbf{g}_t\|^{\beta_0+1}. \end{aligned}$$

Consequently the following holds (note that $\mathbf{g}_t = \mathbf{g}_t$):

$$\begin{aligned} J(\theta_t) &\leq J(\theta_{t-1}) - \langle h_t \mathbf{g}_t, \mathbf{g}_t \rangle + \frac{L}{\beta_0 + 1} \|h_t \mathbf{g}_t\|^{\beta_0+1} \\ &\leq J(\theta_{t-1}) - h_t \|\mathbf{g}_t\|^2 + \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|\mathbf{g}_t\|^{\beta_0+1} \\ &\leq J(\theta_{t-1}) - h_t \|\mathbf{g}_t\| \left[\|\mathbf{g}_t\| - \frac{L}{\beta_0 + 1} h_t^{\beta_0} \|\mathbf{g}_t\|^{\beta_0} \right]. \end{aligned}$$

Subsequently, if we collect such terms, we get the following:

$$\sum_{t=1}^K h_t \|\mathbf{g}_t\|^2 \leq J(\theta_0) - J(\theta_*) + \sum_{t=1}^K \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|\mathbf{g}_t\|^{\beta_0+1}.$$

The final term is a residual term which can only vanish if h_t is sufficiently small.

Now we work with the stochastic form of the gradient to get:

$$\begin{aligned} J(\theta_t) &\leq J(\theta_{t-1}) - \langle h_t \mathbf{g}_t, \mathbf{g}_t \rangle + \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|\mathbf{g}_t - r(s_t, a_t) \psi_\theta(s_t, a_t)\|^{\beta_0+1} + \frac{L}{\beta_0 + 1} \|h_t \mathbf{g}_t\|^{\beta_0+1} \\ &\quad - \left\langle \mathbf{g}_t, h_t \left[\mathbf{g}_t - \frac{1}{B} \sum_{i=1}^B v_{t,i} \psi_\theta(s_{t,i}, a_{t,i}) \right] \right\rangle, \end{aligned}$$

where the final term quantifies the noisy difference between the exact policy gradient, and the estimator used in practice, where $v_{t,i}$ is the return from our Algorithms. (Note that this is unbiased if we draw s_t, a_t from $d_{\theta, \rho}$.)

$$e_{t,i} = (\mathbf{g}_t - v_{t,i} \psi_\theta(s_{t,i}, a_{t,i})). \quad (19)$$

Observe that this term is not white noise, but is dependent on θ and the problem parameters α . To control the gradient, we introduce the following basic lemma on the variance of (one-dimensional) random variables:

Lemma 7. *Let X, Y be two random variables. Then*

$$\text{Var}(XY) \leq 2 \text{Var}(X) \|Y\|_\infty^2 + 2\mathbb{E}[X]^2 \text{Var}(Y).$$

Proof: Let $U = (X - \mathbb{E}[X])Y$ and $V = \mathbb{E}[X]Y$. Then by Young's inequality,

$$\begin{aligned} \text{Var}(XY) &= \text{Var}(U + V) \\ &= 2\text{Var}(U) + 2\text{Var}(V). \end{aligned}$$

The latter term is clearly bounded as $\mathbb{E}[X]$ is constant. For the first term, since X is now centered:

$$\begin{aligned} \text{Var}((X - \mathbb{E}[X])Y) &\leq \mathbb{E}[(X - \mathbb{E}[X])^2 Y^2] \\ &\leq \text{Var}((X - \mathbb{E}[X])) \|Y\|_\infty^2. \end{aligned}$$

This completes the proof. \square .

Now we can show the following lemma concerning the gradient:

Lemma 8. *(Variance of Noisy Policy Gradient) Consider the noise term $e_{t,i}$ in (19). We show that its variance is finite:*

$$\mathbb{E}_{d_{\theta_t}^\rho} [\|\mathbf{g}_t - v_t \psi_\theta(s_t, a_t)\|^2] \leq \sigma^2,$$

where $\sigma^2 \leq \frac{1}{(1-\gamma)} \sqrt{6} \|\alpha\|^2 \psi_\infty$ is a positive constant controlled by our assumptions.

Proof: Applying Lemma 7 to each component:

$$\mathbb{E}_{d_{\theta_t}^\rho} [\|\mathbf{g}_t - v_t \psi_{\theta_t}(s_t, a_t)\|^2] \leq 2\text{Tr}(\text{Var}_{d_{\theta_t}^\rho}(\phi_{\theta_t})) \|\alpha^2\| + 2 \left\| \mathbb{E}_{d_{\theta_t}^\rho} [\phi_{\theta_t}] \right\|^2 \|\alpha^2\|.$$

Then for all $\theta \in \Theta$, we know that

$$\text{Tr}(\text{Var}_{d_\theta^\rho}(\phi_\theta)) = \mathbb{E}_{d_\theta^\rho} [\|\phi_\theta - \mathbb{E}_{d_\theta^\rho} [\phi_\theta]\|^2]$$

$$\leq 2\mathbb{E}_{d_\theta^p} [\|\phi_\theta\|^2] + 2\mathbb{E}_{d_\theta^p} [\|\phi_\theta\|]^2.$$

Then using that $\|\mathbb{E}[\phi_\theta]\|^2 \leq \mathbb{E}[\|\phi_\theta\|^2] \leq \psi_\infty$ by Jensen's inequality, we obtain our final bound. \square
This directly implies the bound on the minibatch gradient.

Lemma 9. (*Variance of Minibatch Policy Gradient*) Consider the noise term

$$e_t = \frac{1}{N} \sum_{i=1}^N v_{t,i} \psi_\theta(s_{t,i}, a_{t,i}) - \mathbf{g}_t, \quad (20)$$

where $s_{t,i}, a_{t,i}$ are i.i.d. sampled from $d_{\theta_t}^p$. We show that its variance is finite:

$$\mathbb{E}_{d_{\theta_t}^p} [\|e_t\|^2] \leq \frac{\sigma^2}{B},$$

where $\sigma^2 = \sqrt{6} \|\alpha\|^2 \psi_\infty$ is a positive constant controlled by our assumptions.

Proof: This follows from Lemma 8 and the i.i.d. assumption.

Remarks: This becomes unnecessary if we assume that $\psi_\theta(s, a)$ is absolutely bounded; in that case, the final term becomes $\leq \psi_\infty$.

Thus the error terms have variance $\mathbb{E}[\|e_t\|^2] \leq \frac{\sigma_{MB}^2}{B}$, and $\mathbb{E}[\|e_t\|^{\beta_0+1}] \leq \left(\frac{\sigma}{\sqrt{B}}\right)^{\beta_0+1}$ by Jensen's inequality; denote σ^{β_0+1} as simply $\tilde{\sigma}$. Consequently we return to the convergence analysis, and see that:

$$\mathbb{E}[\langle \mathbf{g}_t, h_t [\mathbf{g}_t - v_t \psi_\theta(s_t, a_t)] \rangle] = 0,$$

since $\mathbf{g}_t, e_{t,i}$ are uncorrelated by assumption.

