On the Convergence Rates of Log-Linear Policy Gradient Methods

Matin Aghaei*, Anderson de Andrade*, Qiushi Lin*, Sharan Vaswani

{matin_aghaei,anderson_de_andrade,qiushi_lin,sharan_vaswani}@sfu.ca

Simon Fraser University

Abstract

Policy gradient methods are wildly used today because of their versatility. The method decouples the possibly unknown state transition function from the policy parameterization and learning process. Inspired by the functional mirror ascent framework of Vaswani et al. [2022] that generalizes many surrogate functions in the reinforcement learning literature, we propose projected policy gradient and natural policy gradient methods for log-linear parameterizations and analyze their convergence rate in exact and stochastic gradient settings. The proposed methods allows us to extend some of the convergence results presented for policy gradients in a tabular setting to the log-linear setting. For the first time, we have shown for the log-linear policy gradient method a convergence rate of $\mathcal{O}(\sqrt{1/T + b})$ with exact updates, and $\mathcal{O}(\sqrt[3]{1/T + b})$ with inexact updates. The proposed projected NPG method has the usual linear convergence as in the tabular softmax settings but the approximation error does not depend on the distribution mismatch or concentrability coefficients. We also show a convergence rate for the projected NPG method with inexact evaluation. Empirically, we perform experiments in the exact gradient setting to validate our results of convergence rates.

1 Motivation

Policy gradient (PG) methods Williams [1992], Sutton et al. [1999] optimize a parameterized policy with respect to the expected long-term cumulative reward using gradient descent. They are an important class of reinforcement learning methods because of their versatility: (i) the policy representation can be chosen to be useful for the task, (ii) it often has fewer parameters than value-function approaches, (iii) they can be used either *model-free* Williams [1992], Liu et al. [2023] or *model-based* Wang and Dietterich [2003], Deisenroth and Rasmussen [2011], Kurutach et al. [2018]. Model-free methods avoid explicitly estimating the transition probability distribution and the reward function and are the methods with the most prominence in the policy gradient literature.

Many policy gradient methods enjoy good empirical performance but initially lacked strong theoretical guarantees. More recently, theoretical support has been developed for many of these methods under different settings Agarwal et al. [2021]. Convergence rates have been shown for a tabular parameterization, where the space of states and actions is finite Xiao [2022], Mei et al. [2020]. For continuous spaces or problems in which the sets of states and actions are large, function approximation methods Sutton et al. [1999] are necessary. In this setting, convergence rates have been shown for natural policy gradient updates using a log-linear approximation Agarwal et al. [2021], Yuan et al. [2023].

Although the natural policy gradient (NPG) method enjoys faster convergence rates in general, policy gradients (PG) still has advantages in stochastic settings and it is very commonly used in

* denotes equal contributions (alphabetical order)

practice. Thus, the importance of its analysis remains high. However, for the policy gradient method, convergence rates have not been shown yet. Most recently, under strong assumptions on the features, global convergence has been shown Mei et al. [2023].

Having guarantees of convergence allows us to understand existing methods and their flaws, derive new solutions to address their issues, and more generally, guide future research. Some methods can be inherently flawed or proven to be optimal, and as a result, they do not warrant any further development, while others can be improved using the new understanding provided.

2 Problem Formulation

Consider infinite-horizon discounted Markov decision processes (MDP) defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}, \mathcal{A} \subseteq \mathbb{R}$ are the state and action spaces respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, and $\gamma \in [0, 1)$ is a discount factor, where $\Delta(\mathcal{X})$ is the probability simplex for an arbitrary set $\mathcal{X}$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces an *occupancy measure* over states given as:

$$d^{\pi}(s) \triangleq \mathbb{E}_{s_0 \sim \rho}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\pi(\cdot, s_t)}\Big[\mathbb{P}[S_t = s_t | S_0 = s_0]\Big]\right], \tag{1}$$

where $\rho \in \Delta(\mathcal{S})$ is a probability distribution over the initial states, and $\mathbb{P}[S_t = s_t | S_0 = s_0]$ can be deduced by the iteration of $\mathcal{P}$ induced by π. We can extend this measure to actions with $\mu^{\pi}(s, a) \triangleq d^{\pi}(s)\, \pi(a, s)$, which is known as the *state-action occupancy measure*. Furthermore, we can define the value function as $V^{\pi}(\rho) = \langle \mu^{\pi}, r \rangle$. Given ρ, there exists an optimal policy $\pi^* \triangleq \arg\max_{\pi \in \Delta(\mathcal{A})} V^{\pi}(\rho)$ and a best feasible policy $\widehat{\pi} \triangleq \arg\max_{\pi \in \mathbf{\Pi}} V^{\pi}(\rho)$, where $\mathbf{\Pi}$ is a set of feasible policies defined by the policy parameterization. In log-linear policy gradients, we parameterize the policy with $\theta \in \mathbb{R}^d$. In this work, we pointedly choose to *functionally* represent the policy as $\pi_\theta(a, s) = \langle \phi(s, a), \theta \rangle$, where for each state-action pair (s, a), there is a feature mapping $\phi(s, a) \in \mathbb{R}^d$. In a direct representation, the policy probability distribution is defined as $p^{\pi_\theta}(s) = \pi_\theta(\cdot, s)$, whereas in the softmax representation, the functional representation are considered logits to a softmax function, such that:

$$p^{\pi_\theta}(s) = \frac{\exp(\pi_\theta(\cdot, s))}{\sum_{a \in \mathcal{A}} \exp(\pi_\theta(a, s))}. \tag{2}$$

Different policy gradient methods can be derived from the mirror-ascent generalization, in which we take a step in a dual space and then project to a feasible solution in the primal space. Updates for parameters w at step $t + 1$ are given as:

$$w_{t+1} = \arg\max_{w \in \boldsymbol{\mathcal{W}}} \left[\langle \nabla_w J(w_t) | w \rangle - \frac{1}{\eta_t} D_\psi(w, w_t)\right], \tag{3}$$

where $D_\psi(\cdot, \cdot)$ is a distance-like function using a strictly convex, differentiable distance measure function, or mirror map $\psi(\cdot)$, η_t is the step size at update t, and $J(w)$ is an objective function. By setting D_ψ as a quadratic approximation of the Kullback-Leibler (KL) divergence, one recovers the natural policy gradient (NPG) method Kakade [2001], with updates of the form $w_{t+1} = w_t + \eta_t F_\rho(w_t)^{\dagger} \nabla_w J(w_t)$, where $F_\rho(\theta)^{\dagger}$ is the Moore-Penrose psudoinverse of the Fisher information matrix, used as a pre-conditioner to improve the policy gradient direction. If we set D_ψ as the squared Euclidean distance using the square norm as the mirror map, we obtain the projected policy gradient (PPG) method Xiao [2022] with updates $w_{t+1} = \text{proj}_{\mathcal{W}}[w_t + \eta_t \nabla_w J(w_t)]$, where $\mathcal{W}$ is a constrained set of feasible solutions. If we remove the projection, we recover the vanilla policy gradient method, with updates given as $w_{t+1} = w_t + \eta_t \nabla_w J(w_t)$.

In many of the existing convergence rate analyses, two coefficient naturally arise:

$$\left\| \frac{d_\rho^{\pi^*}}{d_\rho^{\pi}} \right\|_\infty = \max_{s \in \mathcal{S}} \frac{d_\rho^{\pi^*}(s)}{d_\rho^{\pi}(s)} \leq \nu_\rho, \quad \text{and} \quad \mathbb{E}_{s \sim d_\rho^*}\left[\left(\frac{d_\rho^{\pi_t}(s)}{d_\rho^{\pi^*}(s)}\right)^2\right] \leq C_\rho. \tag{4}$$

The distribution mismatch coefficient ν_ρ encapsulates the importance of initialization. If states that produce high rewards have no much probability of being explored, the coefficient will be high.

Table 1: Comparison of softmax policy gradient methods with theoretical guarantees.

Algorithm	Parameterization	Gradient	Results
Mei et al. [2020]	tabular	exact	achieved $O(1/T)$ convergence for PG
Mei et al. [2021]	tabular	exact	NPG converges with linear rate
		inexact	NPG may not converge, whereas PG converges in probability
Cen et al. [2022]	tabular	exact	NPG with entropy regularization achieves linear convergence rate
Agarwal et al. [2021]	linear	inexact	NPG converges to the neighbourhood of the optimal policy
Yuan et al. [2023]	linear	inexact	NPG achieves linear convergence to the neighbourhood of the optimal policy
Ours	linear	exact	$\mathcal{O}(\sqrt{1/T + b})$ convergence for PG and linear for NPG with unamplified bias
		inexact	$\mathcal{O}(\sqrt[3]{1/T + b})$ for PG and linear for NPG with unamplified bias

Similarly, if a low probability is placed on a state-action pair of the initial policy, the coefficient will be high if the optimal policy has a high probability assigned to the same state-action pair. Another term shown in convergence analyses is the concentrability coefficient C_ρ Munos [2005]. It measures how much ρ can get amplified in T steps compared to the reference distribution d^{π^*}. A finite concentrability is a restriction placed by the MDP dynamics, while the mismatch ratio does not require restrictions on the MDP dynamics Chen and Jiang [2019], Agarwal et al. [2021].

With the log-linear parameterization defined above, we design projected methods for policy gradients and natural policy gradients and show that as we perform updates to the policy, we converge to the optimal feasible solution π^* at a rate of $\mathcal{O}(1/\sqrt{T})$ in PG and at a linear rate in NPG, using exact updates. In the NPG convergence result, the bias term is not amplified by the distribution mismatch or concentrability coefficients.

