Journal of Machine Learning Research 26 (2025) 1-35 Submitted 9/24; Revised 5/25; Published 7/25

On the Convergence of Projected Policy Gradient for Any Constant Step Sizes

Jiacai Liu 23110980012@m.fudan.edu.cn
School of Data Science
Fudan University
Shanghai, China

Wenye Li 23210980111@m.fudan.edu.cn
School of Data Science
Fudan University
Shanghai, China

Dachao Lin lindachao@pku.edu.cn
Huawei Technologies Shanghai R&D Center
Shanghai, China

Ke Wei kewei@fudan.edu.cn
School of Data Science
Fudan University
Shanghai, China

Zhihua Zhang zhzhang@math.pku.edu.cn
School of Mathematical Sciences
Peking University
Beijing, China

Editor: Alekh Agarwal

Abstract

Projected policy gradient (PPG) is a basic policy optimization method in reinforcement learning. Given access to exact policy evaluations, previous studies have established the sublinear convergence of PPG for sufficiently small step sizes based on the smoothness and the gradient domination properties of the value function. However, as the step size goes to infinity, PPG reduces to the classic policy iteration method, which suggests the convergence of PPG even for large step sizes. In this paper, we fill this gap and show that PPG admits a sublinear convergence *for any constant step sizes*. Due to the existence of the state-wise visitation measure in the expression of policy gradient, the existing optimization-based analysis framework for a preconditioned version of PPG (i.e., projected Q-ascent) is not applicable, to the best of our knowledge. Instead, we proceed the proof by computing the state-wise improvement lower bound of PPG based on its inherent structure. In addition, the finite iteration convergence of PPG for any constant step size is further established, which is also new.

Keywords: projected policy gradient, sublinear convergence, finite iteration convergence, policy optimization, policy iteration

1. Introduction

Reinforcement learning (RL) is essentially about how to make efficient sequential decisions to achieve a long term goal. It has received intensive investigations both from theoretical and algorithmic aspects due to its recent success in many areas, such as games (Mnih et al., 2015; Silver et al., 2016; Berner et al., 2019; Vinyals et al., 2017), robotics (Hwangbo et al., 2019; Lee et al., 2020; Miki et al., 2022) and various other real applications (Agarwal et al., 2016; Chen et al., 2019; Mirhoseini et al., 2021). Typically, RL can be modeled as a discounted Markov decision process (MDP) represented by a tuple $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ denotes the action space, $P(s'|s,a)$ is the transition probability or density from state s to state s' under action a, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, $\gamma \in [0,1)$ is the discounted factor and μ is the probability distribution of the initial state s_0. In this paper, we focus the tabular setting where $\mathcal{S}$ and $\mathcal{A}$ are finite, i.e., $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. Let $\Delta(\mathcal{A})$ be the probability simplex over the set $\mathcal{A}$, defined as

$$\Delta(\mathcal{A}) = \left\{ \theta \in \mathbb{R}^{|\mathcal{A}|} : \theta_i \geq 0, \sum_{i=1}^{|\mathcal{A}|} \theta_i = 1 \right\}. \tag{1}$$

The set of admissible stationary policies (i.e., the direct or simplex parameterization of policies) is given by

$$\Pi := \left\{ \pi = (\pi_s)_{s\in\mathcal{S}} \mid \pi_s \in \Delta(\mathcal{A}) \text{ for all } s \in \mathcal{S} \right\}, \tag{2}$$

where $\pi_s := \pi(\cdot|s) \in \mathbb{R}^{|\mathcal{A}|}$ and $\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Given a policy $\pi \in \Pi$, the state value function at $s \in \mathcal{S}$ is defined as

$$V^\pi(s) := \mathbb{E}\left\{ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, \pi \right\}, \tag{3}$$

while the state-action value function at $(s,a) \in \mathcal{S} \times \mathcal{A}$ are defined as

$$Q^\pi(s,a) := \mathbb{E}\left\{ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a, \pi \right\}. \tag{4}$$

Overall, the goal of RL is to find a policy that maximizes the weighted average of the state values under the initial distribution μ, namely to solve

$$\max_{\pi\in\Pi} V^\pi(\mu). \tag{5}$$

Here $V^\pi(\rho) := \mathbb{E}_{s\sim\rho}[V^\pi(s)]$ for any $\rho \in \Delta(\mathcal{S})$.

Policy optimization refers to a family of effective methods in reinforcement learning. In this paper, we focus on projected policy gradient (PPG) which is likely to be the most direct optimization method for solving (5). Given an initial policy $\pi_0 \in \Pi$, PPG generates a policy sequence $\{\pi^k\}$ for $k = 1, 2, 3, ...$ as follows:

$$\pi^{k+1} = \arg\max_{\pi\in\Pi} \left\{ \eta_k \left\langle \nabla_\pi V^\pi(\mu)|_{\pi=\pi^k}, \pi - \pi^k \right\rangle - \frac{1}{2} \left\| \pi - \pi^k \right\|_2^2 \right\},$$

$$= \arg\max_{\pi\in\Pi} \left\{ \sum_{s\in\mathcal{S}} \left(\eta_k \left\langle \nabla_{\pi_s} V^{\pi}(\mu)\,|_{\pi=\pi^k}\,, \pi_s - \pi_s^k \right\rangle - \frac{1}{2} \left\| \pi_s - \pi_s^k \right\|_2^2 \right) \right\},$$

or state-wisely,

$$\pi_s^{k+1} = \arg\max_{\pi\in\Pi} \left\{ \eta_k \left\langle \nabla_{\pi_s} V^{\pi}(\mu)\,|_{\pi=\pi^k}\,, \pi_s - \pi_s^k \right\rangle - \frac{1}{2} \left\| \pi_s - \pi_s^k \right\|_2^2 \right\}. \tag{6}$$

According to the policy gradient theorem (Sutton et al., 1999),

$$\nabla_{\pi_s} V^{\pi}(\mu) = \frac{d_\mu^{\pi}(s)}{1-\gamma} Q^{\pi}(s,\cdot)$$

where d_μ^π is the state visitation probability defined as

$$d_\mu^{\pi}(s) := (1-\gamma)\,\mathbb{E}\left\{ \sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{[s_t=s]} | s_0 \sim \mu, \pi \right\}. \tag{7}$$

Thus, PPG can be written explicitly in the following form:

$$\text{(PPG)} \quad \pi_s^{k+1} = \mathrm{Proj}_{\Delta(\mathcal{A})} \left(\pi_s^k + \frac{\eta_k d_\mu^k(s)}{1-\gamma} Q^k(s,\cdot) \right), \quad \forall s \in \mathcal{S}, \tag{8}$$

where d_μ^k and $Q^k(s,\cdot)$ are short for $d_\mu^{\pi_k}$ and $Q^{\pi_k}(s,\cdot)$, respectively, and $\mathrm{Proj}_{\Delta(\mathcal{A})}$ denotes the projection onto $\Delta(\mathcal{A})$, i.e., $\mathrm{Proj}_{\Delta(\mathcal{A})}(v) = \arg\min_{p\in\Delta(\mathcal{A})} \|p - v\|_2^2$. Note that removing the visitation measure $d_\mu^k(s)$ in the PPG update leads to the projected Q-ascent (PQA) method,

$$\text{(PQA)} \quad \pi_s^{k+1} = \mathrm{Proj}_{\Delta(\mathcal{A})} \left(\pi_s^k + \eta_k Q^k(s,\cdot) \right), \quad \forall s \in \mathcal{S}. \tag{9}$$

PQA is indeed a special case of policy mirror ascent methods (e.g. Geist et al., 2019; Shani et al., 2020; Lan, 2021; Xiao, 2022; Cen et al., 2022; Zhan et al., 2023; Li et al., 2023; Johnson et al., 2023) where the Bregman distance is the squared ℓ_2-distance and it can also be seen as a preconditioned version of PPG.