Consequently, we can substitute our bound and apply Jensen's inequality to obtain

$$\begin{aligned} \sum_{t=1}^T h_t \|\mathbf{g}_t\|^2 &\leq J(\theta_0) - J(\theta_*) + \sum_{t=1}^T \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \left(\mathbb{E}[\|e_t\|^{\beta_0+1}] + \|\mathbf{g}_t\|^{\beta_0+1} \right) \\ &\leq J(\theta_0) - J(\theta_*) + \sum_{t=1}^T \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \left(\tilde{\sigma} B^{-\frac{\beta_0+1}{2}} + \|\mathbf{g}_t\|^{\beta_0+1} \right). \quad \square \end{aligned}$$

Proof of Corollary 1 Now consider the constant learning rate, $h_t = \lambda T^{\frac{\beta_0-1}{\beta_0+1}}$, when $\beta_0 \neq 1$. (For the case $\beta_0 = 1$, see the constant step-size analysis below). In that case, we get the following order of convergence

$$\begin{aligned} &\sum_{t=1}^T h_t \|\mathbf{g}_t\|^2 - \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \left(\tilde{\sigma} + \|\mathbf{g}_t\|^{\beta_0+1} \right) \\ &= \sum_{t=1}^T \lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^2 - \frac{L \lambda^{\beta_0+1}}{\beta_0 + 1} T^{\beta_0-1} \left(\tilde{\sigma} + \|\mathbf{g}_t\|^{\beta_0+1} \right) \\ &= -\frac{L \lambda^{\beta_0+1} \tilde{\sigma}}{\beta_0 + 1} T^{\beta_0} B^{-\frac{\beta_0+1}{2}} + \sum_{t=1}^T \lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^{\beta_0+1} \left(\|\mathbf{g}_t\|^{1-\beta_0} - \frac{L \lambda^{\beta_0}}{\beta_0 + 1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} \right). \end{aligned}$$

Now consider the following cases: either $\|\mathbf{g}_t\|^{1-\beta_0} \geq 2 \frac{L\lambda^{\beta_0}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}}$, in which case the following holds:

$$\lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^{\beta_0+1} \left(\|\mathbf{g}_t\|^{1-\beta_0} - \frac{L\lambda^{\beta_0}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} \right) \geq \frac{\lambda}{2} T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^2,$$

or $\|\mathbf{g}_t\|^{1-\beta_0} < 2 \frac{L\lambda^{\beta_0}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}}$, in which case we have the following:

$$\begin{aligned} & \lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^{\beta_0+1} \left(\|\mathbf{g}_t\|^{1-\beta_0} - \frac{L\lambda^{\beta_0}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} \right) \\ & \geq \lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^2 - \left(2 \frac{L\lambda^{\beta_0}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} \right)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L\lambda^{\beta_0+1}}{\beta_0+1} T^{\beta_0-1} \right) \\ & \geq \lambda T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^2 - (2\lambda)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L}{\beta_0+1} \right)^{\frac{2}{1-\beta_0}} \frac{1}{T}. \end{aligned}$$

Note that in either case, this implies the following inequality:

$$\sum_{t=1}^T \frac{\lambda}{2} T^{\frac{\beta_0-1}{\beta_0+1}} \|\mathbf{g}_t\|^2 \leq (J(\theta_0) - J(\theta_*)) + (2\lambda)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L}{\beta_0+1} \right)^{\frac{2}{1-\beta_0}} + \frac{L\lambda^{\beta_0+1}\tilde{\sigma}}{\beta_0+1} T^{\beta_0} B^{-\frac{\beta_0+1}{2}},$$

$$\frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \leq \frac{2}{\lambda} T^{-\frac{2\beta_0}{\beta_0+1}} \left(J(\theta_0) - J(\theta_*) + (2\lambda)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L}{\beta_0+1} \right)^{\frac{2}{1-\beta_0}} \right) + \frac{2L\lambda^{\beta_0}\tilde{\sigma}}{\beta_0+1} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} B^{-\frac{\beta_0+1}{2}}. \quad \square$$

Note that we can optimize this expression for λ in order to obtain an optimal bound, but in practice this is infeasible since most of the parameters (such as L, β_0) are not known.

In order for the right hand side to be less than ϵ , we first choose $T = O(\epsilon^{(-\beta_0-1)/(2\beta_0)}(1-\gamma)^{-(\beta_0+1)(\beta_0-\beta_0^2)})$ so that the first term is less than $\frac{\epsilon}{2}$. Subsequently, we choose $B = O(\epsilon^{-1})$, so that the second term is equally small. Thus we get a total sample size of $T \times B = O((1-\gamma)^{-(\beta_0+1)(\beta_0-\beta_0^2)}\epsilon^{(-3\beta_0-1)(2\beta_0)})$, in order for $\frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \leq \epsilon$. We ignore the remaining parameters.

Now consider the time-independent learning rate, $h_t = \lambda$ for some $\lambda > 0$. If we substitute this into our earlier inequalities, we obtain the following cases: either $\|\mathbf{g}_t\|^{1-\beta_0} \geq 2 \frac{L\lambda^{\beta_0}}{\beta_0+1}$ or $\|\mathbf{g}_t\|^{1-\beta_0} < 2 \frac{L\lambda^{\beta_0}}{\beta_0+1}$, with a resulting bound of

$$\frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \leq \frac{2}{\lambda T} (J(\theta_0) - J(\theta_*)) + \frac{2L\lambda^{\beta_0}\tilde{\sigma}}{\beta_0+1} B^{-\frac{\beta_0+1}{2}} + (2\lambda)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L}{\beta_0+1} \right)^{\frac{2}{1-\beta_0}}. \quad \square$$

This results in a bias term that cannot be removed by minibatching.

We can generalize this to a decaying learning rate $h_t = \lambda t^q$ for $q \in (-1, 0]$, and thus we get the following (noting that $\sum_{t=1}^T t^p \leq \frac{T^{p+1}}{p+1}$ for $p \in [-1, 0]$, and $\sum t^p \leq C$ for $p < -1$):

$$\frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \leq \frac{2}{\lambda T} (J(\theta_0) - J(\theta_*)) + (2\lambda)^{\frac{\beta_0+1}{1-\beta_0}} \left(\frac{L}{\beta_0+1} \right)^{\frac{2}{1-\beta_0}} \max \left(cT^{\frac{2q\beta_0}{1-\beta_0}}, T^{-1} \right)$$

$$+ \frac{2L\lambda^{\beta_0}\tilde{\sigma}}{\beta_0 + 1} B^{-\frac{\beta_0+1}{2}} \max\left(cT^{q(\beta_0+1)}, T^{-1}\right). \quad \square$$

We can theoretically consider an optimal learning rate $h_t = \left(\frac{\lambda(\beta_0+1)}{L} \|\mathbf{g}_t\|^{1-\beta_0}\right)^{\frac{1}{\beta_0}}$ when the magnitude of the exact gradient is known, for some $\lambda \in (0, 1)$. This is a rapid order of decay, with $\frac{1-\beta_0}{\beta_0} \rightarrow \infty$ as $\beta_0 \rightarrow 0$. Then for the above analysis, we get

$$\begin{aligned} \sum_{t=1}^T h_t \|\mathbf{g}_t\|^2 - \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|\mathbf{g}_t\|^{\beta_0+1} &= \sum_{t=1}^T h_t \|\mathbf{g}_t\|^{\beta_0+1} \left(\|\mathbf{g}_t\|^{1-\beta_0} - \frac{L}{\beta_0 + 1} h_t^{\beta_0} \right) \\ &= \sum_{t=1}^T (1 - \lambda) \left(\frac{\lambda(\beta_0 + 1)}{L} \right)^{\frac{1}{\beta_0}} \|\mathbf{g}_t\|^{\frac{1+\beta_0}{\beta_0}}. \end{aligned}$$

Thus have completely eliminated the second case from our earlier analysis. If we combine this with the bound on the error term

$$\frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \tilde{\sigma} B^{-\frac{\beta_0+1}{2}} \leq \left(\frac{\beta_0 + 1}{L} \right)^{\frac{1}{\beta_0}} \lambda^{\frac{\beta_0+1}{\beta_0}} \tilde{\sigma} B^{-\frac{\beta_0+1}{2}} \left(1 + \|\mathbf{g}_t\|^{\frac{1+\beta_0}{\beta_0}} \right)$$

We can arrive at the final complexity of:

$$\frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^{\frac{1+\beta_0}{\beta_0}} \leq \left(\frac{L}{\lambda(\beta_0 + 1)} \right)^{\frac{1}{\beta_0}} \left(\frac{J(\theta_0) - J(\theta_*)}{\left(1 - \lambda - \lambda \tilde{\sigma} B^{-\frac{\beta_0+1}{2}}\right) T} \right) + \frac{\lambda \tilde{\sigma}}{\left(1 - \lambda - \lambda \tilde{\sigma} B^{-\frac{\beta_0+1}{2}}\right) B^{\frac{\beta_0+1}{2}}}.$$

This requires us to make the additional choice $\lambda < \frac{1}{2(1+\tilde{\sigma} B^{-(\beta_0+1)/2})}$. \square

Note the problems that arise as $\beta_0 \rightarrow 1$ in most of these scenarios; however, this can be remedied by choosing λ to be sufficiently small, so that the case $\|\mathbf{g}_t\|^{1-\beta_0} < O(h_t^{\beta_0})$ does not occur with high probability. In this work, however, we deal with the worst-case bounds for general λ .