3 Related Work

Table 1 lists the convergence rates proved in recent works under different settings using the *softmax* functional representation. For policy gradients, the work of Mei et al. [2020] shows a rate of $\mathcal{O}(1/t)$ for tabular softmax policies using exact gradients. Their analysis relies on three findings: that the objective satisfies a smoothness constant of $5/2$ for bounded rewards in $[0, 1]$, that it meets the uniform Łojasiewicz condition Lojasiewicz [1963], and that the minimum probability of an optimal action during optimization can be bounded in terms of its initial value.

The subsequent work of Mei et al. [2021] extends the analysis to stochastic gradients. The study concludes that in this setting, an uninformed algorithm such as the PG method converges to a globally optimal policy with probability 1 but at a rate no better than $\mathcal{O}(1/t)$, and that methods with faster convergence in the exact gradient setting, such as natural policy gradients (NPG), might fail to converge in the stochastic setting with some positive probability.

The Q-NPG method is a variation of NPG with the main difference being the use of function approximation for the Q-function instead of advantages. Agarwal et al. [2021] and Yuan et al. [2023] established convergence to the neighbourhood of the optimal policy using the Q-NPG method with linear parameterization in the stochastic setting. In their respective analyses, two error terms arise: the *excess risk* and the *approximation error*. The excess risk is an upper bound on the objective differences between an exact update and an estimated update. In stochastic settings, increasing the number of samples used to estimate the updates reduces this error. The approximation error is an upper bound of the objective differences between the optimal policy and the best feasible policy. In Agarwal et al. [2021], we note that the approximation error scales with the distribution mismatch coefficient. In cite Yuan et al. [2023], the approximation error scales with both the distribution mismatch and concentrability coefficients.

The work of Mei et al. [2023] shows asymptotic convergence for the PG method and a linear rate of convergence for the NPG method in a bandit setting with linear approximation, under strong assumptions on the features and rewards. A bandit setting corresponds to an infinite-horizon MDP in which $|\mathcal{S}| = 1$ and $\gamma = 0$. The PG result of Mei et al. [2023] assumes there exists a θ such

that $r' = \Phi\theta$ preserves the order of r such that for all $i, j \in \{|\mathcal{A}|\}, r(i) > r(j)$ if and only if $r'(i) > r'(j)$. This condition effectively enforces the optimal policy to be feasible. The work uses the same definition of approximation error to analyze both the PG and NPG methods, which does not seem justifiable since the PG method does not use the same regression problem to perform updates. The work shows that an approximation error of zero does not necessarily result in global convergence for the PG method, because it is less directly connected to the concept of approximation error, compared to NPG. Interestingly, their examples show that a non-zero approximation error does not necessarily prevent global convergence in both PG and NPG methods. Consequently, they reject the concept of approximation error, which we believe important to analyze.

The work of Vaswani et al. [2022] introduces a framework that defines the sufficient statistics of a policy as its *functional representation*, and their materialization as its *parameterization*. The same policy can have multiple functional representations. The policy parameterization defines the set of realizable policies that exits for a given parametric model and can be chosen independently of its functional representation. The framework is based on functional mirror ascent and gives rise to an entire family of surrogate functions for policy gradient methods. The work proposes surrogate functions that enable policy improvement guarantees.

4 Main Results

Algorithm 1 Projected PG, Exact Evaluation

Input: Φ, π_1, η, T.
Output: Policies $p^{\pi_t} = \text{softmax}(\pi_t)$.
while $t = 0$ to $T - 1$ **do**
  Improvement: $\pi_{t+1/2} = \pi_t + \eta_t \nabla_\pi V^{\pi_t}(\rho)$
  Projection: $\pi_{t+1} = \phi(\phi^\top \phi)^{-1}\phi^\top \pi_{t+1/2}$
end while

Algorithm 2 Projected NPG, Exact Evaluation

Input: Φ, π_1, η, T.
Output: Policies $p^{\pi_t} = \text{softmax}(\pi_t)$.
for $t = 0$ to $T - 1$ **do**
  Improvement: $\pi_{t+1/2} = \pi_t + \eta_t \frac{A^{\pi_t}}{1-\gamma}$
  Projection: $\pi_{t+1} = \phi(\phi^\top \phi)^{-1}\phi^\top \pi_{t+1/2}$
end for

4.1 Algorithms

Following Vaswani et al. [2021], we instantiate mirror-ascent to perform improvements in the functional space and project to a feasible solution in the parameter space. Eq. (3) can be also formulated in 2-steps. For the first step, or the functional improvement step, setting the value function as the objective function such that $J(\pi) = V^\pi(\rho)$, we follow the standard PG and NPG updates. The functional improvement for the policy gradients method follows the results of Sutton et al. [1999], where the direction of improvement is given by:

$$[\nabla_\pi J(\pi)]_{s,a} = d^\pi(s)\, \pi(s,a)\, \frac{A^\pi(s,a)}{1-\gamma}, \tag{5}$$

with $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ being the advantage function, where Q^π is the Q-function Sutton et al. [1999]. Similarly, for the natural policy gradients method, following Khodadadian et al. [2022], we have the following direction of functional improvement:

$$[F_\pi^\dagger \nabla_\pi J(\pi)]_{s,a} = \frac{A^\pi(s,a)}{1-\gamma}. \tag{6}$$

The second step, the projection step, uses the squared norm as the mirror map, having the following form:

$$\pi_{t+1} = \arg\min_{\pi \in \Pi} \left\| \pi - \pi_{t+1/2} \right\|_2^2, \tag{7}$$

where $\pi_{t+1/2}$ is the result of the intermediate functional improvement. In the linear setting, this projection step is equivalent to:

$$\theta_{t+1} = \arg\min_{\theta} \left\| \phi\, \theta - \pi_{t+1/2} \right\|_2^2, \tag{8}$$

becoming an unconstrained optimization problem with a closed-form solution given by:

$$\theta_{t+1} = (\phi^\top \phi)^{-1}\phi^\top \pi_{t+1/2}. \tag{9}$$

Algorithm 1 and Algorithm 2 describe the optimization procedure in detail for the projected PG and projected NPG methods respectively. The functional update in the projected PG algorithm is given by Eq. (5). A computational complexity analysis is deferred to Appendix D.

Noting that the gradients are taken with respect to the functional representation, established as the logits, the linear approximation is effectively abstracted away and we recover the tabular setting, which properties are well understood and useful in the following analyses. Namely, it has been shown in Mei et al. [2020] that the tabular softmax parameterization satisfies the smoothness and the non-uniform Łojasiewicz condition. Since the parameterization still restricts the set of feasible policies, there exists an error $\left|J(\pi_{t+1/2}) - J(\pi_{t+1})\right| \leq b_t$, incurred when not being able to achieve the optimal policy. We call b_t the projection bias at iteration t and it is akin to the approximation error introduced in some of the previous analyses discussed.

4.2 Convergence Rates

We show a convergence analysis for the proposed methods in both the exact evaluation setting and the inexact evaluation setting. We defer the empirical evaluation of these methods under the exact evaluation setting to Appendix C.

4.2.1 Projected PG with Exact Evaluation

Lemma 4.1 (Smoothness, Lemma 7 in Mei et al. [2020])**.** *$J(\pi)$ is $8/(1-\gamma)^3$-smooth.*

Lemma 4.2 (Non-uniform Lojasiewicz, Lemma 8 in Mei et al. [2020])**.**

$$\left\|\frac{\partial J(\pi)}{\partial \theta}\right\| \geq \frac{\min_s p^\pi(a^*(s)|s)}{\sqrt{S}\cdot\left\|d_\rho^{\pi^*}/d_\rho^\pi\right\|_\infty}\cdot\left[J(\pi^*) - J(\pi)\right] \tag{10}$$

Theorem 4.3. *(Projected PG) Assuming that the bias is bounded $|J(\pi_{t+1}) - J(\pi_{t+1/2})| < b_t$ for all t, after T rounds of the projected PG with $\eta_t = \frac{1}{L}$, we have*

$$\min_{t\in[T-1]}\delta \leq \sqrt{\frac{\delta_0 + \sum_{t=0}^{T-1} b_t}{\frac{\mu^2}{2L}T}}, \tag{11}$$

where $\mu = \inf_t \frac{\min_s p^{\pi_t}(a^*(s)|s)}{\sqrt{S}\cdot\left\|d_\rho^{\pi^*}/d_\rho^{\pi_t}\right\|_\infty}$.