1.1 Motivation and contributions

The convergence of PPG has been investigated given the access to exact policy evaluations (Agarwal et al., 2021; Bhandari and Russo, 2024; Zhang et al., 2020; Xiao, 2022). More precisely, it is shown that PPG converges to a global optimum at an $O(1/\sqrt{k})$ sublinear rate (Agarwal et al., 2021; Bhandari and Russo, 2024), which has been subsequently improved to $O(1/k)$ (Zhang et al., 2020; Xiao, 2022). The analyses in these works all utilize the smoothness property of the value function, and thus require the step size to be smaller than $1/L$, where $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$ is the smoothness coefficient of the value function (Agarwal et al., 2021). However, as η_k goes to infinity, it is easy to see from (6) that PPG approaches the classic policy iteration (PI) method. Therefore, due to the convergence of policy iteration, it is natural to expect PPG also converges for large step sizes.

Motivated by the above observation, we extend the convergence studies of PPG to any constant step sizes in this paper. The main contributions of this paper are summarized as follows:

- The $O(1/k)$ sublinear convergence of PPG has been established for any constant step sizes, see Theorem 15. In order to break the step size limitation hidden in the existing optimization analysis framework, we adopt a different route and leverage the more explicit form of the projection onto the probability simplex to derive a state-wise improvement lower bound for PPG. *It is worth noting that, due to the existence of the visitation measure in PPG, the analysis for PQA within the framework of policy mirror ascent* (Xiao, 2022; Lan, 2021) *is not applicable for PPG, to the best of our knowledge. In fact, the sublinear convergence results of PPG* (*only for sufficiently small step sizes*) *and PQA* (*for any constant step sizes*) *have been established separately based on different techniques in* Xiao (2022).

- We further show that PPG indeed terminates after a finite number of iterations. The finite iteration convergence of PQA for any constant step size can also be obtained in a similar way. Note that, as a special case of a general result, the homotopic PQA can be shown to converge in a finite number of iterations (Li et al., 2023). However, this does not imply the finite convergence of PQA for any constant step sizes and the homotopic PQA basically requires an exponentially increasing step size to converge. As a by-product, we present a new dimension-free bound for the finite iteration convergence of PI and VI, which does not explicitly depend on $|\mathcal{S}|$ and $|\mathcal{A}|$.

In addition to the main contributions, we also give a brief discussion on the γ-rate linear convergence of PPG using non-adaptive geometrically increasing step sizes, as well as the equivalence of PPG and PQA to policy iteration when the step size η_k is larger than a threshold that can be calculated from the current policy π^k. The existing convergence results and our new results on PPG (as well as on PQA for completeness) are summarized in Table 1.1.

Table 1: Convergence results for PPG and PQA.

	PPG	PQA
Existing results	Sublinear convergence when $\eta_k \leq 1/L$ (Agarwal et al., 2021; Bhandari and Russo, 2024; Zhang et al., 2020; Xiao, 2022)	1) Sublinear linear convergence for any constant η_k (Lan, 2021; Xiao, 2022); 2) Finite iteration convergence for homotopic PQA (Li et al., 2023); 3) γ-rate/linear convergence for geometrically increasing step sizes (Xiao, 2022; Johnson et al., 2023)
This paper	1) Sublinear convergence for any constant η_k; 2) Finite iteration convergence for any constant η_k; 3) γ-rate linear convergence for geometrically increasing step sizes	Finite iteration convergence for any constant η_k

1.2 Notation and assumptions

Recalling the definitions of the state value function (3) and the state-action value function (4), the advantage function of a policy π is defined as

$$A^{\pi}(s,a) := Q^{\pi}(s,a) - V^{\pi}(s).$$

It is evident that $A^{\pi}(s,a)$ measures how well a single action is compared with the average state value. Moreover, we use $V^*(s)$, $Q^*(s,a)$ and $A^*(s,a)$ to denote the corresponding value functions associated with the optimal policy π^*, and use $V^k(s)$, $Q^k(s,a)$ and $A^k(s,a)$ to denoted the corresponding value functions associated with the policy output by the algorithm in the k-th iteration. In the sequel we often use the shorthand notation for ease of exposition, for example,

$$\pi_{s,a} := \pi(a|s), \quad \pi_s := \pi(\cdot|s), \quad Q^{\pi}_{s,a} := Q^{\pi}(s,a), \quad \text{and} \quad Q^{\pi}_s := Q^{\pi}(s,\cdot).$$

Given a state $s \in \mathcal{S}$, the set of optimal actions $\mathcal{A}^*_s$ at state s is defined as,

$$\mathcal{A}^*_s = \arg\max_{a\in\mathcal{A}} Q^*(s,a) = \arg\max_{a\in\mathcal{A}} A^*(s,a).$$

Given a policy $\pi \in \Pi$, a state $s \in \mathcal{S}$ and a set $B \subset \mathcal{A}$, define

$$\pi_s(B) = \sum_{a\in B} \pi_s(a)$$

as the probability of π_s on B and denote by b^{π}_s the probability on non-optimal actions,

$$b^{\pi}_s = \pi_s(\mathcal{A} \setminus \mathcal{A}^*_s).$$

When b^{π}_s is small for any $s \in \mathcal{S}$, it is natural to expect that π will be close to be optimal. Thus, b^{π}_s is a very essential optimality measure of a policy. The set of π-optimal actions at state s, denoted $\mathcal{A}^{\pi}_s$, is defined as

$$\mathcal{A}^{\pi}_s = \arg\max_{a\in\mathcal{A}} A^{\pi}(s,a),$$

with $\mathcal{A}^k_s$ being the abbreviation of $\mathcal{A}^{\pi^k}_s$. The following quantity is quite central in the finite iteration convergence analysis, which has also appeared in previous works, see for example Mei et al. (2020); Khodadadian et al. (2021).

Definition 1 *The optimal advantage function gap Δ is defined as follows:*

$$\Delta := \min_{s\in\widetilde{S}, a\notin\mathcal{A}^*_s} |A^*(s,a)|, \tag{10}$$

where $\widetilde{S} = \{s \in \mathcal{S} : \mathcal{A}^*_s \neq \mathcal{A}\}$ denotes the set of states that have non-optimal actions.

Without loss of generality, we assume $\tilde{\mathcal{S}} \neq \emptyset$. It is trivial that $\Delta > 0$ since $A^*(s,a) < 0$ holds for all non-optimal actions. Additionally, we will make the following two standard assumptions about the reward and the initial state distribution.

Assumption 1 (Bounded reward) $r(s,a,s') \in [0,1],\ \forall s,s' \in \mathcal{S},\ a \in \mathcal{A}$.

Assumption 2 (Traversal initial distribution) $\tilde{\mu} := \min_{s\in\mathcal{S}} \mu(s) > 0$.

Recall that d_μ^π defined (7) is the state visitation measure following policy π. We use d_μ^* to the state visitation measure following the optimal policy π^* and use d_μ^k to denote the visitation measure following the policy output by the algorithm in the k-th iteration. For $\pi \in \Pi$, $\mu \in \Delta(\mathcal{S})$ and $s \in \mathcal{S}$, it follows immediately from Assumption 2 that

$$d_\mu^\pi(s) \geq (1-\gamma)\tilde{\mu}. \tag{11}$$

1.3 Organization of the paper

The rest of the paper is outlined as follows. In Section 2, some preliminary results are provided which be used in our later analysis. The sublinear convergence of PPG with any constant step size is discussed in Section 3, followed by the finite convergence in Section 4. The dimension-free bound for the finite iteration convergence of PI and VI is also presented in Section 4. In Section 5, we present the results of linear convergence and equivalence to PI under different step size selection rules.