Finally, we collect the rates required for each of these to be less than ϵ , and place this in the Table. We can also note that $G_T^p = \min_{t \leq T} \|\mathbf{g}_t\|^p \leq \frac{1}{T} \sum \|\mathbf{g}_t\|^p$ for any power p , which allows us to bound the minimum gradient G_T .

B.2 Natural Policy Gradient

First we show the following lemma:

Lemma 10. *For any $\theta \in \Theta$, the following holds:*

$$\left\| (K(\theta_t) + \xi I)^{-1} \mathbf{g}_t - K(\theta_t)^\dagger \mathbf{g}_t \right\| \leq C_{20} \|\mathbf{g}_t\|$$

where $C_{20} = \max(\zeta, \xi^{-1}) > 0$ and $\zeta = \inf_{\theta} \min_{k: \lambda_k > 0} \lambda_k(K(\theta)) \geq 0$ is a policy-dependent regularity constant, which measures the smallest non-zero eigenvalue of K across all policies θ .

Remarks: We adapt this from a similar result presented in [10]. This lemma is used to bound the stable inverse Fisher matrix $(K(\theta_t) + \xi I)^{-1}$ from the unstable pseudo-inverse $K(\theta_t)^\dagger$. Not only is the computation of the inverse far simpler than computation of the pseudo-inverse, but it will also be used to upper bound the rate of convergence. We can rewrite the condition number ζ in terms of the condition number used in [4], but this introduces unnecessary complexity for our purposes.

Table 4: Local convergence results of various learning rate schemes, for both policy gradient and natural policy gradient. We only track the primary dependence in ϵ, γ . For the decaying learning rate, we define the rate coefficients $f(p, \beta_0) = \max(-1, \frac{2q\beta_0}{1-\beta_0})$, $g(p, \beta_0) = \frac{-2}{\beta_0+1}(1 - f(p, \beta_0) \times \max(-1, q(\beta_0 + 1)))$, $h(p, \beta_0) = \frac{-2}{\beta_0+1}(1 - \frac{1-\beta_0}{2}f(p, \beta_0) \times \max(-1, q(\beta_0 + 1)))$

h_t	T	B	Considerations
λ	ϵ^{-1}	$\epsilon^{-\frac{2}{1+\beta_0}}(1-\gamma)^{\frac{-2}{1+\beta_0}}$	Bias term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{-\beta_0-1}{2\beta_0}}(1-\gamma)^{\frac{-(\beta_0+1)}{\beta_0-\beta_0^2}}$	ϵ^{-1}	
λt^q	$\left(\epsilon(1-\gamma)^{\frac{1-\beta_0}{2}}\right)^{f(q, \beta_0)}$	$\epsilon^{g(q, \beta_0)}(1-\gamma)^{h(p, \beta_0)}$	
$O\left(\ \mathbf{g}_t\ ^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{-1}(1-\gamma)^{-\frac{1}{\beta_0}}$	$\epsilon^{-\frac{2}{1+\beta_0}}$	Not practically estimable

Proof: Since $K(\theta)$ is symmetric, it is diagonalizable into $U(\theta)^T \Xi(\theta) U(\theta)$, where $U(\theta) = [u_1, u_2, \dots, u_d]$ is the orthonormal space of eigenvectors (and appropriate null vectors) and $\Xi(\theta) = [\xi_1 e_1, \dots, \xi_d e_d]$ where ξ_i is the i -th eigenvalue (including zeroes) and e_i is the standard basis in \mathbb{R}^d . Thus we can write the following:

$$\begin{aligned} [K(\theta) + \xi I]^{-1} \mathbf{g}_t &= U(\theta)^T (\Xi(\theta)^{-1} + \frac{1}{\xi} I) U(\theta) \mathbf{g}_t, \\ &= U(\theta)^T \left[\frac{1}{\xi_1 + \xi} \langle u_1, \mathbf{g}_t \rangle, \frac{1}{\xi_2 + \xi} \langle u_2, \mathbf{g}_t \rangle \dots \frac{1}{\xi_d + \xi} \langle u_d, \mathbf{g}_t \rangle \right]. \end{aligned}$$

Analogously, the exact pseudo-inverse yields:

$$\begin{aligned} K(\theta)^\dagger \mathbf{g}_t &= U(\theta)^T \Xi(\theta)^\dagger U(\theta) \mathbf{g}_t \\ &= U(\theta)^T \left[\frac{1}{\xi_1} \langle u_1, \mathbf{g}_t \rangle, \frac{1}{\xi_2} \langle u_2, \mathbf{g}_t \rangle \dots \frac{1}{\xi_k} \langle u_k, \mathbf{g}_t \rangle, 0 \dots 0 \right], \end{aligned}$$

where k is the rank of $K(\theta)$. Subsequently their difference is:

$$\begin{aligned} &K(\theta + \xi I)^{-1} \mathbf{g}_t - K(\theta)^\dagger \mathbf{g}_t \\ &\leq U(\theta)^T \left[\left(\frac{1}{\xi_1 + \xi} - \frac{1}{\xi_1} \right) \langle u_1, \mathbf{g}_t \rangle, \dots, \left(\frac{1}{\xi_k + \xi} - \frac{1}{\xi_k} \right) \langle u_k, \mathbf{g}_t \rangle, \frac{1}{\xi} \langle u_{k+1}, \mathbf{g}_t \rangle \dots \frac{1}{\xi} \langle u_d, \mathbf{g}_t \rangle \right] \\ &= -\xi U(\theta)^T \left[\left(\frac{1}{\xi_1(\xi_1 + \xi)} \right) \langle u_1, \mathbf{g}_t \rangle, \dots, \left(\frac{1}{\xi_k(\xi_1 + \xi)} \right) \langle u_k, \mathbf{g}_t \rangle, -\frac{1}{\xi^2} \langle u_{k+1}, \mathbf{g}_t \rangle \dots -\frac{1}{\xi^2} \langle u_d, \mathbf{g}_t \rangle \right]. \end{aligned}$$

Since the information matrix is positive semidefinite, we bound $\xi_1 + \xi \geq \xi_k + \xi \geq \xi$. Thus the norm is bounded:

$$\begin{aligned} \left\| K(\theta + \xi I)^{-1} \mathbf{g}_t - K(\theta)^\dagger \mathbf{g}_t \right\|_2 &\leq |\xi| \|U(\theta)^T\|_2 \max(\xi^{-2}, \xi^{-1}\zeta) \|\nabla J(\theta)\|_2 \\ &\leq \max(\zeta, \xi^{-1}) \|\nabla J(\theta)\|_2 = C_{20} \|\nabla J(\theta)\|_2, \end{aligned}$$

where ζ is the condition number described in the lemma, and we use the fact that U is unitary. \square

Assumption 2 also guarantees that the Fisher information matrix $K(\theta) = \mathbb{E}_{d_\theta} [\psi_\theta(s, a) \psi_\theta(s, a)^T]$ has bounded norm, since by the trace inequality

$$\|K(\theta)\| = \left\| \int \psi_\theta(s, a) \psi_\theta(s, a)^T d_{\theta, \rho} \right\| \leq \int \|\psi_\theta(s, a) \psi_\theta(s, a)^T\| d_{\theta, \rho} \leq \int \|\psi_\theta(s, a)\|_2^2 d_{\theta, \rho} \leq \psi_\infty,$$

This will be important when analyzing the NPG algorithm.