Proof. Using the L-smoothness of J in Lemma 4.1 and the improvement step in Algorithm 1 with $\eta_t = \frac{1}{L}$, we have

$$J(\pi_{t+1/2}) \geq J(\pi_t) + \left\langle \nabla J(\pi_t), \frac{1}{L}\nabla J(\pi_t)\right\rangle - \frac{L}{2}\left\|\frac{1}{L}\nabla J(\pi_t)\right\|^2 \tag{12}$$

$$= J(\pi_t) + \frac{1}{2L}\left\|\nabla J(\pi_t)\right\|^2 \tag{13}$$

Therefore,

$$J(\pi_{t+1}) \geq J(\pi_{t+1/2}) - b_t \tag{14}$$

$$\geq J(\pi_t) + \frac{1}{2L}\left\|\nabla J(\pi_t)\right\|^2 - b_t \tag{15}$$

Using the Lojasiewicz condition in Lemma 4.2, we have

$$\geq J(\pi_t) + \frac{\mu^2}{2L}\left\|J(\pi^*) - J(\pi_t)\right\|^2 - b_t, \tag{16}$$

where $\mu = \inf_t \frac{\min_s p^{\pi_t}(a^*(s)|s)}{\sqrt{S}\cdot\left\|d_\rho^{\pi^*}/d_\rho^{\pi_t}\right\|_\infty}$. Furthermore,

$$J(\pi^*) - J(\pi_{t+1}) \leq J(\pi^*) - J(\pi_t) - \frac{\mu^2}{2L}\left\|J(\pi^*) - J(\pi_t)\right\|^2 + b_t \tag{17}$$

$$\implies \delta_{t+1} \leq \delta_t - \frac{\mu^2}{2L}\delta_t^2 + b_t \tag{18}$$

$$\implies \frac{\mu^2}{2L}\delta_t^2 \leq \delta_t - \delta_{t+1} + b_t \tag{19}$$

Summing up for T iterations and dividing both sides by T, we have

$$\frac{\mu^2}{2L}\min_{t\in[T-1]}\delta_t^2 \leq \frac{\mu^2}{2LT}\sum_{t=0}^{T-1}\delta_t^2 \leq \frac{1}{T}[\delta_0-\delta_T]+\frac{1}{T}\sum_{t=0}^{T-1}b_t \leq \frac{1}{T}[\delta_0]+\frac{1}{T}\sum_{t=0}^{T-1}b_t \tag{20}$$

$$\Longrightarrow \min_{t\in[T-1]}\delta_t \leq \sqrt{\frac{\delta_0+\sum_{t=0}^{T-1}b_t}{\frac{\mu^2}{2L}T}} \tag{21}$$

□

4.2.2 Projected PG with Inexact Evaluation

Theorem 4.4. *Let $\{\pi_t\}_{t\geq 0}$ be generated by using Algorithm 3, i.e., for all $t \geq 0$,*

$$\pi_{t+1/2} = \pi_t + \eta \cdot \widehat{g}_t \tag{22}$$

$$\pi_{t+1} = \phi(\phi^\top\phi)^{-1}\phi^\top\pi_{t+1/2}, \tag{23}$$

with learning rate

$$\eta = \frac{(1-\gamma)^4}{4\cdot C}\cdot\left\|\frac{\partial V^{\pi_t}(\rho)}{\partial\pi}\right\|_2 \tag{24}$$

for all $t \geq 0$, and

$$C \coloneqq \left[3+\frac{2\cdot(C_\infty-(1-\gamma))}{(1-\gamma)\cdot\gamma}\right]\cdot\sqrt{S}, \tag{25}$$

where $C_\infty \coloneqq \max_\pi\left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s\rho(s)} < \infty$. Denote $C'_\infty \coloneqq \max_\pi\left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty$. We have

$$\min_{t\in[T-1]}\mathbb{E}\left[V^*(\rho)-V^{\pi_t}(\rho)\right] \leq \sqrt[3]{\frac{\mathbb{E}\left[V^*(\rho)-V^{\pi_0}(\rho)\right]+\sum_{t=0}^{T-1}\mathbb{E}\left[b_t\right]}{\frac{(1-\gamma)^7}{8\cdot C}\cdot\left\|\frac{d_\rho^{\pi^*}}{\rho}\right\|_\infty^{-3}\cdot\frac{c}{S\cdot\sqrt{S}}\cdot T}}, \tag{26}$$

where $c > 0$ is independent with T.

The proof is included in Appendix A.

4.2.3 Projected NPG with Exact Evaluation

Here, we analyze the convergence rate for the project linear NPG with exact policy evaluation. Mei et al. has proven the Non-uniform Łojasiewicz inequality (Lemma 10) for natural policy gradient in tabular softmax settings. Based on that, we can easily derive the following lemma for the improvement step of the projected NPG.

Lemma 4.5 (Natural NŁInequality)**.** *Considering every improvement step of Algorithm 2, for all $s \in \mathcal{S}$, we have:*

$$J_s(\pi_{t+1/2})-J_s(\pi_t) \geq C(\pi_t)\cdot(1-\gamma)\cdot\left\|\frac{d_\rho^{\pi^*}}{\rho}\right\|_\infty^{-1}\left[J_s(\pi^*)-J_s(\pi_t)\right], \tag{27}$$

where $C(\pi_t)$ is given by

$$C(\pi_t) \coloneqq \min_{s\in\mathcal{S}}\left[1-\frac{1}{\pi_t(s,\bar{a}_t(s))\cdot(\exp(\eta\cdot\Delta_t(s))-1)+1}\right] \in (0,1), \tag{28}$$

and $\bar{a}_t(s) \coloneqq \arg\max_{a\in\mathcal{A}}Q^{\pi_t}(s,a)$ and $\Delta_t(s) \coloneqq Q^{\pi_t}(s,\bar{a}_t(s))-\max_{a\neq\bar{a}_t(s)}\{Q^{\pi_t}(s,a)\}$.

As RHS ≥ 0 in Eq. (27), we know that NPG in tabular softmax settings will result in monotonic value improvement, meaning that $J_s(\pi_{t+1/2}) \geq J_s(\pi_t)$ for all $t \geq 1$. We can further address the monotonic improvement for the correspond Q-values:

$$Q^{\pi_{t+1/2}}(s,a)-Q^{\pi_t}(s,a) = \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\left[J_s(\pi_{t+1})-J_s(\pi_t)\right] \geq 0. \tag{29}$$

Therefore, we have $Q^{\pi_{t+1/2}}(s,a) \geq Q^{\pi_t}(s,a)$ for all $t > 1$. Using this, we present the theorem for the convergence rate of the projected NPG as follows.

Theorem 4.6. *(Projected NPG) We define the step size as:*

$$\eta_t \geq \frac{1}{c_t} \max_{s \in \mathcal{S}} \left\{ \min_{p^\pi \in \Pi_t^s} D_\Phi(p^\pi, p^{\pi_t}) \right\}, \tag{30}$$

where c_t is a sequence of positive reals, $\Pi_t^s = \{\pi^s | \pi^s = \arg\max_{p^s \in \Delta(\mathcal{A})} \langle Q^{\pi_t}(s,\cdot), p^s \rangle\}$ is a set of policies w.r.t $Q^{\pi_t}(s,\cdot)$, and D_Φ denotes the Bregman divergence under the mirror map Φ. Assuming that the bias is bounded $|J(\pi_{t+1}) - J(\pi_{t+1/2})| < b_t$ for all t, after T rounds of the projected NPG with η_t, we have

$$\|J(\pi^*) - J(\pi_T)\|_\infty \leq \gamma^T \left[\|J(\pi^*) - J(\pi_0)\|_\infty + \sum_{t=1}^{T} \gamma^{-t}(c_t + b_t) \right]. \tag{31}$$

Proof. As mentioned above, we already know that $Q^{\pi_t}(s,\cdot) \leq Q^{\pi_{t+1/2}}(s,\cdot)$ for the projected NPG. Then, we have $\langle Q^{\pi_t}(s,\cdot), \pi^s_{t+1/2} \rangle \leq \langle Q^{\pi_{t+1/2}}(s,\cdot), \pi^s_{t+1/2} \rangle = J_s(\pi_{t+1/2})$. Using this, we have

$$\begin{aligned}
\langle Q^{\pi_t}(s,\cdot), \pi^{*s} - \pi^s_{t+1/2} \rangle &\geq \langle Q^{\pi_t}(s,\cdot), \pi^{*s} \rangle - J_s(\pi_{t+1/2}) && (32)\\
&\geq \langle Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot), \pi^{*s} \rangle + \langle Q^*(s,\cdot), \pi^{*s} \rangle - J_s(\pi_{t+1/2}) && (33)\\
&\geq -\|Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot)\|_\infty + J_s(\pi^*) - J_s(\pi_{t+1/2}) \quad \text{(Hölder's inequality)} && (34)
\end{aligned}$$

Next, we upper bound $\|Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot)\|_\infty$.

$$Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot) = \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)[J_{s'}(\pi_t) - J_{s'}(\pi^*)] \leq \gamma \|J(\pi_t) - J(\pi^*)\|_\infty. \tag{35}$$

Putting them together, we have

$$\begin{aligned}
&-\gamma \|J(\pi_t) - J(\pi^*)\|_\infty + J_{s'}(\pi^*) - J_{s'}(\pi_{t+1/2}) && (36)\\
&\leq \langle Q^{\pi_t}(s,\cdot), \pi^{*s} - \pi^s_{t+1/2} \rangle \leq Q^{\pi_t}(s,\cdot), \tilde{\pi}^s - \pi^s_{t+1/2} \rangle \quad \text{(define } \tilde{\pi} \text{ as the greedy policy w.r.t } Q^{\pi_t}(s,\cdot)) && (37)\\
&\leq \frac{D_\Phi(\tilde{\pi}^s, \pi^s_t) - D_\Phi(\tilde{\pi}^s, \pi^s_{t+1/2}) - D_\Phi(\pi^s_{t+1/2}, \pi^s_t)}{\eta_t} \quad \text{(Three-Point Descent Lemma Xiao [2022])} && (38)\\
&\leq \frac{D_\Phi(\tilde{\pi}^s, \pi^s_t)}{\eta_t} \leq \min_{p^\pi \in \Pi_t^s} \frac{D_\Phi(p^\pi, \pi^s_t)}{\eta_t} \leq c_t. \quad \text{(definition of } \eta_t) && (39)
\end{aligned}$$

From the above inequality, $-\gamma \|J(\pi_t) - J(\pi^*)\|_\infty + J_{s'}(\pi^*) - J_{s'}(\pi_{t+1/2}) \leq c_t$. Combining the bias bound, we have

$$\begin{aligned}
-\gamma \|J(\pi_t) - J(\pi^*)\|_\infty + J_{s'}(\pi^*) - J_{s'}(\pi_{t+1}) &\leq c_t + J_{s'}(\pi_{t+1/2}) - J_{s'}(\pi_{t+1}) && (40)\\
&\leq c_t + |J_{s'}(\pi_{t+1/2}) - J_{s'}(\pi_{t+1})| && (41)\\
&\leq c_t + b_t && (42)\\
\implies \|J(\pi^*) - J(\pi_{t+1})\|_\infty &\leq \gamma \|J(\pi_t) - J(\pi^*)\|_\infty + c_t + b_t && (43)
\end{aligned}$$

Unravelling this recursion yields

$$\|J(\pi^*) - J(\pi_T)\|_\infty \leq \gamma^T \left[\|J(\pi^*) - J(\pi_0)\|_\infty + \sum_{t=1}^{T} \gamma^{-t}(c_t + b_t) \right]. \tag{44}$$

□

This theorem states that the project linear NPG can approach the neighborhood of optimal policy with a linear convergence rate. We can use the geometrically increasing step size by setting $c_t = \gamma^t c$ for some constant c, then term $\gamma^T \sum_{t=1}^{T} \gamma^{-t} c_t \leq \gamma^T T c$ will diminish linearly. Moreover, we will be able to limit the neighbourhood size to $\max_t b_t / 1-\gamma$.