2. Preliminaries

2.1 Useful lemmas

As we assume the reward function r is bounded in Assumption 1, all the value functions are bounded as they are discounted summations of rewards.

Lemma 2 *For any policy $\pi \in \Pi$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$V^\pi(s) \in \left[0, \frac{1}{1-\gamma}\right],\ \ Q^\pi(s,a) \in \left[0, \frac{1}{1-\gamma}\right],\ \ A^\pi(s,a) \in \left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right].$$

By leveraging the structure property of the MDP, we further have the lemma below.

Lemma 3 *For any policy π ,*

- $\|Q^* - Q^\pi\|_\infty \leq \gamma \|V^* - V^\pi\|_\infty$
- $\|A^* - A^\pi\|_\infty \leq \|V^* - V^\pi\|_\infty$
- $\|V^* - V^\pi\|_\infty \leq \frac{V^*(\rho) - V^\pi(\rho)}{\tilde{\rho}}$ *for any $\rho \in \Delta(\mathcal{S})$ such that $\tilde{\rho} := \min_{s\in\mathcal{S}} \rho(s) > 0$.*

Proof Recalling the definition of state-action value function (4), one has

$$|Q^\pi(s,a) - Q^*(s,a)| = \gamma \left| \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[V^\pi(s') - V^*(s') \right] \right| \leq \gamma \|V^\pi - V^*\|_\infty.$$

For the advantage function, one has

$$A^{\pi}(s,a) - A^{*}(s,a) = \left(V^{*}(s) - V^{\pi}(s)\right) - \left(Q^{*}(s,a) - Q^{\pi}(s,a)\right).$$

On the one hand,

$$A^{\pi}(s,a) - A^{*}(s,a) \leq V^{*}(s) - V^{\pi}(s) \leq \left\|V^{*} - V^{\pi}\right\|_{\infty}.$$

On the other hand,

$$A^{*}(s,a) - A^{\pi}(s,a) \leq Q^{*}(s,a) - Q^{\pi}(s,a) \leq \gamma \left\|V^{*} - V^{\pi}\right\|_{\infty}.$$

Thus $\left\|A^{\pi} - A^{*}\right\|_{\infty} \leq \left\|V^{*} - V^{\pi}\right\|_{\infty}$. For the bound on $\left\|V^{*} - V^{\pi}\right\|_{\infty}$, a direct computation yields

$$\left\|V^{*} - V^{\pi}\right\|_{\infty} \leq \sum_{s} \frac{\rho(s)}{\rho(s)} \left(V^{*}(s) - V^{\pi}(s)\right) \leq \frac{V^{*}(\rho) - V^{\pi}(\rho)}{\tilde{\rho}},$$

which concludes the proof. ■

The performance difference lemma below is a fundamental lemma in the analysis of RL algorithms (e.g. Agarwal et al., 2021; Mei et al., 2020; Khodadadian et al., 2021; Liu et al., 2023; Xiao, 2022). It characterizes the difference between the value functions of two arbitrary policies can be represented as the weighted average of the advantages.

Lemma 4 (Performance Difference Lemma, (Kakade and Langford, 2002)) *For any two policies π_1, π_2, and any $\rho \in \Delta(\mathcal{S})$, one has*

$$V^{\pi_1}(\rho) - V^{\pi_2}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\rho}^{\pi_1}} \left[\mathbb{E}_{a \sim \pi_1(\cdot|s)} \left[A^{\pi_2}(s,a)\right]\right].$$

Recall that b_s^{π} denotes the probability on non-optimal actions which can be viewed as an essential measure for the optimality of a policy. The following two lemmas establish the relation between b_s^{π} and the mismatch $V^{*}(\rho) - V^{\pi}(\rho)$.

Lemma 5 (Khodadadian et al., 2021, Theorem 3.1) *For any policy $\pi \in \Pi$ and $\rho \in \Delta(\mathcal{S})$,*

$$V^{*}(\rho) - V^{\pi}(\rho) \leq \frac{1}{(1-\gamma)^2} \cdot \mathbb{E}_{s \sim d_{\rho}^{\pi}} \left[b_s^{\pi}\right].$$

Lemma 6 *For any policy $\pi \in \Pi$ and $\rho \in \Delta(\mathcal{S})$,*

$$\mathbb{E}_{s \sim \rho} \left[b_s^{\pi}\right] \leq \frac{V^{*}(\rho) - V^{\pi}(\rho)}{\Delta}.$$

Proof According to the performance difference lemma,

$$\begin{aligned}
&V^*(\rho) - V^\pi(\rho) = -\left(V^\pi(\rho) - V^*(\rho)\right) \\
&= \frac{1}{1-\gamma}\sum_s d_\rho^\pi(s)\sum_a \pi(a|s)\cdot(-A^*(s,a)) \\
&= \frac{1}{1-\gamma}\sum_{s\in\tilde{\mathcal{S}}} d_\rho^\pi(s)\sum_{a\notin\mathcal{A}_s^*} \pi(a|s)\cdot|A^*(s,a)| \\
&\geq \frac{1}{1-\gamma}\sum_{s\in\tilde{\mathcal{S}}} d_\rho^\pi(s)\sum_{a\notin\mathcal{A}_s^*} \pi(a|s)\cdot\Delta \\
&= \frac{\Delta}{1-\gamma}\sum_{s\in\tilde{\mathcal{S}}} d_\rho^\pi(s)\, b_s^\pi \geq \Delta\cdot\mathbb{E}_{s\sim\rho}\left[b_s^\pi\right].
\end{aligned}$$

The proof is complete after rearrangement. ■

2.2 Basic facts about projection onto probability simplex

Recall that Euclidian projection onto the probability simplex is defined as

$$\mathrm{Proj}_{\Delta(\mathcal{A})}(p) = \arg\min_{y\in\Delta(\mathcal{A})}\|y-p\|^2.$$

This projection has an explicit expression, presented in the following lemma.

Lemma 7 *For arbitrary vector $p = (p_a)_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$,*

$$\mathrm{Proj}_{\Delta(\mathcal{A})}(p) = (p+\lambda\mathbf{1})_+$$

where λ is a constant such that $\sum_{a\in\mathcal{A}}(p_a+\lambda)_+ = 1$.

Proof This fact can be obtained by studying the KKT condition of the projection problem, see for example Wang and Carreira-Perpiñán (2013) for details. ■

Remark 8 *It's trivial to see that the projection onto probability simplex has a shift-invariant property. That is, $\mathrm{Proj}_{\Delta(\mathcal{A})}(p) = \mathrm{Proj}_{\Delta(\mathcal{A})}(p+c\mathbf{1})$ holds for arbitrary constant $c\in\mathbb{R}$. Therefore, PPG and PQA can also be expressed in terms of advantages functions. For example, we have the following alternative expression for PPG:*

$$\pi_s^{k+1} = \mathrm{Proj}_{\Delta(\mathcal{A})}\left(\pi_s^k + \frac{\eta_k d_\mu^k(s)}{1-\gamma}A^k(s,\cdot)\right), \quad \forall s\in\mathcal{S}.$$

Lemma 7 implies that the projection onto the probability simplex can be computed by first translating the vector with an offset, followed by truncating those negative values to zeros. Moreover, the next lemma provides a characterization on the support of the projection, which will be used frequently in our analysis.