Proof for Theorem 1, NPG: Note that for the natural policy gradient we choose the learning rate to be instead: $h_t = \lambda \left[\frac{(\beta_0+1)\xi^{\beta_0-1}}{L} \right]^{\frac{1}{\beta_0}} \|\mathbf{g}_t\|^{\frac{1-\beta_0}{\beta_0}}$ where again λ is a sufficiently small constant. We now perform the same analysis for the exact natural policy gradient, assuming an oracle for \mathbf{g}_t :

$$\begin{aligned} J(\theta_t) &\leq J(\theta_{t-1}) - \langle \mathbf{g}_t, h_t(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t \rangle + \frac{L}{\beta_0 + 1} \|h_t(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|^{\beta_0+1} \\ &\leq J(\theta_{t-1}) - h_t \|(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|^2 + \frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|^{\beta_0+1} \\ &\leq J(\theta_{t-1}) - h_t \left[\|(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|^2 - \frac{L}{\beta_0 + 1} h_t^{\beta_0} \|(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|^{\beta_0+1} \right] \\ &\leq J(\theta_{t-1}) - h_t \|\mathbf{g}_t\| \left[\frac{1}{(\psi_\infty + \xi)^2} \|\mathbf{g}_t\| - \frac{L}{(\beta_0 + 1)\xi^{\beta_0+1}} h_t^{\beta_0} h_t^{\beta_0} \|\mathbf{g}_t\|^{\beta_0} \right], \end{aligned}$$

where we use the fact that $0 \leq \|K(\theta)\| \leq \psi_\infty$. This is entirely analogous to the standard policy gradient case, except for the terms $(K(\theta_t) + \xi I)$ which appears from the natural policy gradient. We can either lower bound or upper bound this dependent on the sign of term inside the brackets.

Similar to policy gradient, we can bound the error term with

$$\frac{L}{\beta_0 + 1} h_t^{\beta_0+1} \|(K(\theta_{t-1} + \xi I)^{-1} \mathbf{g}_t\|^{\beta_0+1} \leq \frac{L}{(\beta_0 + 1)\xi^{\beta_0+1}} h_t^{\beta_0+1} \tilde{\sigma} B^{-\frac{\beta_0+1}{2}}$$

Finally, we sum for all $t = 1 \dots T$ to get:

$$\frac{1}{T} \sum_{t=1}^T h_t \|\mathbf{g}_t\|^2 \leq \frac{1}{T} (J(\theta_0) - J(\theta_*)) + \frac{L(\psi_\infty + \xi)^2}{(\beta_0 + 1)\xi^{\beta_0+1}} h_t^{\beta_0+1} \left(\tilde{\sigma} B^{-\frac{\beta_0+1}{2}} + \frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t\|^{\beta_0+1} \right) \square$$

Proof of Corollary 1, NPG:

The proof techniques for generating the convergence rates are identical to policy gradient, except for the terms involving ξ, ψ_∞ which are introduced by the matrix K . As a result, the rates are unchanged in ϵ, γ . \square .

C Optimality

We now ask under what conditions will the stationary points be optimal within our policy class. Note that for standard policy gradient, the stationary points have previously been demonstrated to be non-convex [4]. However, under the following assumptions, we can show some form of global optimality; this is done by defining a global "domination" condition, under which optimality is satisfied. One such condition is:

Definition 2. (*Polyak-Łojasiewicz Condition*) A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies a Polyak-Łojasiewicz (PL) condition if, for any x in the minimizing set \mathcal{X}^* the following holds:

$$f(x) \geq f(\tilde{x}) + \frac{\mu}{2} F(x, \nabla f(\tilde{x})) \quad \forall \tilde{x},$$

where $\mu \geq 0$ is an arbitrary non-negative constant, and F is some suitable function or norm applied to the first-order term.

First we introduce an auxiliary lemma, which quantifies the gap between any two policies in a policy class. This is a standard lemma, see e.g. [4]; this is pivotal in the proof for optimality.

Lemma 11. (*Performance Difference Lemma*) *For all policies $\pi_{\theta_1}, \pi_{\theta_2}$ and states $s \in \mathcal{S}$:*

$$V_{\theta_1}(s) - V_{\theta_2}(s) = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_{\theta_1, \delta(s)}} \left[\mathbb{E}_{a \sim \pi_{\theta_1}(\cdot|s)} [A_{\theta_2}(s, a)] \right].$$

We first observe that the value function is equal to:

$$V_{\theta}(s) = \mathbb{E}_{\theta} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right],$$

so then we write the difference as:

$$\begin{aligned} V_{\theta_1}(s) - V_{\theta_2}(s) &= \mathbb{E}_{\theta_1} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] - V_{\theta_2}(s) \\ &= \mathbb{E}_{\theta_1} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + V_{\theta_2}(s_t) - V_{\theta_2}(s_t)) \right] - V_{\theta_2}(s) \\ &\stackrel{(i)}{=} \mathbb{E}_{\theta_1} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma V_{\theta_2}(s_{t+1}) - V_{\theta_2}(s_t)) + V_{\theta_2}(s_0) - V_{\theta_2}(s_0) \right] \\ &\stackrel{(ii)}{=} \mathbb{E}_{\theta_1} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma \mathbb{E}[V_{\theta_2}(s_{t+1})|s_t, a_t] - V_{\theta_2}(s_t)) \right] \\ &\stackrel{(iii)}{=} \mathbb{E}_{\theta_1} \left[\sum_{t=0}^{\infty} \gamma^t A_{\theta_2}(s_t) \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_{\theta_1, \delta(s)}} \mathbb{E}_{a \sim \pi_{\theta_1}(\cdot|s)} [A_{\theta_2}(s, a)] \\ &= \frac{1}{1-\gamma} \int_{\mathcal{S}} \int_{\mathcal{A}} \pi_{\theta_1}(a|s) d_{\theta_1, \delta(s)}(s) A_{\theta_2}(s, a) ds da, \end{aligned}$$

where in (i) we pull the leading term out of the summation, in (ii) we use the tower property of expectations, and in (iii) we substitute the definition of $A_{\theta}(s_t)$. Finally, we use the definitions of d_{θ_1} as the generator of the Markov chain, to arrive at the final equation. This concludes the proof. \square

We note that the performance difference lemma allows us to absorb the difference term into purely multiplicative analysis. We now state a global condition on the value function, which will suffice to prove optimality.

Lemma 12. (*Gradient Domination Lemma*) *Under Assumption 4, then for any policy π_{θ} , the following global bound holds:*

$$V_{\theta_1}(\rho) - V_{\theta_2}(\rho) \leq \frac{1}{1-\gamma} \left\| \frac{d_{\theta_1, \rho}}{d_{\theta_2, \rho}}(s) \right\|_{\infty} (\nabla_{\theta} J(\pi_{\theta_2})),$$

for any two policies $\pi_{\theta_1}, \pi_{\theta_2}$, any state distribution $\rho \in \Delta(\mathcal{S})$.

Remarks: This is similar to the gradient domination inequality proven by [4]; however, we consider only fixed distributions ρ , and additionally must work in the parameterized space. Thus the proof necessitates an additional assumption to incorporate the score function ψ .

Proof: The proof of this theorem is similar to that found in other works on policy gradient. Recall that the gradient of any (exponential) policy can be written as:

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1-\gamma} \int \int A_{\pi_{\theta}}(s, a) \psi_{\pi_{\theta}}(s, a) d_{\pi_{\theta}, \rho}(s) \pi(a|s) dad s.$$