4.2.4 Projected NPG with Inexact Evaluation

Here, we extend the convergence rate for the project linear NPG in the inexact settings. We assume that we have a Q-estimator that is τ-accurate, i.e., for all π and s, $\left\|Q^{\pi}(s,\cdot) - \widehat{Q}^{\pi}(s,\cdot)\right\|_{\infty} \leq \tau$. The improvement step becomes $\pi_{t+1/2} = \pi_t + \eta_t \frac{\widehat{A}^{\pi_t}}{1-\gamma}$, where $\widehat{A}^{\pi_t}$ is induced by $\widehat{Q}^{\pi_t}$ and π_t. We then bound the inexact Q-values for NPG updates derived from Lemma A.5 in Johnson et al. [2023].

Lemma 4.7. *Consider the policies produced by Algorithm 2, if* $\left\|Q^{\pi}(s,\cdot) - \widehat{Q}^{\pi}(s,\cdot)\right\|_{\infty} \leq \tau$ *for any* π *and* s*, we have*

$$\widehat{Q}^{\pi_{t+1/2}}(s,a) \geq \widehat{Q}^{\pi_t}(s,a) - \frac{2\tau\gamma}{1-\gamma}. \tag{45}$$

Based on that, we can introduce a convergence rate for inexact projected NPG that is similar to Theorem 4.6.

Theorem 4.8. *(Inexact Projected NPG) We define the step size as:*

$$\eta_t \geq \max_{s\in\mathcal{S}} \left\{ \min_{p^{\pi}\in\Pi_t^s} \frac{D_{\Phi}(p^{\pi}, p^{\pi_t})}{\gamma^{2k+1}} \right\}, \tag{46}$$

where $\Pi_t^s = \{\pi^s | \pi^s = \arg\max_{p^s\in\Delta(\mathcal{A})} \langle Q^{\pi_t}(s,\cdot), p^s\rangle\}$ *is a set of policies w.r.t* $Q^{\pi_t}(s,\cdot)$*, and* D_{Φ} *denotes the Bregman divergence under the mirror map* Φ*. Assuming that the bias is bounded* $|J(\pi_{t+1}) - J(\pi_{t+1/2})| < b_t$*, after* T *rounds of the projected linear NPG with* η_t *and* τ*-accurate* Q*-estimates, we have:*

$$\|J(\pi^*) - J(\pi_T)\|_{\infty} \leq \gamma^T \left(\|J(\pi^*) - J(\pi_0)\|_{\infty} + \frac{1}{1-\gamma} \right) + \frac{4\gamma\tau}{(1-\gamma)^2} + \sum_{t=0}^{T-1} \gamma^t b_t. \tag{47}$$

The proof of this theorem can be found in Appendix B. This theorem shows that the projected linear NPG can also converge to the neighborhood of the optimal policy in a linear rate. The neighbourhood size depend on both the inexactness $4\gamma\tau/(1-\gamma)^2$ and the projection error $\sum_{t=0}^{T-1}\gamma^t b_t$, which can be bounded by $4\gamma\tau/(1-\gamma)^2 + \max_t b_t/1-\gamma$.

5 Conclusion and Future Work

In this work we designed a projected PG method and a projected NPG method for the linear approximation setting. We analyzed their convergence by extending the results from the tabular policy gradient methods. We achieved a convergence rate for policy gradients in a linear setting in both exact and inexact settings for the first time. For the NPG method, we achieved a convergence rate in which the bias is not amplified by the mismatch ratio or concentrability coefficient for the first time. We performed an empirical evaluation in which we showed the performance of these methods in an exact gradient setting and compared them to the non-projected PG and NPG methods. The results are consistent with theoretical analysis. In future work, we hope to extend the empirical evaluation of these methods to the inexact setting.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *JMLR*, 22:98:1–98:76, 2021.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *PMLR*, volume 97, pages 1042–1051, 2019.

Marc Peter Deisenroth and Carl Edward Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *ICML*, pages 465–472, 2011.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. *arXiv preprint arXiv:2302.11381*, 2023.

Sham M. Kakade. A natural policy gradient. In *NIPS*, pages 1531–1538, 2001.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *ICLR*, 2018.

Shixuan Liu, Yanghe Feng, Keyu Wu, Guangquan Cheng, Jincai Huang, and Zhong Liu. Graph-attention-based casual discovery with trust region-navigated clipping policy optimization. *IEEE Transactions on Cybernetics*, 53(4):2311–2324, 2023.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *PMLR*, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021.

Jincheng Mei, Bo Dai, Alekh Agarwal, Mohammad Ghavamzadeh, Csaba Szepesvari, and Dale Schuurmans. Ordering-based conditions for global convergence of policy gradient methods. In *NeurIPS*, 2023.

Rémi Munos. Error bounds for approximate value iteration. In *AAAI*, pages 1006–1011, 2005.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pages 1057–1063, 1999.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Mueller, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A functional mirror ascent view of policy gradient methods with function approximation. *arXiv preprint arXiv:2108.05828*, 2021.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C. Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. In *PMLR*, 2022.

Xin Wang and Thomas G. Dietterich. Model-based policy gradient reinforcement learning. In *ICML*, pages 776–783, 2003.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Lin Xiao. On the convergence rates of policy gradient methods. *HMLR*, 23:282:1–282:36, 2022.

Rui Yuan, Simon Shaolei Du, Robert M. Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *ICLR*, 2023.

A Proof of Theorem 4.4

Algorithm 3 Projected PG, Inexact Evaluation

Input: Φ, π_1, η, T.
Output: Policies $p^{\pi_t} = \text{softmax}(\pi_t)$.
for $t = 0$ to $T-1$ **do**
  Sample $a_t(s) \sim p^{\pi_t}(\cdot|s)$ for all $s \in \mathcal{S}$.
  $\widehat{Q}^{\pi_t}(s,a) \leftarrow \frac{\mathcal{I}\{a_t(s)=a\}}{p^{\pi_t}(a|s)} \cdot Q^{\pi_t}(s,a)$.
  $\widehat{g}_t(s,\cdot) \leftarrow \frac{1}{1-\gamma} \cdot d_\rho^{\pi_t}(s) \cdot \Big[\sum_a \frac{\partial p^{\pi_t}(a|s)}{\partial \pi(s,\cdot)} \cdot \widehat{Q}^{\pi_t}(s,a)\Big]$.
  Improvement: $\pi_{t+1/2} \leftarrow \pi_t + \eta \cdot \widehat{g}_t$.
  Projection: $\pi_{t+1} = \phi(\phi^\top \phi)^{-1}\phi^\top \pi_{t+1/2}$
end for

Proof. This proof is mainly based on the proof of Theorem 13 of Mei et al. [2021]. First note that for any π and ρ,

$$
\begin{aligned}
d_\rho^\pi(s) &= \mathbb{E}_{s_0\sim\rho}\left[d_\rho^\pi(s)\right] && (48)\\
&= \mathbb{E}_{s_0\sim\rho}\left[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t \text{Pr}(s_t = s|s_0,\pi,\mathcal{P})\right] && (49)\\
&\geq \mathbb{E}_{s_0\sim\rho}\left[(1-\gamma)\cdot\text{Pr}(s_0 = s|s_0)\right] && (50)\\
&= (1-\gamma)\cdot\rho(s). && (51)
\end{aligned}
$$

Next, according to [Mei et al., 2021, Lemma 20], we have,

$$
\begin{aligned}
V^*(\rho) - V^\pi(\rho) &= \frac{1}{1-\gamma}\sum_s d_\rho^\pi(s)\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) && (52)\\
&= \frac{1}{1-\gamma}\sum_s \frac{d_\rho^\pi(s)}{d_\rho^\pi(s)}\cdot d_\rho^\pi(s)\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) && (53)\\
&\leq \frac{1}{1-\gamma}\cdot\left\|\frac{d_\rho^\pi}{d_\rho^\pi}\right\|_\infty \sum_s d_\rho^\pi(s)\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) \qquad \left(\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) \geq 0\right) && (54)\\
&\leq \frac{1}{(1-\gamma)^2}\cdot\left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \sum_s d_\rho^\pi(s)\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) \qquad \Big(\text{by Equation 48 and } \min_s \rho(s) > 0\Big) && (55)\\
&\leq \frac{1}{(1-\gamma)^2}\cdot C'_\infty \cdot \sum_s d_\rho^\pi(s)\sum_a \Big(p^{\pi^*}(a|s) - p^\pi(a|s)\Big)\cdot Q^*(s,a) && (56)\\
&= \frac{1}{1-\gamma}\cdot C'_\infty \cdot \left[V^*(\rho) - V^\pi(\rho)\right]. \qquad \text{(by [Mei et al., 2021, Lemma 20])} && (57)
\end{aligned}
$$