Lemma 9 (Gap property) *Let $\mathcal{B}$ and $\mathcal{C}$ be two disjoint non-empty sets such that $\mathcal{A} = \mathcal{B} \cup \mathcal{C}$. Given an arbitrary vector $p = (p_a)_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, let $y = \mathrm{Proj}_{\Delta(\mathcal{A})}(p)$. Then*

$$\forall a' \in \mathcal{C},\ y_{a'} = 0 \Leftrightarrow \sum_{a\in\mathcal{B}} \left(p_a - \max_{a'\in\mathcal{C}} p_{a'}\right)_+ \geq 1.$$

Roughly speaking, this lemma indicates that if the entries of p in the index set $\mathcal{C}$ are generally smaller than those in the index set $\mathcal{B}$ and the cumulative gap is larger than 1, then the index set $\mathcal{C}$ will be excluded from the support set of the projection y. The proof of this lemma is essentially contained in the argument for Theorem 1 in Wang and Carreira-Perpiñán (2013). To keep the presentation self-contained, we give a short proof below.

Proof [Proof of Lemma 9] First note that $\forall a \in \mathcal{A}$,

$$y_a = 0 \overset{(a)}{\Longleftrightarrow} \lambda \leq -p_a \overset{(b)}{\Longleftrightarrow} \sum_{a'\in\mathcal{A}} (p_{a'} - p_a)_+ \geq 1,$$

where (a) follows from Lemma 7 and (b) is due to $\sum_{a\in\mathcal{A}}(p_a+\lambda)_+ = 1$ and the monotonicity of $(\cdot)_+$. Thus we have

$$\begin{aligned} y_{a'} = 0,\ \ \forall a' \in \mathcal{C} \Leftrightarrow \min_{a'\in\mathcal{C}} \sum_{a\in\mathcal{A}} (p_a - p_{a'})_+ \geq 1 &\Leftrightarrow \sum_{a\in\mathcal{A}} \left(p_a - \max_{a'\in\mathcal{C}} p_{a'}\right)_+ \geq 1 \\ &\Leftrightarrow \sum_{a\in\mathcal{B}} \left(p_a - \max_{a'\in\mathcal{C}} p_{a'}\right)_+ \geq 1, \end{aligned}$$

which completes the proof. ■

Based on Lemma 7, we can now rewrite the one-step update of PPG in (8) and PQA in (9) into a unified framework with an explicit expression for the projection. This is the prototype update that will be mainly analysed. Given an input policy $\pi \in \Pi$ and step size $\eta > 0$, the new policy π^+ is generated via

$$\text{(Prototype Update)} \qquad \pi^+_{s,a}(\eta_s) = \left(\pi_{s,a} + \eta_s A^\pi_{s,a} + \lambda_s\right)_+, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \qquad (12)$$

where λ_s is a constant such that $\sum_a \pi^+_{s,a} = 1$. Note that, given a constant step size η, $\eta_s = \frac{\eta d^\pi_\mu(s)}{1-\gamma}$ for PPG while $\eta_s = \eta$ for PQA.

2.3 Basic properties of prototype update

Before presenting the sublinear convergence of PPG, we give a brief discussion on the basic properties of the prototype update. The lemma below presents all the possibilities for the support set of the new policy $\pi^+_s(\eta)$ (the subscript s in η_s will be omitted in this section for simplicity).

Lemma 10 *Consider the prototype update in (12). Denote by $\mathcal{B}_s(\eta)$ the support set of π^+_s:*

$$\mathcal{B}_s(\eta) := \left\{a : \pi^+_{s,a}(\eta) > 0\right\}.$$

Then for any $\eta > 0$, $\mathcal{B}_s(\eta)$ admits one of the following three forms:

1. $\mathcal{B}_s(\eta) \subsetneqq \mathcal{A}_s^\pi$,
2. $\mathcal{B}_s(\eta) = \mathcal{A}_s^\pi$,
3. $\mathcal{B}_s(\eta) = \mathcal{A}_s^\pi \cup \mathcal{C}_s(\eta)$, *where* $\mathcal{C}_s(\eta) \subseteq \mathcal{A} \setminus \mathcal{A}_s^\pi$ *is not empty.*

Proof Without loss of generality, assume $\mathcal{A}_s^\pi \neq \mathcal{A}$. Then it suffices to show that if $\mathcal{B}_s(\eta)$ contains an action $a' \notin \mathcal{A}_s^\pi$, all π-optimal actions are included in $\mathcal{B}_s(\eta)$. Given any $\hat{a} \in \mathcal{A}_s^\pi$, define

$$\hat{\mathcal{A}}_s = \left\{a \in \mathcal{A} : \pi_{s,a} + \eta A_{s,a}^\pi \geq \pi_{s,\hat{a}} + \eta A_{s,\hat{a}}^\pi\right\}.$$

Next we will show that

$$\begin{aligned} I :=& \sum_{a\in\mathcal{A}} \left(\pi_{s,a} + \eta A_{s,a}^\pi - \left(\pi_{s,\hat{a}} + \eta A_{s,\hat{a}}^\pi\right)\right)_+ \\ =& \sum_{a\in\hat{\mathcal{A}}_s} \left(\pi_{s,a} + \eta A_{s,a}^\pi - \left(\pi_{s,\hat{a}} + \eta A_{s,\hat{a}}^\pi\right)\right) \\ =& \sum_{a\in\hat{\mathcal{A}}_s\cap\mathcal{A}_s^\pi} (\pi_{s,a} - \pi_{s,\hat{a}}) + \sum_{a\in\hat{\mathcal{A}}_s\setminus\mathcal{A}_s^\pi} \left(\pi_{s,a} - \pi_{s,\hat{a}} + \eta\left(A_{s,a}^\pi - A_{s,\hat{a}}^\pi\right)\right) < 1, \end{aligned}$$

from which the claim follows immediately using Lemma 9.

If $\hat{\mathcal{A}}_s \setminus \mathcal{A}_s^\pi \neq \emptyset$, then

$$\sum_{a\in\hat{\mathcal{A}}_s\setminus\mathcal{A}_s^\pi} \left(\pi_{s,a} - \pi_{s,\hat{a}} + \eta\left(A_{s,a}^\pi - A_{s,\hat{a}}^\pi\right)\right) < \sum_{a\in\hat{\mathcal{A}}_s\setminus\mathcal{A}_s^\pi} (\pi_{s,a} - \pi_{s,\hat{a}}),$$

since $A_{s,a}^\pi < A_{s,\hat{a}}^\pi$ for $a \in \hat{\mathcal{A}}_s \setminus \mathcal{A}_s^\pi$. Consequently,

$$I < \sum_{a\in\hat{\mathcal{A}}_s\cap\mathcal{A}_s^\pi} (\pi_{s,a} - \pi_{s,\hat{a}}) + \sum_{a\in\hat{\mathcal{A}}_s\setminus\mathcal{A}_s^\pi} (\pi_{s,a} - \pi_{s,\hat{a}}) \leq 1.$$

On the other hand, if $\hat{\mathcal{A}}_s \setminus \mathcal{A}_s^\pi = \emptyset$, one has $\hat{\mathcal{A}}_s \subset \mathcal{A}_s^\pi$. In this case,

$$I = \sum_{a\in\mathcal{A}_s^\pi} (\pi_{s,a} - \pi_{s,\hat{a}}) \leq \pi_s\left(\mathcal{A}_s^\pi\right) < 1,$$

where the last inequality is due to the fact that $\mathcal{B}_s(\eta)$ contains an action $a' \notin \mathcal{A}_s^\pi$. ■

The last lemma implies that at least one π-optimal action is included in the support set $\mathcal{B}_s(\eta)$. In addition, it is not hard to verify that when step size η goes to infinity every π-suboptimal actions will be excluded from the support set of π_s^+, which suggests that $\mathcal{B}_s(\eta)$ might shrink as η increases. The following lemma confirms that this observation is indeed true.