By the performance difference lemma, we find:

$$\begin{aligned} J_{\theta_1}(s) - J_{\theta_2}(s) &= \frac{1}{1-\gamma} \int_{\mathcal{S}} \int_{\mathcal{A}} d_{\theta_1, \rho}(s) \pi_{\theta_1}(a|s) [A_{\theta_2}(s, a)] dad s \\ &= \frac{1}{1-\beta_0} \int_{\mathcal{S}} \int_{\mathcal{A}} d_{\theta_1, \rho}(s) \pi_{\theta_2}(a|s) \frac{\pi_{\theta_1}(a|s)}{\pi_{\theta_2}(a|s)} A_{\theta_2}(s, a) dad s \\ &\stackrel{(i)}{\leq} \frac{1}{1-\gamma} \int_{\mathcal{S}} \int_{\mathcal{A}} d_{\theta_1, \rho}(s) \|\psi_{\theta_2}(a|s)\| \pi_{\theta_2}(a|s) A_{\theta_2}(s, a) \\ &\leq \frac{1}{1-\gamma} \int_{\mathcal{S}} \int_{\mathcal{A}} \frac{d_{\theta_1, \rho}(s)}{d_{\theta_2, \rho}(s)} d_{\theta_2, \rho}(s) \|\psi_{\theta_2}(a|s)\| \pi_{\theta_2}(a|s) A_{\theta_2}(s, a) \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_{\theta_1, \rho}(s)}{d_{\theta_2, \rho}(s)} \right\|_{\infty} \int_{\mathcal{S}} \int_{\mathcal{A}} d_{\theta_2, \rho}(s) \|\psi_{\theta_2}(a|s)\| \pi_{\theta_2}(a|s) A_{\theta_2}(s, a) \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_{\theta_1, \rho}(s)}{d_{\theta_2, \rho}(s)} \right\|_{\infty} \|\nabla J(\pi_{\theta_2})\|, \end{aligned}$$

where in (i) we use Assumption 4 on the ratio of policies (which is the exponential of $\nu_{\theta_1} - \nu_{\theta_2}$), in (ii) we use the fact that the Radon-Nikodym derivative is absolutely bounded on \mathcal{S} . \square

This serves as the Polyak-Łojasiewicz result that will be required to show global optimality. We note that, instead of directly assuming the gradient domination condition (which is not easily verified), we provide a more verifiable assumption. Although not trivial to confirm, our assumption is significantly more interpretable in terms of the problem parameters, when compared to the result.

C.1 Policy Gradient

Finally, we use this result to show the global optimality of our algorithm for vanilla policy gradient, under loose assumptions.

Proof of Theorem 2 for PG: From the gradient domination lemma, we find that:

$$\begin{aligned} \min_{t=1, \dots, T} J_{\theta_*} - J_{\theta_t} &= \frac{1}{1-\gamma} \left\| \frac{d_{\theta_*, \rho}(s)}{d_{\theta_t, \rho}(s)} \right\|_{\infty} \min_{t=1, \dots, T} \|\mathbf{g}_t\| \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_{\theta_*, \rho}(s)}{\rho} \right\|_{\infty} \min_{t=1, \dots, T} \|\mathbf{g}_t\|. \quad \square \end{aligned}$$

To show the corollary, it remains to substitute the bounds for $\min_{t=1, \dots, T} \|\mathbf{g}_t\|$ that we obtain from Corollary 1. When converted to a sample complexity bound, this yields the final result.

Proof of Corollary 2, PG:

It remains now to substitute the learning rate and gradient bounds from each of the cases we analyzed earlier; doing so obtains the rates found in the following table. Since $\beta_1 \geq \frac{1}{2}$, the final term always dominates the order if ϵ is sufficiently small. Finally, we use everywhere that $(\sum_i x_i)^p \leq \sum_i x_i^p$ for $p \in [0, 1]$.

Case I, $h_t = \lambda$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^2 = O(T^{-1}) + O\left(\frac{1}{1-\gamma} B^{-\frac{\beta_0+1}{2}}\right) + O\left((1-\gamma)^{\frac{-2}{1-\beta_0}}\right)$.

Table 5: 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}	Considerations
λ	$\epsilon^2(1-\gamma)^2$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{6}{1+\beta_0}}$	Growing bias term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(2-\beta_0)(\beta_0+1)}{(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{4\beta_0-2}{\beta_0+1}}$	
$O\left(\ \mathbf{g}_t\ ^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{\beta_0+2}{\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{2}{\beta_0}}$	Not practically estimable

Subsequently we require $T^{-\frac{1}{2}} \leq \epsilon(1-\gamma)$ and $B^{-\frac{\beta_0+1}{4}} \leq \epsilon(1-\gamma)^{\frac{3}{2}}$, to obtain the following:

$$\min J_* - \mathbb{E}[J(\theta_t)] \leq \epsilon + O\left((1-\gamma)^{\frac{-1}{\beta_0-1}}\right).$$

Case II, $h_t = \lambda T^{\frac{\beta_0-1}{\beta_0+1}}$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^2 = O((1-\gamma)^{\frac{-2}{1-\beta_0}} T^{-\frac{2\beta_0}{\beta_0+1}}) + O\left(\frac{1}{1-\gamma} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} B^{-\frac{\beta_0+1}{2}}\right)$.

Subsequently we require $T^{-\frac{\beta_0}{\beta_0+1}} \leq \epsilon(1-\gamma)^{1+\frac{1}{1-\beta_0}}$ and $B^{-\frac{\beta_0+1}{4}} \leq \epsilon^{\beta_0}(1-\gamma)^{\frac{2\beta_0-1}{2}}$, to obtain the following:

$$\min J_* - \mathbb{E}[J(\theta_t)] \leq \epsilon. \quad (21)$$

Case III, $h_t = \left(\frac{\lambda(\beta_0+1)}{L} \|\mathbf{g}_t\|^{1-\beta_0}\right)^{\frac{1}{\beta_0}}$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^{\frac{\beta_0+1}{\beta_0}} = O((1-\gamma)^{\frac{-1}{\beta_0}} T^{-1}) + O\left(B^{-\frac{\beta_0+1}{2}}\right)$.

Subsequently we require $T^{-\frac{\beta_0}{\beta_0+1}} \leq \epsilon(1-\gamma)^{\frac{\beta_0+2}{\beta_0+1}}$, and $B^{-\frac{\beta_0}{2}} \leq \epsilon(1-\gamma)$, to obtain (21).

C.2 Natural Policy Gradient

Proof of Theorem 2 for NPG: In terms of optimality of the natural policy gradient, consider the KL metric: $D(\theta_t, s) = D_{KL}(\pi_*(\cdot|s), \pi_{\theta_t}(\cdot|s))$ where π^* is any optimal policy (uniqueness of this policy is possible under some conditions [4]). Let $d_* \in \Delta(\mathcal{S} \times \mathcal{A})$ be the distribution measure of π^* , $d_*(\mathcal{S} \times \mathcal{A}) = \sum_t \beta_0^t \mathbb{P}_{\pi^*}(s_t \in \mathcal{S}, a_t \in \mathcal{A})$. We subsequently write:

$$\begin{aligned} & D_t(\theta_t, s) - D_{t-1}(\theta_{t-1}, s) \\ &= \mathbb{E}_{s, a \sim d_*} [\log \pi_{\theta_{t-1}}(a|s) - \log \pi_{\theta_t}(a|s)] \\ &\stackrel{(i)}{\geq} \mathbb{E}_{s, a \sim d_*} [\nabla \log \pi_{\theta_{t-1}}(a|s)]^T (\theta_t - \theta_{t-1}) - \frac{L}{\beta_0 + 1} \|\theta_t - \theta_{t-1}\|^{\beta_0+1} \\ &= \mathbb{E}_{s, a \sim d_*} [\psi_{\theta_t}(s_t, a_t)]^T (\theta_t - \theta_{t-1}) - \frac{L}{\beta_0 + 1} \|\theta_t - \theta_{t-1}\|^{\beta_0+1}, \end{aligned}$$

where in (i) we use the Hölder-smoothness condition on ψ . Let $\mathbf{g}_t = r(s_t, a_t) \psi_{\theta_t}(s_t, a_t)$ be the approximate gradient. Concentrating on the first term, we can make use of the triangle inequality

to obtain the following:

$$\begin{aligned}
&= \mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (\theta_t - \theta_{t-1}) \\
&= \mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K_t + \xi I)^{-1} \tilde{\mathbf{g}}_t) \\
&\leq \underbrace{\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K_t + \xi I)^{-1} \mathbf{g}_t - h_t(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t)}_{(a)} \\
&\quad + \underbrace{\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t - h_t K(\theta_{t-1})^\dagger \mathbf{g}_t)}_{(b)} \\
&\quad + \underbrace{\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T h_t K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t}(s, a)}_{(c)} + \underbrace{\mathbb{E}[A_{\theta_t}(s, a)]}_{(d)} \\
&\quad + \underbrace{\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K_t + \xi I)^{-1} \mathbf{e}_t)}_{(e)}.
\end{aligned}$$