Denote $\delta(\pi_t) \coloneqq V^*(\rho) - V^{\pi_t}(\rho)$. Let We have, for all $t \geq 1$,

$$
\begin{aligned}
&\delta(\pi_{t+1/2}) - \delta(\pi_t) && (58)\\
&= -V^{\pi_{t+1/2}}(\rho) + V^{\pi_t}(\rho) + \left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \pi_{t+1/2} - \pi_t\right\rangle - \left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \pi_{t+1/2} - \pi_t\right\rangle && (59)\\
&\leq C\cdot\left\|\frac{\partial V^{\pi_t}(\rho)}{\partial \pi}\right\|_2 \cdot \|\pi_{t+1/2} - \pi_t\|_2^2 - \left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \pi_{t+1/2} - \pi_t\right\rangle \qquad \text{(by [Mei et al., 2021, Lemma 12])} && (60)\\
&= C\cdot\eta^2\cdot\left\|\frac{\partial V^{\pi_t}(\rho)}{\partial \pi}\right\|_2 \cdot \|\widehat{g}_t\|_2^2 - \eta\cdot\left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \widehat{g}_t\right\rangle. \qquad \text{(using Eq. (22))} && (61)
\end{aligned}
$$

Since $\delta(\pi_t) - \delta(\pi_{t+1/2}) = V^{\pi_{t+1/2}}(\rho) - V^{\pi_{t+1}}(\rho) \leq b_t$, we have

$$\delta(\pi_{t+1}) - \delta(\pi_t) \leq \delta(\pi_{t+1/2}) - \delta(\pi_t) + b_t \tag{62}$$

$$\leq C \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2 \cdot \|\widehat{g}_t\|_2^2 - \eta \cdot \left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \widehat{g}_t \right\rangle + b_t \tag{63}$$

Next, taking expectation over the random sampling on Eq. (58), we have,

$$\mathbb{E}\left[\delta(\pi_{t+1})\right] - \mathbb{E}\left[\delta(\pi_t)\right] \leq C \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2 \cdot \mathbb{E}\left[\|\widehat{g}_t\|_2^2\right] - \eta \cdot \left\langle \frac{\partial V^{\pi_t}(\rho)}{\partial \pi}, \mathbb{E}\left[\widehat{g}_t\right] \right\rangle + \mathbb{E}\left[b_t\right] \tag{64}$$

$$= C \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2 \cdot \mathbb{E}\left[\|\widehat{g}_t\|_2^2\right] - \eta \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2^2 + \mathbb{E}\left[b_t\right] \qquad \text{(unbiased PG, by [Mei et al., 2021, Lemma 11])} \tag{65}$$

$$\leq \frac{2 \cdot C}{(1-\gamma)^4} \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2 - \eta \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2^2 + \mathbb{E}\left[b_t\right] \qquad \text{(bounded PG, by [Mei et al., 2021, Lemma 11])} \tag{66}$$

$$= -\frac{(1-\gamma)^4}{8 \cdot C} \cdot \left\| \frac{\partial V^{\pi_t}(\rho)}{\partial \pi} \right\|_2^3 + \mathbb{E}\left[b_t\right] \qquad \text{(by Eq. (24))} \tag{67}$$

$$\leq -\frac{(1-\gamma)^4}{8 \cdot C} \cdot \mathbb{E}\left[\min_s p^{\pi_t}(a^*(s)|s)^3\right] \cdot \mathbb{E}\left[\delta(\pi_t)^3\right] \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\rho^{\pi_t}} \right\|_\infty^{-3} \cdot \frac{1}{S \cdot \sqrt{S}} + \mathbb{E}\left[b_t\right] \qquad \text{(by [Mei et al., 2021, Lemma 9])} \tag{68}$$

$$\leq -\frac{(1-\gamma)^4}{8 \cdot C} \cdot \left(\mathbb{E}\left[\delta(\pi_t)\right]\right)^3 \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\rho^{\pi_t}} \right\|_\infty^{-3} \cdot \frac{c}{S \cdot \sqrt{S}} + \mathbb{E}\left[b_t\right], \qquad \text{(by Jensen's inequality)} \tag{69}$$

where

$$c \coloneqq \inf_{t \geq 1} \mathbb{E}\left[\min_s p^{\pi_t}(a^*(s)|s)^3\right] \tag{70}$$

$$\geq \inf_{t \geq 1} \left(\mathbb{E}\left[\min_s p^{\pi_t}(a^*(s)|s)\right]\right)^3 \qquad \text{(by Jensen's inequality)} \tag{71}$$

$$> 0, \tag{72}$$

and the last inequality is from [Mei et al., 2020, Lemma 9], since the expected iteration equals the true gradient update, which converges to global optimal policy. According to Eq. (48), we have:

$$\mathbb{E}\left[\delta(\pi_{t+1})\right] - \mathbb{E}\left[\delta(\pi_t)\right] \leq -\frac{(1-\gamma)^7}{8 \cdot C} \cdot \left(\mathbb{E}\left[\delta(\pi_t)\right]\right)^3 \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-3} \cdot \frac{c}{S \cdot \sqrt{S}} + \mathbb{E}\left[b_t\right]. \tag{73}$$

Denote $\tilde{\delta}(\pi_t) \coloneqq \mathbb{E}\left[\delta(\pi_t)\right]$ and $\tilde{b}_t \coloneqq \mathbb{E}\left[b_t\right]$. We have:

$$\frac{(1-\gamma)^7}{8 \cdot C} \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-3} \cdot \frac{c}{S \cdot \sqrt{S}} \cdot \tilde{\delta}(\pi_t)^3 \leq \tilde{\delta}(\pi_t) - \tilde{\delta}(\pi_{t+1}) + \tilde{b}_t. \tag{74}$$

Summing up for T iterations and dividing both sides by T, we have:

$$\frac{(1-\gamma)^7}{8 \cdot C} \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-3} \cdot \frac{c}{S \cdot \sqrt{S}} \cdot \min_{t \in [T-1]} \tilde{\delta}(\pi_t)^3 \leq \frac{1}{T}[\tilde{\delta}(\pi_0) - \tilde{\delta}(\pi_T)] + \frac{1}{T} \sum_{t=0}^{T-1} \tilde{b}_t \tag{75}$$

$$\leq \frac{1}{T}[\tilde{\delta}(\pi_0)] + \frac{1}{T} \sum_{t=0}^{T-1} \tilde{b}_t. \tag{76}$$

Therefore,

$$\min_{t\in[T-1]} \tilde{\delta}(\pi_t) \leq \sqrt[3]{\frac{\tilde{\delta}(\pi_0) + \sum_{t=0}^{T-1} \tilde{b}_t}{\frac{(1-\gamma)^7}{8\cdot C} \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-3} \cdot \frac{c}{S\cdot\sqrt{S}} \cdot T}} \tag{77}$$

$$\implies \min_{t\in[T-1]} \mathbb{E}\left[V^*(\rho) - V^{\pi_t}(\rho)\right] \leq \sqrt[3]{\frac{\mathbb{E}\left[V^*(\rho) - V^{\pi_0}(\rho)\right] + \sum_{t=0}^{T-1} \mathbb{E}\left[b_t\right]}{\frac{(1-\gamma)^7}{8\cdot C} \cdot \left\| \frac{d_\rho^{\pi^*}}{\rho} \right\|_\infty^{-3} \cdot \frac{c}{S\cdot\sqrt{S}} \cdot T}} \tag{78}$$

□

B Proof of Theorem 4.8

Here, we derive the convergence rate for the projected NPG with inexact policy evaluation, as stated in Theorem 4.8.

Proof. Based on Lemma 4.7, we have

$$\begin{aligned}
&\langle \widehat{Q}^{\pi_t}(s,\cdot), \pi^{*s} - \pi^s_{t+1/2}\rangle && (79)\\
=&\langle Q^{\pi_t}(s,\cdot), \pi^{*s} - \pi^s_{t+1/2}\rangle + \langle \widehat{Q}^{\pi_t}(s,\cdot) - Q^{\pi_t}(s,\cdot), \pi^{*s} - \pi^s_{t+1/2}\rangle && (80)\\
\geq&\langle Q^{\pi_t}(s,\cdot), \pi^{*s}\rangle - \langle Q^{\pi_t}(s,\cdot), \pi^s_{t+1/2}\rangle - \left\|\widehat{Q}^{\pi_t}(s,\cdot) - Q^{\pi_t}(s,\cdot)\right\|_\infty \left\|\pi^{*s} - \pi^s_{t+1/2}\right\|_1 && (81)\\
\geq&\langle Q^{\pi_t}(s,\cdot), \pi^{*s}\rangle - \langle Q^{\pi_t}(s,\cdot), \pi^s_{t+1/2}\rangle - \frac{2\gamma\tau}{1-\gamma} - 2\tau && (82)\\
\geq&\langle Q^{\pi_t}(s,\cdot), \pi^{*s}\rangle - J_s(\pi_{t+1/2}) - \frac{4\gamma\tau}{1-\gamma} && (83)\\
\geq&\langle Q^{\pi_t}(s,\cdot), \pi^{*s}\rangle - J_s(\pi_{t+1}) - b_t - \frac{4\gamma\tau}{1-\gamma} && (84)\\
\geq&\langle Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot), \pi^{*s}\rangle + J_s(\pi^*) - J_s(\pi_{t+1}) - b_t - \frac{4\gamma\tau}{1-\gamma} && (85)\\
\geq&-\left\|Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot)\right\|_\infty + J_s(\pi^*) - J_s(\pi_{t+1}) - b_t - \frac{4\gamma\tau}{1-\gamma} \qquad \text{(Hölder's inequality)} && (86)
\end{aligned}$$