Lemma 11 *For* $\eta_1 > \eta_2 > 0$ *we have*

$$\mathcal{B}_s(\eta_1) \subseteq \mathcal{B}_s(\eta_2).$$

Proof Since the relation holds trivially when $\mathcal{B}_s(\eta_2) = \mathcal{A}$, we only consider the case $\mathcal{B}_s(\eta_2) \neq \mathcal{A}$. First the application of Lemma 9 yields that

$$\sum_{a\in\mathcal{B}_s(\eta)} \left[\pi_{s,a} + \eta A^\pi_{s,a} - \max_{a'\notin\mathcal{B}_s(\eta)} (\pi_{s,a'} + \eta A^\pi_{s,a'})\right]_+ \geq 1 \tag{13}$$

and

$$a' \notin \mathcal{B}_s(\eta) \Longleftrightarrow \sum_{a\neq a'} \left[\pi_{s,a} + \eta A^\pi_{s,a} - (\pi_{s,a'} + \eta A^\pi_{s,a'})\right]_+ \geq 1. \tag{14}$$

If $\mathcal{B}_s(\eta_2) \subseteq \mathcal{A}^\pi_s$ (i.e. the first two cases in Lemma 10), then for any $a' \notin \mathcal{B}_s(\eta_2)$ we have

$$\begin{aligned}
&\sum_{a\neq a'} \left[\pi_{s,a} + \eta_1 A^\pi_{s,a} - (\pi_{s,a'} + \eta_1 A^\pi_{s,a'})\right]_+ \\
&\geq \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_1 A^\pi_{s,a} - (\pi_{s,a'} + \eta_1 A^\pi_{s,a'})\right]_+ \\
&\overset{(a)}{\geq} \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_2 A^\pi_{s,a} - (\pi_{s,a'} + \eta_2 A^\pi_{s,a'})\right]_+ \geq 1,
\end{aligned}$$

where (a) is due to the fact $(A^\pi_{s,a} - A^\pi_{s,a'}) \geq 0$ for $\forall a \in \mathcal{B}_s(\eta_2) \subseteq \mathcal{A}^\pi_s$. This implies that $a' \notin \mathcal{B}_s(\eta_1)$, and thus $\mathcal{B}_s(\eta_1) \subseteq \mathcal{B}_s(\eta_2)$.

For the case that $\mathcal{B}_s(\eta_2) = \mathcal{A}_s \cup \mathcal{C}_s(\eta_2)$, fixing $a' \notin \mathcal{B}_s(\eta_2)$, it follows from (13) that

$$\begin{aligned}
&\sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_2 A^\pi_{s,a} - (\pi_{s,a'} + \eta_2 A^\pi_{s,a'})\right]_+ \\
&\geq \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_2 A^\pi_{s,a} - \max_{a'\notin\mathcal{B}_s(\eta_2)} (\pi_{s,a'} + \eta_2 A^\pi_{s,a'})\right]_+ \geq 1.
\end{aligned}$$

Furthermore, since $a' \notin \mathcal{B}_s(\eta_2)$ it is trivial to see that $\forall a \in \mathcal{B}_s(\eta_2)$, $\pi_{s,a} + \eta_2 A^\pi_{s,a} > \pi_{s,a'} + \eta_2 A^\pi_{s,a'}$. Therefore,

$$\begin{aligned}
&\sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_2 A^\pi_{s,a} - (\pi_{s,a'} + \eta_2 A^\pi_{s,a'})\right]_+ \\
&= \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} + \eta_2 A^\pi_{s,a} - (\pi_{s,a'} + \eta_2 A^\pi_{s,a'})\right] \\
&= \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_2 (A^\pi_{s,a} - A^\pi_{s,a'})\right] \\
&\geq 1,
\end{aligned}$$

which yields $\sum_{a\in\mathcal{B}_s(\eta_2)} (A^\pi_{s,a} - A^\pi_{s,a'}) \geq 0$, as $\eta_2 > 0$. Consequently,

$$\sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_1 (A^\pi_{s,a} - A^\pi_{s,a'})\right]_+ - \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_2 (A^\pi_{s,a} - A^\pi_{s,a'})\right]_+$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_1(A^\pi_{s,a} - A^\pi_{s,a'})\right]_+ - \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_2(A^\pi_{s,a} - A^\pi_{s,a'})\right] \\
&\geq \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_1(A^\pi_{s,a} - A^\pi_{s,a'})\right] - \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_2(A^\pi_{s,a} - A^\pi_{s,a'})\right] \\
&= (\eta_1 - \eta_2) \sum_{a\in\mathcal{B}_s(\eta_2)} (A^\pi_{s,a} - A^\pi_{s,a'}) > 0,
\end{aligned}
$$

yielding

$$
\begin{aligned}
&\sum_{a\neq a'} \left[\pi_{s,a} - \pi_{s,a'} + \eta_1(A^\pi_{s,a} - A^\pi_{s,a'})\right]_+ \geq \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_1(A^\pi_{s,a} - A^\pi_{s,a'})\right]_+ \\
&\geq \sum_{a\in\mathcal{B}_s(\eta_2)} \left[\pi_{s,a} - \pi_{s,a'} + \eta_2(A^\pi_{s,a} - A^\pi_{s,a'})\right]_+ \geq 1.
\end{aligned}
$$

Together with (14) we have $a' \notin \mathcal{B}_s(\eta_1)$, which implies $\mathcal{B}_s(\eta_1) \subseteq \mathcal{B}_s(\eta_2)$. ■

The following lemma shows that $A^\pi_{s,a}$ should be sufficiently large in order to be included in the support set of π^+_s, which is reasonable.

Lemma 12 *Consider the prototype update in* (12)*. We have*

$$
A^\pi_{s,a} \geq \max_{\tilde{a}\in\mathcal{A}} A^\pi_{s,\tilde{a}} - \frac{2\pi_s\left(\mathcal{A}\setminus\mathcal{A}^\pi_s\right)}{\eta}, \quad \forall a \in \mathcal{B}_s(\eta).
$$

Proof It suffices to consider the third case in Lemma 10. According to Lemma 9, for any action $a \in \mathcal{B}_s(\eta)\setminus\mathcal{A}^\pi_s$, we have

$$
\begin{aligned}
1 &> \sum_{a'\in\mathcal{A}} \left(\pi_{s,a'} + \eta A^\pi_{s,a'} - \pi_{s,a} - \eta A^\pi_{s,a}\right)_+ \\
&\geq \sum_{a'\in\mathcal{A}^\pi_s} \left(\pi_{s,a'} + \eta A^\pi_{s,a'} - \pi_{s,a} - \eta A^\pi_{s,a}\right)_+ \\
&= \sum_{a'\in\mathcal{A}^\pi_s} \left(\pi_{s,a'} - \pi_{s,a} + \eta\left(\max_{\tilde{a}\in\mathcal{A}} A^\pi_{s,\tilde{a}} - A^\pi_{s,a}\right)\right)_+ \\
&\geq \sum_{a'\in\mathcal{A}^\pi_s} \left(\pi_{s,a'} - \pi_{s,a} + \eta\left(\max_{\tilde{a}\in\mathcal{A}} A^\pi_{s,\tilde{a}} - A^\pi_{s,a}\right)\right) \\
&= \pi_s\left(\mathcal{A}^\pi_s\right) + \left|\mathcal{A}^\pi_s\right|\left(\eta\left(\max_{\tilde{a}\in\mathcal{A}} A^\pi_{s,\tilde{a}} - A^\pi_{s,a}\right) - \pi_{s,a}\right).
\end{aligned}
$$