Consequently we bound each of these terms individually. We begin with (b):

$$\begin{aligned}
(b) : \quad &\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K(\theta_{t-1})^{-1} + \xi I) \mathbf{g}_t - h_t(K(\theta_{t-1})^\dagger \mathbf{g}_t)) \\
&\stackrel{(i)}{\leq} \|\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]\| \|h_t [\psi_{\theta_t}(s, a)]^T (K(\theta_{t-1})^\dagger - (K(\theta_{t-1}) + \xi I)^{-1}) \mathbf{g}_t\| \\
&\leq \sqrt{\mathbb{E}_{s,a \sim d_*} [\|\psi_{\theta_t}(s, a)\|^2]} h_t C_{20} \xi^2 \|\mathbf{g}_t\| \\
&\stackrel{(ii)}{\leq} \sqrt{\left\| \frac{d_*}{d_{\theta_t}}(s, a) \right\|_\infty} \mathbb{E}_{s,a \sim d_{\theta_t}} [\|\psi_{\theta_t}(s, a)\|^2 + 1] h_t C_{20} \xi^2 \|\mathbf{g}_t\| \\
&\stackrel{(iii)}{\leq} h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} \sqrt{\psi_{\infty,1} + 1} C_{20} \xi^2 \|\mathbf{g}_t\|,
\end{aligned}$$

where in (i) we apply Lemma 10, and in (ii) we change the measure from d_* to d_{θ_t} and use the Triangle inequality, and in (iii) we apply that $d_\theta \geq (1-\gamma)\rho$ for all θ . For (a), we get:

$$\begin{aligned}
(a) : \quad &\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T (h_t(K_t + \xi I)^{-1} \mathbf{g}_t - h_t(K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t) \\
&\stackrel{(i)}{\leq} h_t \|\mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]\|_2 \|(K_t + \xi I)^{-1} \mathbf{g}_t - (K(\theta_{t-1}) + \xi I)^{-1} \mathbf{g}_t\|_2 \\
&\stackrel{(ii)}{\leq} h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} \sqrt{\psi_{\infty,1} + 1} \left[\frac{2}{\xi} \right] \|\mathbf{g}_t\|,
\end{aligned}$$

where in (i) we use the Cauchy-Schwarz inequality, in (ii) we use the boundedness $\|K_t + \xi I\| \geq \xi$, $\|K(\theta_t) + \xi I\| \geq \xi$.

$$\begin{aligned}
(c) : \quad & \mathbb{E}_{s,a \sim d_*} \left[\psi_{\theta_t}(s, a)^T h_t K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t}(s, a) \right] \\
&= h_t \mathbb{E}_{s,a \sim d_{\theta_t}} \left[\left[\frac{d_*(s, a)}{D d_{\theta_t}(s, a)} \right] \psi_{\theta_t}(s, a)^T K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t}(s, a) \right] \\
&\stackrel{(i)}{\leq} h_t \sqrt{\left\| \frac{d_*}{D d_{\theta_t}}(s, a) \right\|_\infty} \sqrt{\mathbb{E}_{s,a \sim d_{\theta_t}} ((\psi_{\theta_t}(s, a)^T K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t}(s, a))^2)} \\
&\stackrel{(ii)}{\leq} h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} \sqrt{\mathbb{E}_{s,a \sim d_{\theta_t}} [(\psi_{\theta_t}(s, a)^T K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t}(s, a))^2]} \\
&\stackrel{(iii)}{\leq} h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} \sqrt{E_\Pi},
\end{aligned}$$

where in (i) we use our observation that the Radon-Nikodym derivative is bounded, and in (ii) we use that $\left\| \frac{d_*(s, a)}{d_{\theta_t}(s, a)} \right\|_\infty \leq \frac{1}{1-\beta_0} \left\| \frac{d_*(s, a)}{d_{\theta_0}(s, a)} \right\|_\infty$ (this is a simple extension of $d_{\theta, \rho}(s, a) \geq (1-\beta_0)\rho(s)$, and in (iii) we substitute the condition number of the policy class: $E_\Pi = \max_{\theta_t} \left\| \psi_{\theta_t}^T K(\theta_{t-1})^\dagger \mathbf{g}_t - A_{\theta_t} \right\|_{d_{\theta_t}}^2$.

$$\begin{aligned}
(d) : \quad & \mathbb{E}_{s,a \sim d_*} [h_t A_{\theta_t}(s, a)]^T = h_t(1-\gamma) \mathbb{E}_{s,a \sim d_*} [Q_{\theta_t}(s, a) - V_{\theta_t}(s)] \\
&\stackrel{(i)}{=} h_t(1-\gamma) [J(\pi_*) - J(\pi_{\theta_t})],
\end{aligned}$$

by the performance difference lemma. Lastly we recall that the noise term is bounded by:

$$(e) : \mathbb{E}_{s,a \sim d_*} [\psi_{\theta_t}(s, a)]^T \mathbb{E}(h_t(K_t + \xi I)^{-1} e_t) \leq h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} \frac{\sigma}{\xi \sqrt{B}}.$$

Finally, combining all of these inequalities together, we get:

$$\begin{aligned}
& \mathbb{E} [D_t(\theta_t, s) - D_{t-1}(\theta_{t-1}, s)] \\
&\geq h_t(1-\gamma) \mathbb{E} [J(\pi_*) - J(\pi_{\theta_t})] - \mathbb{E} \frac{L h_t^{\beta_0+1}}{\beta_0+1} \|(K_t + \xi I)^{-1} \tilde{\mathbf{g}}_t\|^{\beta_0+1} \\
&\quad - h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} (\psi_{\infty,1} + 1) \left[\frac{\sigma}{\xi \sqrt{B}} + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1} + 1}} + \left(\frac{C_{20} \xi^2 + 2}{\xi} \right) \|\mathbf{g}_t\| \right] \\
&\geq h_t(1-\gamma) \mathbb{E} [J(\pi_*) - J(\pi_{\theta_t})] - \frac{L h_t^{\beta_0+1}}{\xi(\beta_0+1)} \mathbb{E} [\|e_t\|^{\beta_0+1}] - \frac{L h_t^{\beta_0+1}}{\xi(\beta_0+1)} \|\mathbf{g}_t\|^{\beta_0+1} \\
&\quad - h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} (\psi_{\infty,1} + 1) \left[\frac{\sigma}{\xi \sqrt{B}} + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1} + 1}} + \left(\frac{C_{20} \xi^2 + 2}{\xi} \right) \|\mathbf{g}_t\| \right] \\
&\stackrel{(i)}{\geq} h_t(1-\gamma) \mathbb{E} [J(\pi_*) - J(\pi_{\theta_t})] - \frac{L \tilde{\sigma} h_t^{\beta_0+1} B^{-\frac{\beta_0+1}{2}}}{\xi(\beta_0+1)} - \frac{L h_t^{\beta_0+1}}{\xi(\beta_0+1)} \|\mathbf{g}_t\|^{\beta_0+1} \\
&\quad - h_t \sqrt{\frac{1}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty} (\psi_{\infty,1} + 1) \left[\frac{\sigma}{\xi \sqrt{B}} + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1} + 1}} + \left(\frac{C_{20} \xi^2 + 2}{\xi} \right) \|\mathbf{g}_t\| \right],
\end{aligned}$$

where (i) follows by interpolation (note that e_t is L_1 by assumption). We replace the term $\left\| \frac{d_*}{\rho}(s, a) \right\|_\infty$ with C_{21} for notational convenience.

However, noting that $D_t(\theta_t, s) - D_{t-1}(\theta_{t-1}, s) \leq \frac{C_{\nu,1}}{1-\gamma} \left\| \frac{d_*}{\rho}(s, a) \right\|_\infty \|\theta_t - \theta_{t-1}\|^{2\beta_0}$ by Assumption 1, we find the following inequality:

$$\begin{aligned} J(\pi_*) - \mathbb{E}[J(\theta_t)] &\leq \frac{2C_{\nu,1}C_{21}}{(1-\gamma)^2} h_t^{2\beta_0-1} \left(\sigma^{2\beta_0} B^{-\beta_0} + \|\mathbf{g}_t\|^{2\beta_0} \right) + \frac{Lh_t^{\beta_0}}{\xi(\beta_0+1)(1-\gamma)} \left(\tilde{\sigma} B^{-\frac{\beta_0+1}{2}} + \|\mathbf{g}_t\|^{\beta+1} \right) \\ &\quad + \sqrt{\frac{(\psi_{\infty,1}+1)}{(1-\gamma)^3}} C_{21} \left(\frac{\sigma}{\xi\sqrt{B}} + \frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1}+1}} + \left(\frac{C_{20}\xi^2+2}{\xi} \right) \|\mathbf{g}_t\| \right). \end{aligned}$$

It remains to take the minimum of the gradient over $t = 1 \dots T$.