Using the same approach to bound $Q^{\pi_t}(s,\cdot) - Q^*(s,\cdot)$, rearrange the above inequality, and set $c_t = \gamma^{2t+1}$, we have

$$\left\|J(\pi^*) - J(\pi_{t+1})\right\|_\infty \leq \gamma \left\|J(\pi_t) - J(\pi^*)\right\|_\infty + \gamma^{2t+1} + \frac{4\gamma\tau}{1-\gamma} + b_t \tag{87}$$

Unravelling the recursion give us

$$\left\|J(\pi^*) - J(\pi_T)\right\|_\infty \leq \gamma^T \left(\left\|J(\pi^*) - J(\pi_0)\right\|_\infty + \sum_{t=1}^{T} \gamma^{-t}\gamma^{2(t-1)+1}\right) + \frac{4\gamma\tau}{1-\gamma}\sum_{t=0}^{T-1}\gamma^t + \sum_{t=0}^{T-1}\gamma^t b_t \tag{88}$$

$$\leq \gamma^T \left(\left\|J(\pi^*) - J(\pi_0)\right\|_\infty + \frac{1}{1-\gamma}\right) + \frac{4\gamma\tau}{(1-\gamma)^2} + \sum_{t=0}^{T-1}\gamma^t b_t, \tag{89}$$

which concludes the proof. □

C Empirical Evaluation

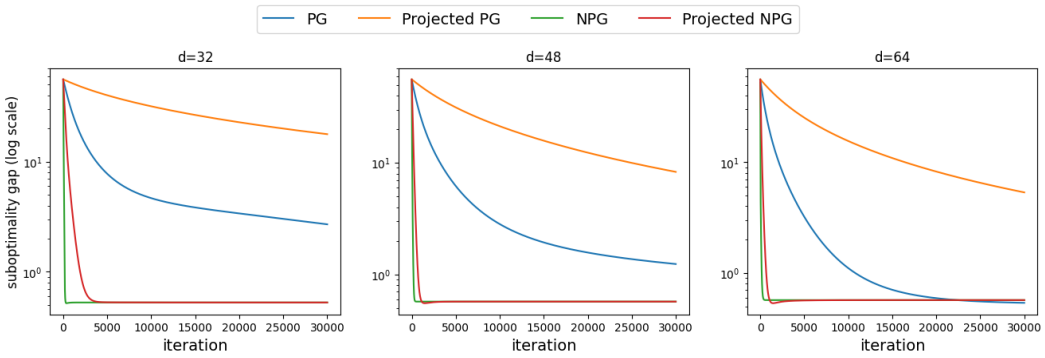

Figure 1: Comparison of different policy gradient methods for 3 different number of features d. The sub-optimality gap is computed as $V^*(\rho) - V^{\pi_t}(\rho)$ and presented in log scale.

Here, we present some preliminary results to demonstrate the convergence rates of different methods. We run experiments in the exact update setting on the Cliff World environment of Sutton and Barto [1998]. This problem has 21 states and 4 actions. We generate the features randomly by sampling from a uniform distribution such that $\Phi \sim \mathcal{U}(0, 1)$. We compare our proposed projected methods against the standard PG and NPG methods. All methods share the same features under each experiment. We run each algorithm for 30,000 iterations with $\gamma = 0.9$ and $\eta_t = 1 \times 10^{-5}$.

Figure 1 shows the results of our experiments. The PG method outperforms the projected PG method in both decreasing the sub-optimality gap and having a faster convergence rate. We can see the difference in converge rates between the PG and NPG methods. As we increase the number of features, the difference in performance between these two methods is larger and with $d = 64$, the standard PG method reaches the sub-optimality gap achieved by the NPG methods. Across all experiments, the projected NPG method matches the standard NPG method closely but slightly outperforms it.

D Computational Complexity Analysis

Because the Moore–Penrose inversion of Φ can be computed just once outside of the scope of the iterative process, the complexities of our proposed methods are very similar to their counterparts. This inversion is done in $\mathcal{O}(d^3 + d|\mathcal{S}||\mathcal{A}|)$ steps. Multiplying $\pi_{t+1/2}$ by this inverse replaces the multiplication of the tabular gradient by Φ in the PG and NPG methods. Both operations take $\mathcal{O}(d|\mathcal{S}||\mathcal{A}|)$ steps. Thus, the projection step has a complexity of $\mathcal{O}(d^3 + Td|\mathcal{S}||\mathcal{A}|)$.

Assuming that the MDP is known, computing the advantage function A^{π_t} and the state occupancy measure d^{π_t} involves $\mathcal{O}(|\mathcal{S}|^3 + |\mathcal{S}|^2|\mathcal{A}|)$ steps. Each metric requires the computation of the transition probabilities $P_{\pi_t}[s'|s]$ under the policy π_t in $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$ steps, and the inversion of the Neumann series matrix, which takes $\mathcal{O}(|\mathcal{S}|^3)$ plus a constant amount of matrix-vector multiplications subsumed by $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$. Computing d^{π_t} has the same complexity as A^{π_t}. Thus, the PG and NPG methods have the same complexity, up to a constant.

These metrics are needed for the functional step, which takes an additional amount of $\mathcal{O}(d|\mathcal{S}||\mathcal{A}|)$ steps in the form of matrix-vector multiplications needed to compute the logits, policy, and the rest of the update. This complexity is the same as what is needed to compute tabular gradient (before including Φ) in the PG and NPG methods. Putting everything together, we have a complexity of:

$$\mathcal{O}\Big[d^3 + T\big(|\mathcal{S}|^3 + |\mathcal{S}|^2|\mathcal{A}| + d|\mathcal{S}||\mathcal{A}|\big)\Big], \tag{90}$$

for both the proposed projected PG and NPG methods, which compared to their counterparts:

$$\mathcal{O}\Big[T\big(|\mathcal{S}|^3 + |\mathcal{S}|^2|\mathcal{A}| + d|\mathcal{S}||\mathcal{A}|\big)\Big], \tag{91}$$

is only different by a d^3 term. The terms $|\mathcal{S}|^3 + |\mathcal{S}|^2|\mathcal{A}|$ can be replaced depending on the method used to estimate the advantage function and the state occupancy measure.

E Analysis of policy gradients for the bandit setting

We start an analysis of the policy gradient method in the bandit setting to potentially arrive at a convergence rate. In similar work, the results obtained for the bandit setting are easily extendable to MDPs.

Let $\Phi \in \mathbb{R}^{d\times|\mathcal{A}|}$ be a feature matrix for each action. Throughout this section, we use the standard notation of $\pi_\theta = \text{softmax}(z)$, where $z = \Phi^\top\theta$ are the logits. In the bandit case, the objective function reduces to:

$$\max_\theta \mathbb{E}_{a\sim\pi_\theta}\big[r(a)\big] = \max_\theta \pi_\theta^\top\, r, \tag{92}$$

with:

$$\frac{\partial \pi_\theta^\top r}{\partial\theta} = \Phi\big[\text{diag}(\pi_\theta) - \pi_\theta\,\pi_\theta^\top\big]\, r = \Phi\big[\pi_\theta \circ \big(r - \pi_\theta^\top r\mathbf{1}\big)\big]. \tag{93}$$

E.1 Smoothness

We show that the objective function $\theta \to \pi_\theta^\top r$ is $5/2\|\Phi\|_2^2$ smooth. Under different assumptions on the reward function, a similar result has been shown in Mei et al. [2023] but the L-smooth coefficient and the proof are different.

Let $\Phi_i \in \mathbb{R}^{|\mathcal{A}|}$ be the feature position i across all actions and let $S \triangleq S(r,\theta) \in \mathbb{R}^{d\times d}$ be the Hessian of the objective function, such that:

$$\begin{aligned}
S_{i,j} &= \Phi_i^\top \frac{\partial[\pi_\theta \circ (r - \pi_\theta^\top r\mathbf{1})]}{\partial\theta_j} && (94)\\
&= \Phi_i^\top\left[\frac{\partial\pi_\theta}{\partial\theta_j} \circ (r - \pi_\theta^\top r\mathbf{1}) + \pi_\theta \circ \frac{\partial[r - \pi_\theta^\top r\mathbf{1}]}{\partial\theta_j}\right] && (95)\\
&= \Phi_i^\top\Big[H(\pi_\theta)\,\Phi_j \circ (r - \pi_\theta^\top r\mathbf{1}) - (H(\pi_\theta)\,\Phi_j)^\top r\,\pi_\theta\Big] && (96)\\
&= \Phi_i^\top\Big[H(\pi_\theta)\,\Phi_j \circ (r - \pi_\theta^\top r\mathbf{1}) - \Phi_j^\top H(\pi_\theta)\,r\,\pi_\theta\Big] && (H^\top = H) \quad (97)\\
&= \Phi_i^\top\Big[H(\pi_\theta)\,\Phi_j \circ (r - \pi_\theta^\top r\mathbf{1}) - r^\top H(\pi_\theta)\,\Phi_j\,\pi_\theta\Big] && (u^\top Hv = v^\top Hu) \quad (98)\\
&= \Phi_i^\top\Big[\text{diag}(r)\,H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta)\Big]\Phi_j. && (x^\top yz = zx^\top y) \quad (99)
\end{aligned}$$

Any operator norm is consistent with the vector norms that induce it. Thus, we have that:

$$\|Ax\|_2 \le \|A\|_2\|x\|_2. \tag{100}$$

Moreover, we have the following norm inequalities:

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1. \tag{101}$$

Assuming that $r \in [0,1]^{|\mathcal{A}|}$, we note that $\|\pi_\theta^\top r\| \le 1$ and that $\|r^\top H(\pi_\theta)\|_1 \le 1$ as shown in Mei et al. [2020]. Also note that:

$$\|H(\pi_\theta)\|_1 = \max_{1\le j\le m}\|H_{i,\cdot}(\pi_\theta)\|_1 \le 1/2, \tag{102}$$

where $\|H_{i,:}(\pi_\theta)\|_1 \leq 1/2$ is shown in Mei et al. [2020]. We upper bound the spectral norm of the Hessian as follows:

$$
\begin{aligned}
&|y^\top S y| && (103)\\
&= \left| \sum_{i=1}^{m} \sum_{j=1}^{m} y_i \, S_{i,j} \, y_j \right| && (104)\\
&= \left| \sum_{i=1}^{m} \sum_{j=1}^{m} y_i \, \Phi_i^\top \Big[\mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \Big] \Phi_j \, y_j \right| && (105)\\
&= \left| y^\top \Phi^\top \Big[\mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \Big] \Phi \, y \right| && (106)\\
&\leq \left\| y^\top \Phi^\top \right\|_\infty \left\| \left[\mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \right] \Phi \, y \right\|_1 && \text{(Hölder's ineq.)} \quad (107)\\
&\leq \left\| y^\top \Phi^\top \right\|_\infty \left\| \mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \right\|_2 \left\| \Phi \, y \right\|_2 && \text{(Eq. 100)} \quad (108)\\
&\leq \left\| y^\top \Phi^\top \right\|_2 \left\| \mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \right\|_1 \left\| \Phi \, y \right\|_2 && \text{(Eq. 101)} \quad (109)\\
&\leq \|\Phi\|_2^2 \, \|y\|_2^2 \left\| \mathrm{diag}(r)\, H(\pi_\theta) - \pi_\theta^\top r H(\pi_\theta) - \pi_\theta r^\top H(\pi_\theta) \right\|_1 && \text{(Cauchy ineq.)} \quad (110)\\
&\leq \|\Phi\|_2^2 \, \|y\|_2^2 \Big[\|\mathrm{diag}(r)\, H(\pi_\theta)\|_1 + \|\pi_\theta^\top r H(\pi_\theta)\|_1 + \|\pi_\theta r^\top H(\pi_\theta)\|_1 \Big] && \text{(triangle ineq.)} \quad (111)\\
&\leq \|\Phi\|_2^2 \, \|y\|_2^2 \Big[\|r^\top H(\pi_\theta)\|_1 + |\pi_\theta^\top r| \, \|H(\pi_\theta)\|_1 + \|\pi_\theta\|_\infty \|r^\top H(\pi_\theta)\|_1 \Big] && (112)\\
&\leq \|\Phi\|_2^2 \, \|y\|_2^2 \, [1 + 1/2 + 1] && \text{(Eq. 102)} \quad (113)\\
&= 5/2 \, \|\Phi\|_2^2 \, \|y\|_2^2. && (114)
\end{aligned}
$$

Thus, using Taylor's theorem, we have that:

$$
\left| \left(\pi_{\theta_{t+1}} - \pi_{\theta_t} \right)^\top r - \left\langle \frac{\partial \pi_{\theta_t}^\top r}{\partial \theta_t} \middle| \theta_{t+1} - \theta_t \right\rangle \right| \leq \frac{5}{4} \|\Phi\|_2^2 \|\theta_{t+1} - \theta_t\|_2^2. \tag{115}
$$

E.2 Gradient dominance

We show that the gradient norm is non-zero for all non-optimal feasible policies that are non-deterministic. Let $w(\pi) = \pi \odot (r - \pi^\top r)$ such that:

$$
\frac{\partial \pi_\theta^\top r}{\partial \theta} = \Phi \big[\pi_\theta \odot (r - \pi_\theta^\top r) \big] = \Phi w(\pi_\theta). \tag{116}
$$

Deterministic policies have $w(\pi)_i = 0 \, \forall i \in \{1, ..., |\mathcal{A}|\}$. Let $\pi^\dagger$ be a deterministic policy such that $\pi_i^\dagger = 1, \pi_{j \neq i}^\dagger = 0$. We have that:

$$
\begin{aligned}
&\pi_i^\dagger = 1, \pi_{j \neq i}^\dagger = 0 \implies w(\pi)_i = \pi_i^\dagger r_i - \pi_i^\dagger \pi^{\dagger\top} r = \pi_i^\dagger r_i - r_i = r_i - r_i = 0 && (117)\\
&\pi_i^\dagger = 1, \pi_{j \neq i}^\dagger = 0 \implies w(\pi)_j = \pi_j^\dagger r_j - \pi_j^\dagger \pi^{\dagger\top} r = 0 - 0 = 0. && (118)
\end{aligned}
$$

As a result $\Phi \pi^\dagger = 0$. Thus, for gradient dominance and the PL condition to hold, we need to exclude non-optimal deterministic policies from the feasible set. We assume that $\mathbf{\Pi}$ is formed such that $\pi(a^*) > 0 \, \forall \pi \in \mathbf{\Pi}; a^* = \arg\max_a \pi^*(a) = 1$, where $\pi^* = \arg\max_{\pi \in \Delta(|\mathcal{A}|)} \pi^\top r$ is *any* optimal policy.

If any of the optimal feasible policies $\widehat{\pi} = \arg\max_{\pi \in \mathbf{\Pi}} \pi^\top r$ are non-deterministic, then we must have that:

$$
\Phi w(\widehat{\pi}) = 0. \tag{119}
$$

If $\widehat{\pi}$ is non-deterministic and non-optimal in the sense that $\widehat{\pi}^\top r < \pi^{*\top} r$, then $\exists i : w(\widehat{\pi})_i \neq 0$. We show this as follows. The optimal feasible policies have $\widehat{\pi}(a^*) \geq 1/|\mathcal{A}|$, with equality when $\theta = 0$ produces logits $\Phi\theta = 0$, and since $\exp(0) = 1$, we have after the softmax that $\widehat{\pi} = 1/|\mathcal{A}|$. For a policy

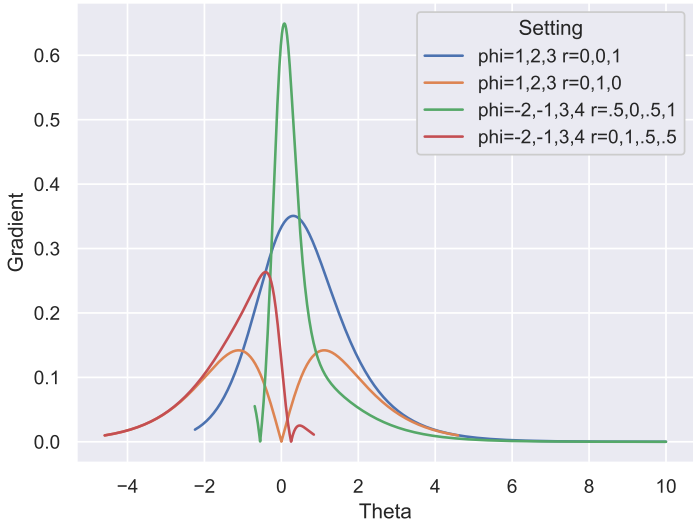

Figure 2: Absolute gradient across different parameter values θ, for a single feature vector ϕ, under different settings. Policies with $\pi(a^*) > 0.01$ are removed from the plot to highlight the optimal solution from a feasible set that does not consider such policies. The blue line corresponds to a a setting in which the optimal solution is deterministic. The orange line corresponds to a setting in which the optimal solution is a uniform distribution across all actions with $\theta = 0$. The green and red lines correspond to non-deterministic policies.

to be non-deterministic and non-optimal with respect to $\Delta(|\mathcal{A}|)$ as described, it must have probability $\delta \in (0, 1)$ on non-optimal actions with reward $r(a^*) - \epsilon$; $\epsilon \in (0, r(a^*)]$. Thus, we have that:

$$\widehat{\pi}^\top r < \pi^{*\top} r \implies \exists i : \widehat{\pi}_i \big[r_i - \widehat{\pi}^\top r \big] \tag{120}$$

$$= \delta \big[(r(a^*) - \epsilon) - \big[\delta \, (r(a^*) - \epsilon) + (1 - \delta) \, r(a^*) \big] \big] \tag{121}$$

$$= \delta \big[-\epsilon + \delta \, \epsilon \big] < 0. \tag{122}$$

Hence, since a non-deterministic optimal and feasible policy has a non-zero $w(\widehat{\pi})$ and because its gradient must be zero, then we must have that:

$$\widehat{\pi}^\top r < \pi^{*\top} r \implies w(\widehat{\pi}) \in \ker \Phi, \tag{123}$$

where $\ker \Phi$ is the nullspace or kernel of Φ.