It follows that

$$
\eta\left(\max_{\tilde{a}\in\mathcal{A}} A^\pi_{s,\tilde{a}} - A^\pi_{s,a}\right) \leq \pi_{s,a} + \frac{1-\pi_s\left(\mathcal{A}^\pi_s\right)}{\left|\mathcal{A}^\pi_s\right|} \leq \left(1+\frac{1}{\left|\mathcal{A}^\pi_s\right|}\right)\pi_s\left(\mathcal{A}\setminus\mathcal{A}^\pi_s\right) \leq 2\pi_s\left(\mathcal{A}\setminus\mathcal{A}^\pi_s\right).
$$

The proof is complete after rearrangement. ■

3. Sublinear convergence of PPG for any constant step size

As already mentioned, the sublinear convergence of PQA for any constant step size has already been developed (see for example Xiao, 2022; Lan, 2021. *Even though PPG and PQA are overall similar to each other, to the best of our knowledge, the technique for the sublinear convergence analysis of PQA cannot be used to establish the sublinear convergence of PPG for any constant step sizes due to the existence of the visitation measure.* Instead, we fill this gap by utilizing the explicit form of the projection onto the probability simplex to establish the lower bound for the one-step improvement,

$$\sum_{a\in\mathcal{A}} \pi_{s,a}^{k+1} A_{s,a}^k \geq \frac{\left(\max\limits_{a\in\mathcal{A}} A_{s,a}^k\right)^2}{\max\limits_{a\in\mathcal{A}} A_{s,a}^k + C}, \quad \forall s \in \mathcal{S}.$$

Combining this result with the performance difference lemma yields that

$$V^{k+1}(\rho) - V^k(\rho) \geq \mathcal{O}\left(\left(V^*(\rho) - V^k(\rho)\right)^2\right),$$

which directly implies the sublinear convergence of PPG.

Following the notation in the prototype update, the key ingredient in our analysis is

$$f_s(\eta_s) := \sum_{a\in\mathcal{A}} \pi_{s,a}^+(\eta_s) A_{s,a}^\pi,$$

where $\eta_s = \frac{\eta d_\mu^\pi(s)}{1-\gamma}$ for PPG. We first give an expression for $f_s(\eta_s)$.

Lemma 13 (Improvement expression) *Consider the prototype update in* (12)*. For any* $\eta_s > 0$ *one has*

$$\begin{aligned} f_s(\eta_s) = \eta_s &\left(\sum_{a\in\mathcal{B}_s(\eta_s)} \left(A_{s,a}^\pi\right)^2 - \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{a\in\mathcal{B}_s(\eta_s)} A_{s,a}^\pi\right)^2\right) \\ &+ \sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a'} \left(\frac{1}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)} (A_{s,a}^\pi - A_{s,a'}^\pi)\right). \end{aligned}$$

The proof of Lemma 13 is deferred to Section 3.1. Based on this lemma, we are able to derive a lower bound for $f_s(\eta_s)$, as stated in the next lemma whose proof is deferred to Section 3.2.

Theorem 14 (Improvement lower bound) *Consider the update in* (12)*. For any* $\eta_s > 0$ *one has*

$$f_s(\eta_s) \geq \frac{\left(\max_a A_{s,a}^\pi\right)^2}{\max_a A_{s,a}^\pi + \frac{2+5|\mathcal{A}|}{\eta_s}}.$$

With this lower bound, the sublinear convergence of PPG can be established together with the performance difference lemma.

Theorem 15 (Sublinear convergence of PPG) *With any constant step size $\eta_k = \eta$ and distribution $\rho \in \Delta(\mathcal{S})$, the policy sequence π^k generated by PPG satisfies*

$$V^*(\rho) - V^k(\rho) \leq \frac{1}{k}\frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^*}{\rho}\right\|_\infty \left(1 + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}\right). \tag{15}$$

Remark 16 *Note that Assumption 2 on the initial distribution is necessary for us to establish the convergence of PPG. Otherwise, consider the following two-state MDP,*

$$\begin{gathered}\mathcal{S} = \{s_0, s_1\}, \quad \mathcal{A} = \{a_0, a_1\},\\ P(s_0|s_0,a_0) = P(s_0|s_0,a_1) = P(s_1|s_1,a_0) = P(s_1|s_1,a_1) = 1,\\ r(s_0,a_0) = r(s_1,a_0) = 0, \quad r(s_0,a_1) = r(s_1,a_1) = 1.\end{gathered}$$

That is, s_0 and s_1 are not connected and a_1 is the optimal action for both the states. Assume $\mu(s_0) = 0$ and $\mu(s_1) = 1$. It can be easily verified that for any $\pi \in \Pi$ there holds $d_\mu^\pi(s_0) = 0$, thus PPG (equation (8)) does not update the policy on s_0, i.e., $\pi^k(\cdot|s_0) = \pi^0(\cdot|s_0)$ for all k. Therefore, V^k generated by PPG cannot converge to V^* if $\pi^0(a_1|s_0) < 1$. Despite this negative example, we would like to point out it is not clear whether the traversal initial distribution is still needed if the states are connected.

Remark 17 *Compared with the previous results* (Xiao, 2022; Agarwal et al., 2021; Zhang et al., 2020)*, the result in* (15) *removes the constraint $\eta_k \leq \frac{1}{L}$ on the step size, where $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$ is the smoothness coefficient of the value function. The best sublinear convergence rate for PPG in prior works is achieved when $\eta_k = \frac{1}{L}$, leading to the result*

$$V^*(\mu) - V^k(\mu) \leq \mathcal{O}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^5\,k}\left\|\frac{d_\mu^*}{\mu}\right\|_\infty^2\right). \tag{16}$$

By setting $\eta_k = \frac{1}{L} = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ and $\rho = \mu$ in (15) *we can obtain the bound*

$$\begin{aligned}V^*(\mu) - V^k(\mu) &\leq \frac{1}{k}\frac{1}{(1-\gamma)^2}\left\|\frac{d_\mu^*}{\mu}\right\|_\infty \left(1 + \frac{2\gamma\,|\mathcal{A}|\,(2+5\,|\mathcal{A}|)}{(1-\gamma)^3\,\tilde{\mu}}\right)\\ &= \mathcal{O}\left(\frac{1}{k}\frac{|\mathcal{A}|^2}{(1-\gamma)^5}\left\|\frac{d_\mu^*}{\mu}\right\|_\infty \frac{1}{\tilde{\mu}}\right).\end{aligned}$$

Compared with (16)*, the new bound has the same dependency on the discounted factor. Moreover, Theorem 15 suggests that the best sublinear convergence rate for PPG is indeed achieved when $\eta_k = \eta \geq \frac{2+5|\mathcal{A}|}{\tilde{\mu}}$ rather than $\eta_k = \frac{1}{L}$, yielding the rate*

$$\mathcal{O}\left(\frac{1}{k}\frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^*}{\rho}\right\|_\infty\right).$$

Remark 18 *Our analysis technique is also available for the establishment of the sublinear convergence of PQA. However, the result is not as tight as the one obtained in* Xiao (2022); Lan (2021) *based on the particular structure of PQA within the framework of policy mirror ascent. Thus, we omit the details. It is worth emphasizing again that, to the best of our knowledge, the analysis technique in* Xiao (2022); Lan (2021) *for PQA is not applicable for PPG due the existence of the visitation measure.*