Proof of Corollary 2, NPG:

It remains now to substitute the learning rate and gradient bounds from each of the cases we analyzed earlier; doing so obtains the rates found in the following table. Since $\beta_1 \geq 1$, the final term always dominates the order if ϵ is sufficiently small. Finally, we use everywhere that $(\sum_i x_i)^p \leq \sum_i x_i^p$ for $p \in [0, 1]$.

Case I, $h_t = \lambda$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^2 = O(T^{-1}) + O\left(\frac{1}{1-\gamma} B^{-\frac{\beta_0+1}{2}}\right) + O\left((1-\gamma)^{\frac{-2}{1-\beta_0}}\right)$.

Subsequently we require $T^{-\frac{1}{2}} \leq \epsilon(1-\gamma)^{3/2}$ and $B^{-\frac{\beta_0+1}{4}} \leq \epsilon(1-\gamma)^2$, to obtain the following:

$$\min J_* - \mathbb{E}[J(\theta_t)] \leq \epsilon + \sqrt{\frac{(\psi_{\infty,1}+1)}{(1-\gamma)^3}} C_{21} \left(\frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1}+1}} + O\left((1-\gamma)^{\frac{-1}{\beta_0-1}}\right) \right).$$

Case II, $h_t = \lambda T^{\frac{\beta_0-1}{\beta_0+1}}$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^2 = O((1-\gamma)^{\frac{-2}{1-\beta_0}} T^{-\frac{2\beta_0}{\beta_0+1}}) + O\left(\frac{1}{1-\gamma} T^{\frac{\beta_0^2-\beta_0}{\beta_0+1}} B^{-\frac{\beta_0+1}{2}}\right)$.

Subsequently we require $T^{-\frac{\beta_0}{\beta_0+1}} \leq \epsilon(1-\gamma)^{\frac{3}{2} + \frac{1}{1-\beta_0}}$ and $B^{-\frac{\beta_0+1}{4}} \leq \epsilon^{\beta_0} (1-\gamma)^{\frac{3\beta_0-1}{2}}$, to obtain the following:

$$\min J_* - \mathbb{E}[J(\theta_t)] \leq \epsilon + \sqrt{\frac{(\psi_{\infty,1}+1)}{(1-\gamma)^3}} C_{21} \frac{\sqrt{E_\Pi}}{\sqrt{\psi_{\infty,1}+1}}. \quad (22)$$

Case III, $h_t = \left(\frac{\lambda(\beta_0+1)}{L} \|\mathbf{g}_t\|^{1-\beta_0}\right)^{\frac{1}{\beta_0}}$: Recall that in this case, $T^{-1} \sum_{t \leq T} \|\mathbf{g}_t\|^{\frac{\beta_0+1}{\beta_0}} = O((1-\gamma)^{\frac{-1}{\beta_0}} T^{-1}) + O\left(B^{-\frac{\beta_0+1}{2}}\right)$.

Subsequently we require $T^{-\frac{\beta_0}{\beta_0+1}} \leq \epsilon(1-\gamma)^{\frac{5+2\beta_0}{2}}$, and $B^{-\frac{\beta_0}{2}} \leq \epsilon(1-\gamma)^{\frac{3}{2}}$, to obtain (22).

D Learning Rate Estimators

We discuss the difficulties in estimating the optimal learning rate without strong additional assumptions.

Table 6: 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	$\epsilon^2(1-\gamma)^3$	$\epsilon^{\frac{4}{1+\beta_0}}(1-\gamma)^{\frac{8}{1+\beta_0}}$	Growing bias term
$\lambda T^{\frac{\beta_0-1}{\beta_0+1}}$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5-3\beta_0)(\beta_0+1)}{2(\beta_0-\beta_0^2)}}$	$\epsilon^{\frac{4\beta_0}{\beta_0+1}}(1-\gamma)^{\frac{6\beta_0-2}{\beta_0+1}}$	
$O\left(\ \mathbf{g}_t\ ^{\frac{1-\beta_0}{\beta_0}}\right)$	$\epsilon^{\frac{\beta_0+1}{\beta_0}}(1-\gamma)^{\frac{(5+2\beta_0)}{2\beta_0}}$	$\epsilon^{\frac{2}{\beta_0}}(1-\gamma)^{\frac{3}{\beta_0}}$	Not practically estimable

Suppose we set the learning rate using the sampled gradient, with an additional factor c :

$$\hat{h}_t = c\lambda \left[\frac{\beta_0 + 1}{L} \right]^{\frac{1}{\beta_0}} \|\tilde{\mathbf{g}}_t\|^{\frac{1-\beta_0}{\beta_0}}, \quad (23)$$

where $\tilde{\mathbf{g}}_t = r(s_t, a_t)\psi_{\theta_t}(s_t, a_t)$ is the sampled gradient. We choose $c = \min(1, \frac{\beta_0}{1-\beta_0})$ to manage the subsequent inequalities. We show that policy gradient using this learning rate attains roughly the same rate as with the true unknown parameter h_t , if we can apply strong upper bounds on the error $\|e_t\|$.

Our analysis becomes the following:

$$\begin{aligned} \hat{h}_t &= c\lambda \left[\frac{\beta_0 + 1}{L} \right]^{\frac{1}{\beta_0}} \|\mathbf{g}_t + e_t\|^{\frac{1-\beta_0}{\beta_0}} \\ &\leq \lambda \left[\frac{\beta_0 + 1}{L} \right]^{\frac{1}{\beta_0}} \left[\|\mathbf{g}_t\|^{\frac{1-\beta_0}{\beta_0}} + \|e_t\|^{\frac{1-\beta_0}{\beta_0}} \right], \end{aligned} \quad (24)$$

using a Young's inequality result (since $\frac{1-\beta_0}{\beta_0} < 1$ for $\beta_0 \in (1/2, 1]$). We would like to apply a lower bound $\hat{h}_t \geq \lambda \left[\frac{\beta_0+1}{L} \right]^{\frac{1}{\beta_0}} \left[\|\mathbf{g}_t\|^{\frac{1-\beta_0}{\beta_0}} - \frac{\kappa}{\sigma} \right]$, by assuming that $\|e_t\|$ is bounded. However, this is difficult without additional assumptions; we can attempt to bound the magnitude of $\|e_t\|$ via Chebyshev's inequality; it is required that this κ be extremely small for tight bounds, which requires strong bounds for Assumption 2. Instead we directly apply the following assumption:

Assumption 8. *The sample-based learning rate*

$$\hat{h}_t = c\lambda \left[\frac{\beta_0 + 1}{L} \right]^{\frac{1}{\beta_0}} \|\tilde{\mathbf{g}}_t\|^{\frac{1-\beta_0}{\beta_0}},$$

satisfies

$$\hat{h}_t \leq \lambda \left[\frac{\beta_0 + 1}{L} \right]^{\frac{1}{\beta_0}} \left[\|\mathbf{g}_t\|^{\frac{1-\beta_0}{\beta_0}} - \frac{\kappa}{\sigma} \right],$$

where κ_t is some sequence such that the sum

$$\sum_{t=0}^{\infty} \|\mathbf{g}_t\|^2 \frac{\kappa_t}{\sigma} \leq \frac{\epsilon}{2},$$

is finite for our ϵ of choice.