We now show that non-optimal policies are not in the nullspace of Φ. First we show that $\sum_i w(\pi)_i = 0$ for any policy as follows:

$$\sum_i w(\pi)_i = \sum_i \pi_i \left(r_i - \sum_j \pi_j r_j \right) = \sum_i \pi_i \, r_i - \sum_j \pi_j r_j \sum_i \pi_i \tag{124}$$

$$= \sum_i \pi_i \, r_i - \sum_j \pi_j r_j = 0. \tag{125}$$

There can be an infinite amount of vectors w in the nullspace of Φ. However, once we consider the constraint $\sum_i w(\pi)_i = 0$, only scaled versions of an optimal feasible policy vector $\widehat{w} = w(\widehat{\pi}) = \widehat{\pi} \odot (r - \widehat{\pi}^\top r)$ are in the nullspace. We prove this as follows. Without loss of generality, we focus on a single row $\phi \in \Phi$. The constraints $\phi^\top w = 0$ and $\sum_i w(\pi)_i = 0$ defines the following system:

$$\phi_1 w_1(\pi) + \cdots + \phi_{|\mathcal{A}|} w_{|\mathcal{A}|}(\pi) = 0 \tag{126}$$

$$w_1(\pi) + \cdots + w_{|\mathcal{A}|}(\pi) = 0. \tag{127}$$

This is an underdetermined system for $|\mathcal{A}| > 2$. When the system is consistent, it has an infinitude of solutions. Subtracting both equations and solving for one variable, we obtain:

$$(\phi_1 - 1) w_1(\pi) + \cdots + (\phi_{|\mathcal{A}|} - 1) w_{|\mathcal{A}|}(\pi) = 0 \tag{128}$$

$$\implies w_i(\pi) = \frac{1}{(1 - \phi_i)} \sum_{i \neq j} (\phi_j - 1) \, w_j(\pi). \tag{129}$$

Solving for any $w_i(\pi)\,\forall i \in \{1, ..., |\mathcal{A}|\}$ results in a linear equation with no constant term so that it only depends on the other variables. Given a solution $\widehat{w}$, when scaling a variable $\widehat{w}_i$ by λ, the other variables must also scale by λ to satisfy the constraints. Thus, the system has infinite solutions corresponding to $\lambda\widehat{w}\;\forall\lambda \in \mathbb{R}$.

We show now that $w(\pi) \neq \lambda\widehat{w}\,\forall\pi \in \mathbf{\Pi} : \pi \neq \widehat{\pi}$. Thus, optimal feasible policies are the only policies with a corresponding $w(\widehat{\pi}) \in \ker\Phi$ and therefore only their gradients can be zero. We split the analysis into two cases, one with optimal feasible policies with non-zero probability in more than 2 actions such that $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} > 2$ and the other with the converse, such that $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} = 2$.

For $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} > 2$, note that:

$$\widehat{w} = \lambda w(\pi) \implies \widehat{\pi} \odot (r - \widehat{\pi}^\top r) = \lambda\pi \odot (r - \pi^\top r) \tag{130}$$

$$\implies (\mathrm{diag}(\widehat{\pi}) - \widehat{\pi}\widehat{\pi}^\top)\, r = \lambda(\mathrm{diag}(\pi) - \pi\pi^\top)\, r \tag{131}$$

$$\implies H(\widehat{\pi}) = \lambda H(\pi). \tag{132}$$

We focus on the diagonal where $H_{i,i}(\pi) = \pi_i - \pi_i^2$. This is a quadratic concave function with a maximum of $1/4$ at $\pi_i = 1/2$. The domain of this function is $[0, 1]$ and the image is $[0, 1/4]$. The derivative is given as:

$$\frac{dH_{i,i}(\pi)}{d\pi_i} = 1 - 2\pi_i. \tag{133}$$

Note that the derivative for any point π_i has different magnitude, except at point $1 - \pi_i$. As we scale $H_{i,i}(\pi)$ by $\lambda \neq 0$, all other points $H_{j,j}(\pi) : i \neq j$ must scale at the same rate, to satisfy:

$$\frac{H_{i,i}(\widehat{\pi})}{H_{i,i}(\pi)} = \frac{H_{j,j}(\widehat{\pi})}{H_{j,j}(\pi)} \implies \frac{\widehat{\pi}_i(1-\widehat{\pi}_i)}{\pi_i(1-\pi_i)} = \frac{\widehat{\pi}_j(1-\widehat{\pi}_j)}{\pi_j(1-\pi_j)} \implies \frac{\widehat{\pi}_i(1-\widehat{\pi}_i)}{\widehat{\pi}_j(1-\widehat{\pi}_j)} = \frac{\pi_i(1-\pi_i)}{\pi_j(1-\pi_j)}. \tag{134}$$

If we set $\widehat{\pi}_j = (1 - \widehat{\pi}_i)$ we can find another policy π that satisfies this constraint. However this will mean that only two actions have all the probability mass. Under the assumption that $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} > 2$, we cannot find another policy that satisfies Eq. (134). Because the derivative magnitude is only the same at points π_i and $1 - \pi_i$, the underlying probabilities will change at a different rate when scaling $H(\pi)$, breaking the constraint that they must sum to 1. Hence, there are no other policies that can produce $\lambda\widehat{w}$ when $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} > 2$.

For policies in which $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} = 2$, there are an infinite amount of policies such that $H(\widehat{\pi}) = \lambda H(\pi)$ for $\lambda \in \mathbf{\Lambda} \subset \mathbb{R}$. However, from this set of policies, we have that all the feasible ones are optimal. We divide the analysis in two cases based on whether the rewards for the two actions with non-zero probabilities are the same or different. When the rewards are the same they also are the highest rewards such that $r(a_1) = r(a_2) = r(a*)$, because policies with zero probability on the best action are not in the feasible set as per our constraint. Thus, we have that $w(\pi) = 0\,\forall\pi \in \Delta(|\mathcal{A}|) : \pi_1 = (1 - \pi_2)$:

$$\pi_2 = (1 - \pi_1), r_1 = r_2 \implies \pi^\top r = \pi_1 r_1 + \pi_2 r_2 = \pi_1 r_1 + (1 - \pi_1)\, r_2 = r_2 = r_1 \tag{135}$$

$$\implies w(\pi)_1 = \pi_1(r_1 - \pi^\top r) = \pi_1(r_1 - r_1) = 0 \tag{136}$$

$$\implies w(\pi)_2 = \pi_2(r_2 - \pi^\top r) = \pi_2(r_2 - r_2) = 0 \tag{137}$$

$$\implies w(\pi)_{j\notin\{1,2\}} = \pi_j(r_j - \pi^\top r) = 0 \cdot (r_j - \pi^\top r) = 0. \tag{138}$$

All these policies are optimal and produce the zero vector for $w(\pi)$, which is in the nullspace of Φ. If the rewards are not the same then if the optimal feasible solution $\widehat{\pi}$ in non-deterministic, we have that $w(\widehat{\pi})_1 = -w(\widehat{\pi})_2, w(\widehat{\pi})_{j\notin\{1,2\}} = 0$:

$$\pi_2 = (1 - \pi_1), \pi_1, \pi_2 > 0, r_1 \neq r_2 \tag{139}$$

$$\implies \pi^\top r = \pi_1 r_1 + \pi_2 r_2 = \pi_1(r_1 - r_2) + r_2 \tag{140}$$

$$\implies w(\pi)_1 = \pi_1(r_1 - \pi_1(r_1 - r_2) - r_2) = \pi_1(1 - \pi_1)(r_1 - r_2) \neq 0 \qquad (\pi_1, \pi_2 > 0) \tag{141}$$

$$\implies w(\pi)_{j\notin\{1,2\}} = \pi_j(r_j - \pi^\top r) = 0 \cdot (r_j - \pi^\top r) = 0 \tag{142}$$

$$\implies w(\pi)_2 = -\pi_1(r_1 - \pi_1(r_1 - r_2) - r_2) \neq 0. \qquad \left(\textstyle\sum_i w_i = 0\right) \tag{143}$$

Since $w(\pi) \neq \mathbf{0}$, for it to be in the nullspace of Φ, we must have that $\phi_1 = \phi_2\,\forall\phi \in \Phi$. If the features are the same, then the only feasible policy that can be produced is one in which $\pi_1 = \pi_2$, and it

would be the optimal feasible policy $\widehat{\pi}$. Otherwise the best feasible policy would be deterministic and it would produce $w(\widehat{\pi}) = \mathbf{0}$. Thus, when $\sum_i \mathcal{I}\{\widehat{\pi}_i \neq 0\} = 2$, all feasible policies are optimal. This concludes our proof, showing that:

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\| > 0 \; \forall \pi_\theta \in \Pi : \pi_\theta(a^*) > 0 \wedge \pi_\theta \neq \widehat{\pi}_\theta. \tag{144}$$

Figure 2 shows a plot of the gradient under different settings to demonstrate the gradient dominance.

E.3 Non-uniform Łojasiewicz condition

We start an analysis that can potentially arrive at the PL condition:

$$\begin{aligned}
\left\| \frac{\partial \pi_\theta^\top r}{\partial \theta} \right\|_2^2 &= \left\| \Phi \left[\pi_\theta \odot \left(r - \pi_\theta^\top r \right) \right] \right\|_2^2 && (145)\\
&\geq \left| \phi^\top \left[\pi_\theta \odot \left(r - \pi_\theta^\top r \right) \right] \right| \qquad (\phi \in \Phi) && (146)\\
&= \left| \phi^\top \left[\pi_\theta \odot \left(r - \pi_\theta^\top r + \widehat{\pi}_\theta^\top r - \widehat{\pi}_\theta^\top r \right) \right] \right| && (147)\\
&= \left| \phi^\top \left[\pi_\theta \odot \left(\widehat{\pi}_\theta^\top r - \pi_\theta^\top r \right) + \pi_\theta \odot \left(r - \widehat{\pi}_\theta^\top r \right) \right] \right| && (148)\\
&= \left| \phi^\top \pi_\theta \left(\widehat{\pi}_\theta - \pi_\theta \right)^\top r + \phi^\top \left[\pi_\theta \odot \left(r - \widehat{\pi}_\theta^\top r \right) \right] \right|. && (149)
\end{aligned}$$

We notice that the first term in Eq. (149) goes to zero as $\pi_\theta \to \widehat{\pi}_\theta$. The scaling factor is $\phi^\top \pi_\theta$. For the second term, when $\pi_\theta = \widehat{\pi}_\theta$, we recover the expression that we analyzed in the previous gradient dominance section, where we have shown that the resulting $w(\widehat{\pi}_\theta)$ is in the nullspace of ϕ and the entire term is zero.