Proof [Proof of Theorem 15] By the performance difference lemma (Lemma 4) and Theorem 14, one has

$$\begin{aligned}
V^{k+1}(\rho) - V^{k}(\rho) &= \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} d_{\rho}^{k+1}(s)\sum_{a\in\mathcal{A}}\pi_{s,a}^{k+1}A_{s,a}^{k} \geq \mathbb{E}_{s\sim\rho}\left[\frac{\left(\max\limits_{a\in\mathcal{A}} A_{s,a}^{k}\right)^{2}}{\max\limits_{a\in\mathcal{A}} A_{s,a}^{k} + \frac{2+5|\mathcal{A}|}{\eta_s^k}}\right] \\
&\overset{(a)}{\geq} \mathbb{E}_{s\sim\rho}\left[\frac{\left(\max\limits_{a\in\mathcal{A}} A_{s,a}^{k}\right)^{2}}{\max\limits_{a\in\mathcal{A}} A_{s,a}^{k} + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}}\right] = \mathbb{E}_{s\sim\rho}\left[g\left(\max_{a\in\mathcal{A}} A_{s,a}^{k}\right)\right] \\
&\geq \left\|\frac{d_{\rho}^{*}}{\rho}\right\|_{\infty}^{-1}\mathbb{E}_{s\sim d_{\rho}^{*}}\left[g\left(\max_{a\in\mathcal{A}} A_{s,a}^{k}\right)\right] \\
&\overset{(b)}{\geq} \left\|\frac{d_{\rho}^{*}}{\rho}\right\|_{\infty}^{-1}\cdot g\left(\mathbb{E}_{s\sim d_{\rho}^{*}}\left[\max_{a\in\mathcal{A}} A_{s,a}^{k}\right]\right),
\end{aligned} \tag{17}$$

where $g(x) := \frac{x^2}{x+\frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}}$ is a convex and monotonically increasing function when $x \geq 0$, (a) is due to $\eta_s^k = \frac{\eta d_{\mu}^{k}(s)}{1-\gamma} \geq \eta\tilde{\mu}$ according to inequality (11), and (b) is due to the Jensen Inequality. Notice that

$$\begin{aligned}
\mathbb{E}_{s\sim d_{\rho}^{*}}\left[\max_{a\in\mathcal{A}} A_{s,a}^{k}\right] &\geq \mathbb{E}_{s\sim d_{\rho}^{*}}\left[\mathbb{E}_{a\sim\pi_s^{*}}\left[A_{s,a}^{k}\right]\right] = (1-\gamma)\left[\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\rho}^{*}}\left[\mathbb{E}_{a\sim\pi_s^{*}}\left[A_{s,a}^{k}\right]\right]\right] \\
&= (1-\gamma)\left(V^{*}(\rho) - V^{k}(\rho)\right).
\end{aligned} \tag{18}$$

Let $\delta_k := V^{*}(\rho) - V^{k}(\rho)$. As g is monotonically increasing, plugging (17) into (18) yields that

$$\delta_k - \delta_{k+1} \geq \left\|\frac{d_{\rho}^{*}}{\rho}\right\|_{\infty}^{-1}\cdot g\left((1-\gamma)\delta_k\right) = \left\|\frac{d_{\rho}^{*}}{\rho}\right\|_{\infty}^{-1}\cdot\frac{(1-\gamma)^2\delta_k^2}{(1-\gamma)\delta_k + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}}. \tag{19}$$

Since $\delta_k \leq \frac{1}{1-\gamma}$ by Lemma 2, we have

$$\delta_k - \delta_{k+1} \geq \left\|\frac{d_{\rho}^{*}}{\rho}\right\|_{\infty}^{-1}\cdot\frac{(1-\gamma)^2\delta_k^2}{1 + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}}.$$

This inequality implies that δ_k is monotonically decreasing. Dividing both sides by δ_k^2 yields

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} = \frac{\delta_k - \delta_{k+1}}{\delta_k \delta_{k+1}} \geq \frac{\delta_k - \delta_{k+1}}{\delta_k^2} \geq \left\| \frac{d_\rho^*}{\rho} \right\|_\infty^{-1} \cdot \frac{(1-\gamma)^2}{1 + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}}.$$

Consequently,

$$\begin{aligned}\delta_k = \frac{1}{\frac{1}{\delta_k}} = \frac{1}{\delta_0 + \sum_{i=0}^{k-1}\left(\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}\right)} &\leq \frac{1}{\sum_{i=0}^{k-1}\left(\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}\right)} \\ &\leq \frac{1}{k}\left\| \frac{d_\rho^*}{\rho} \right\|_\infty \left(\frac{1}{(1-\gamma)^2} + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}(1-\gamma)^2} \right),\end{aligned}$$

and the proof is complete. ■

3.1 Proof of Lemma 13

Noting that

$$\begin{aligned}1 = \sum_{a\in\mathcal{A}} \pi_{s,a}^+ = \sum_{a\in\mathcal{B}_s(\eta_s)} \pi_{s,a}^+(\eta_s) &= \sum_{a\in\mathcal{B}_s(\eta_s)} \left[\pi_{s,a} + \eta_s A_{s,a}^\pi + \lambda_s(\eta_s)\right]_+ \\ &= \sum_{a\in\mathcal{B}_s(\eta_s)} \left[\pi_{s,a} + \eta_s A_{s,a}^\pi + \lambda_s(\eta_s)\right],\end{aligned}$$

we have

$$\begin{aligned}\lambda_s(\eta_s) &= \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(1 - \sum_{a\in\mathcal{B}_s(\eta_s)} \left[\pi_{s,a} + \eta_s A_{s,a}^\pi\right]\right) \\ &= \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{a\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a} - \eta_s \sum_{a\in\mathcal{B}_s(\eta_s)} A_{s,a}^\pi\right). \qquad (20)\end{aligned}$$

It follows that

$$\begin{aligned}f_s(\eta_s) = \sum_{a\in\mathcal{B}_s(\eta_s)} \pi_{s,a}^+(\eta_s) A_{s,a}^\pi &= \sum_{a\in\mathcal{B}_s(\eta_s)} \left[\pi_{s,a} + \eta_s A_{s,a}^\pi + \lambda_s(\eta_s)\right] A_{s,a}^\pi \\ &\overset{(a)}{=} \lambda_s(\eta_s) \sum_{a\in\mathcal{B}_s(\eta_s)} A_{s,a}^\pi + \eta_s \sum_{a\in\mathcal{B}_s(\eta_s)} \left(A_{s,a}^\pi\right)^2 - \sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a'} A_{s,a'}^\pi \\ &= \eta_s \left(\sum_{a\in\mathcal{B}_s(\eta_s)} (A_{s,a}^\pi)^2 - \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{a\in\mathcal{B}_s(\eta_s)} A_{s,a}^\pi \right)^2 \right) \\ &+ \sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a'} \left(\frac{1}{|\mathcal{B}_s(\eta_s)|} \sum_{a\in\mathcal{B}_s(\eta_s)} (A_{s,a}^\pi - A_{s,a'}^\pi) \right),\end{aligned}$$

where (a) utilizes the fact $\sum_{a\in\mathcal{B}_s(\eta_s)} \pi_{s,a} A_{s,a}^\pi = 0$.