We then note that, for example, in policy gradient, the analysis now becomes:

$$\begin{aligned}
J(\theta_t) &\leq J(\theta_{t-1}) - \langle \hat{h}_t \mathbf{g}_t, \mathbf{g}_t \rangle + \frac{L}{\beta_0 + 1} \hat{h}_t^{\beta_0+1} \|\mathbf{g}_t - v_t \psi_\theta(s_t, a_t)\|^{\beta_0+1} + \frac{L}{\beta_0 + 1} \|\hat{h}_t \mathbf{g}_t\|^{\beta_0+1} \\
&\leq J(\theta_{t-1}) - \hat{h}_t \left[\|\mathbf{g}_t\|^2 - \frac{L}{\beta_0 + 1} \hat{h}_t^{\beta_0} \left[\|e_t\|^{\beta_0+1} + \|\mathbf{g}_t\|^{\beta_0+1} \right] \right] \\
&\stackrel{(i)}{\leq} J(\theta_{t-1}) - \hat{h}_t \left[\|\mathbf{g}_t\|^2 - \lambda^{\beta_0} \left[\|\mathbf{g}_t\|^{1-\beta_0} + \|e_t\|^{1-\beta_0} \right] \left[\|e_t\|^{\beta_0+1} + \|\mathbf{g}_t\|^{\beta_0+1} \right] \right] \\
&\stackrel{(ii)}{\leq} J(\theta_{t-1}) - \hat{h}_t \left[\|\mathbf{g}_t\|^2 - \lambda^{\beta_0} \left[\|e_t\|^2 + (1 + \|\mathbf{g}_t\|^2)(\|e_t\|^{1-\beta_0} + \|e_t\|^{\beta_0+1}) + \|\mathbf{g}_t\|^2 \right] \right] \\
&\stackrel{(iii)}{\leq} J(\theta_{t-1}) - \hat{h}_t \left[\|\mathbf{g}_t\|^2 - 3\lambda^{\beta_0} \left[\|e_t\|^2 + \|\mathbf{g}_t\|^2 (1 + \|e_t\|^2) + 1 \right] \right],
\end{aligned}$$

where in (i) we use another Young's inequality and that $\beta \leq 1$, and in (ii) we use that $\|\mathbf{g}_t\|^\alpha < (1 + \|\mathbf{g}_t\|^2)$ for $\alpha < 2$, and in (iii) we combine terms (somewhat loosely). Now we rearrange, and take expectation to get:

$$\mathbb{E} \left[\hat{h}_t \left[\|\mathbf{g}_t\|^2 - 3\lambda^{\beta_0} \left[\|e_t\|^2 + \|\mathbf{g}_t\|^2 (\|e_t\|^2 + 1) + 1 \right] \right] \right] \leq \mathbb{E} [J(\theta_{t-1}) - J(\theta_t)]. \quad (25)$$

Clearly if we choose λ_t to be sufficiently small, the difference in the interior becomes positive and thus we have contraction. Telescoping, we find:

$$\begin{aligned}
\mathbb{E} [J(\theta_0)] - J(\theta_*) &\geq \mathbb{E} \left[\sum_t \hat{h}_t \left[\|\mathbf{g}_t\|^2 - 3\lambda^{\beta_0} \left[\|e_t\|^2 + \|\mathbf{g}_t\|^2 (\|e_t\|^2 + 1) + 1 \right] \right] \right] \\
&\geq \mathbb{E} \left[\sum_t h_t \left[\|\mathbf{g}_t\|^2 - 3\lambda^{\beta_0} \left[\sum \|e_t\|^2 + \sum \|\mathbf{g}_t\|^2 (\|e_t\|^2 + 1) + 1 \right] \right] \right] - \sum_t \mathbb{E} \left[\|\mathbf{g}_t\|^2 \right] \frac{\kappa_t}{\sigma},
\end{aligned} \quad (26)$$

which is identical to the analysis we found earlier in the optimization portion, with slight adjustments to the noise and gradient terms, and the addition of a new error term due the error in the gradient estimator. Since the contribution of this term is bounded by $\frac{\epsilon}{2}$, we are free to choose a similar T as before.

Note that in general, it is difficult to minimize the additional error caused by estimating $\|\mathbf{g}_t\|$ at each iteration, since the variance on $\frac{e_t}{\mathbf{g}_t}$ cannot be easily controlled non-asymptotically; we must aggregate over time, or consider multiple queries for the gradient: $\hat{\mathbf{g}}_t^{(k)} = r(s_t^{(k)}, a_t^{(k)}) \psi_{\theta_t}(s_t^{(k)}, a_t^{(k)})$ where $s_t^{(k)} \sim d_{\theta_t, \rho}, a_t^{(k)} \sim \pi(s_t^{(k)})$ are sampled i.i.d. from their distributions. Under additional moment assumptions on the noise term, we can apply a Central Limit Theorem result to each \hat{h}_t and control the variance with high probability.

In such cases, we are required to pay additional sample complexity cost to control \hat{h} ; these additional samples can also be used to reduce the variance of the gradient, but we leave such analysis to future work. Furthermore, there are other ways to estimate the norm of \mathbf{g}_t with higher efficiency, but we only demonstrate the simplest such scheme here for the sake of completeness.

E Experiments

All results of these experiments can be found in the two Figures in the main text, and they are primarily intended to demonstrate practical examples of policies that satisfy our assumptions. All experiments were run locally on a single CPU, with 1Gb of dedicated RAM. The experiments are not computationally intensive and are primarily illustrative.

E.1 Generalized Gaussian Policies

For Figure 1(a), we sampled 4000 actions from a generalized Gaussian with mean $\langle s, \theta \rangle$ with $\theta \in \mathbb{R}^2$ and parameter $\kappa = 1.2$, for a state $[1.0, 0]$ in \mathbb{R}^2 . We let the first coordinate $\theta^{(1)}$ range from $[-2, 2]$ and compute the score function is then computed for these actions. The magnitude of the difference (compared to the reference point, $\theta^{(1)} = 0$) is then generated, and plotted against the standard Gaussian ($\kappa = 2$). The tail growth properties as well as local smoothness can be seen clearly from the Figure.

For the exploration problem, we implement the reward function described in Equation (10) with parameter $\theta^* = 3.9$ and initial parameter $\theta_0 = 0$. We aggregate gradients using a batch size of 1000 and update using our policy gradient rule. The performance is then reported on a batch of 1000, with standard deviations reported within this batch. We ran this experiment for 5 seeds and found that the qualitative behaviour does not change (although the inflection point of the policy behaviour is highly random, the Gaussian policy is always significantly worse than the generalized Gaussian).

E.2 Unbounded Gradient

We devise a simple policy and show that its gradient appears to be unbounded (in fact scaling logarithmically with the number of samples), but has a bounded integral according to its own measure π . For this, we simply implement the policy given in Example 3, and sample larger batches of actions. We then compute the gradients, and compare two norms that have been used in assumptions, that is the norm $\max_{n=1, \dots, K} \|\psi(s, a_n)\|$ and our assumption $\frac{1}{K} \sum_{n=1}^K \|\psi(s, a_n)\|^2$. We find that the former is unbounded for larger sets while the latter converges smoothly to a fixed value. As we explain in our paper, this shows that interesting policies can be covered by our assumptions while not being allowed in prior work.

E.3 Ergodicity

We measure a necessary condition for ergodicity for the Mountain Car environment, which can be found in many open-source repositories. We make some modifications to allow the dynamics to be more continuous; we remove the clipping at the boundary found in the standard implementation, and instead make the walls of the system scale quadratically (i.e. the cart experiences a force proportional to its displacement from the origin). This becomes a strong attractor system. We can also add a dampening term to the velocity, and this improves ergodicity; this is not done for the result found in the Figure.

Then, we measure the convergence of the test function $\zeta = \mathbb{E}[\|\phi(s, a)\|]$ by computing its average within 20000 trajectories. Since the distribution is not symmetric, instead of reporting standard deviations, we measure a 0.95 confidence interval within each batch and plot this in the shaded region. This value converges smoothly to a stable limit as $t \rightarrow \infty$. We can also propose to measure other metrics by varying the test function, however we cannot establish ergodicity with complete certainty through such empirical tests. It can also be seen that the estimates are quite noisy even for such large batch sizes; in general ergodicity cannot be empirically verified and can only be estimated based on loose argumentation.