3.2 Proof of Theorem 14

Without loss of generality, we only consider the case $\mathcal{B}_s(\eta_s) \setminus \mathcal{A}_s^\pi \neq \emptyset$. First recall the expression of $f_s(\eta_s)$ in Lemma 13:

$$f_s(\eta_s) = \eta_s \Bigg(\underbrace{\sum_{a\in\mathcal{B}_s(\eta_s)} (A_{s,a}^\pi)^2 - \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{a\in\mathcal{B}_s(\eta_s)} A_{s,a}^\pi\right)^2}_{I_1} \Bigg) + \underbrace{\sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a'}\left(\frac{1}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)}(A_{s,a}^\pi - A_{s,a'}^\pi)\right)}_{I_2}.$$

For the term I_1, it is evident that

$$I_1 = |\mathcal{B}_s(\eta_s)|\left(\mathbb{E}_{a\sim U}\left[\left(A_{s,a}^\pi\right)^2\right] - \left(\mathbb{E}_{a\sim U}\left[A_{s,a}^\pi\right]\right)^2\right) = |\mathcal{B}_s(\eta_s)|\cdot \mathrm{Var}_{a\sim U}\left[A_{s,a}^\pi\right],$$

where U denotes the uniform distribution on $\mathcal{B}_s(\eta_s)$. Letting $\Delta_{s,a}^\pi := \max_{a'\in\mathcal{A}} A_{s,a'}^\pi - A_{s,a}^\pi$,

$$\begin{aligned}
I_1 &= |\mathcal{B}_s(\eta_s)|\cdot \mathrm{Var}_{a\sim U}\left[A_{s,a}^\pi\right] = |\mathcal{B}_s(\eta_s)|\cdot \mathrm{Var}_{a\sim U}\left[\Delta_{s,a}^\pi\right] \\
&= \sum_{a\in\mathcal{B}_s(\eta_s)} (\Delta_{s,a}^\pi)^2 - \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{a\in\mathcal{B}_s(\eta_s)} \Delta_{s,a}^\pi\right)^2 \\
&\overset{(a)}{=} \sum_{a\in\mathcal{B}_s(\eta_s)} (\Delta_{s,a}^\pi)^2 - \frac{1}{|\mathcal{B}_s(\eta_s)|}\left(\sum_{\tilde{a}\in\mathcal{B}_s(\eta_s)\setminus\mathcal{A}_s^\pi} \Delta_{s,\tilde{a}}^\pi\right)^2 \\
&\geq \sum_{a\in\mathcal{B}_s(\eta_s)} (\Delta_{s,a}^\pi)^2 - \frac{|\mathcal{B}_s(\eta_s)\setminus\mathcal{A}_s^\pi|}{|\mathcal{B}_s(\eta_s)|}\sum_{\tilde{a}\in\mathcal{B}_s(\eta_s)\setminus\mathcal{A}_s^\pi} (\Delta_{s,\tilde{a}}^\pi)^2 \\
&= \left(1 - \frac{|\mathcal{B}_s(\eta_s)\setminus\mathcal{A}_s^\pi|}{|\mathcal{B}_s(\eta_s)|}\right)\sum_{a\in\mathcal{B}_s(\eta_s)} (\Delta_{s,a}^\pi)^2 = \frac{|\mathcal{B}_s(\eta_s)\cap A_s^\pi|}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)} \left(\Delta_{s,a}^\pi\right)^2 \\
&\geq \frac{1}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)} \left(\Delta_{s,a}^\pi\right)^2,
\end{aligned}$$

where (a) leverages the property that $\Delta_{s,a}^\pi = 0$ for π-optimal actions $a \in \mathcal{A}_s^\pi$.

For the term I_2, we can rewrite it through the notation Δ^π as follows:

$$I_2 = \sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)} \pi_{s,a'}\left(\Delta_{s,a'}^\pi - \bar{\Delta}_s^\pi\right),$$

where $\bar{\Delta}_s^\pi := \frac{1}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)}\Delta_{s,a}^\pi$. By lemma 9, for any action $a' \notin \mathcal{B}_s(\eta)$ we have

$$\begin{aligned}
&\sum_{a\in\mathcal{B}_s(\eta)}\left[\pi_{s,a}+\eta A_{s,a}^\pi-(\pi_{s,a'}+\eta A_{s,a'}^\pi)\right] \geq \sum_{a\in\mathcal{B}_s(\eta)}\left[\pi_{s,a}+\eta A_{s,a}^\pi-\max_{a'\notin\mathcal{B}_s(\eta)}(\pi_{s,a'}+\eta A_{s,a'}^\pi)\right]\\
&\overset{(a)}{=}\sum_{a\in\mathcal{B}_s(\eta)}\left[\pi_{s,a}+\eta A_{s,a}^\pi-\max_{a'\notin\mathcal{B}_s(\eta)}(\pi_{s,a'}+\eta A_{s,a'}^\pi)\right]_+ \geq 1,
\end{aligned}$$

where (a) is due to $a\in\mathcal{B}_s(\eta)$ and $a'\notin\mathcal{B}_s(\eta)$, implying $\pi_{s,a}+\eta A_{s,a}^\pi > \pi_{s,a'}+\eta A_{s,a'}^\pi$. It follows that

$$\begin{aligned}
1 &\leq \sum_{a\in\mathcal{B}_s(\eta)}\left[\pi_{s,a}-\pi_{s,a'}+\eta(A_{s,a}^\pi-A_{s,a'}^\pi)\right]\\
&= \eta\sum_{a\in\mathcal{B}_s(\eta)}(A_{s,a}^\pi-A_{s,a'}^\pi)-|\mathcal{B}_s(\eta)|\pi_{s,a'}+\sum_{a\in\mathcal{B}_s(\eta)}\pi_{s,a},
\end{aligned}$$

which yields

$$\frac{1}{|\mathcal{B}_s(\eta)|}\sum_{a\in\mathcal{B}_s(\eta)}\left(A_{s,a}^\pi-A_{s,a'}^\pi\right) \geq \frac{1}{\eta\,|\mathcal{B}_s(\eta)|}\left(1-\sum_{a\in\mathcal{B}_s(\eta)}\pi_{s,a}+|\mathcal{B}_s(\eta)|\,\pi_{s,a'}\right) \geq \frac{\pi_{s,a'}}{\eta}.$$

Using the notation of $\Delta_{s,a}^\pi$ and $\bar{\Delta}_s^\pi$, this inequality can be reformulated as

$$\forall a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s):\quad \Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi \geq \frac{\pi_{s,a'}}{\eta_s}. \tag{21}$$

Let $\mathcal{B}_s(0) := \mathrm{supp}(\pi_s) = \{a:\pi_{s,a}>0\}$. By (21) we know that

$$\forall a'\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0):\quad \Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi \geq \frac{\pi_{s,a'}}{\eta_s} > 0. \tag{22}$$

Furthermore,

$$I_2 = \sum_{a'\in\mathcal{A}\setminus\mathcal{B}_s(\eta_s)}\pi_{s,a'}\left(\Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi\right) = \sum_{a'\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)}\pi_{s,a'}\left(\Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi\right).$$

Combining I_1 and I_2 together, we have

$$f_s(\eta_s) \geq \frac{\eta_s}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)}\left(\Delta_{s,a}^\pi\right)^2+\sum_{a'\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)}\pi_{s,a'}(\Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi).$$

By Cauchy-Schwarz Inequality,

$$\begin{aligned}
&f_s(\eta_s)\times\left(\frac{|\mathcal{B}_s(\eta_s)|}{\eta_s}\sum_{a\in\mathcal{B}_s(\eta_s)}(\pi_{s,a})^2+\sum_{a'\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)}\pi_{s,a'}\frac{\left(\Delta_{s,a'}^\pi\right)^2}{\Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi}\right)\\
&\geq\left(\frac{\eta_s}{|\mathcal{B}_s(\eta_s)|}\sum_{a\in\mathcal{B}_s(\eta_s)}\left(\Delta_{s,a}^\pi\right)^2+\sum_{a'\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)}\pi_{s,a'}\left(\Delta_{s,a'}^\pi-\bar{\Delta}_s^\pi\right)\right)
\end{aligned}$$

$$\times \left(\frac{|\mathcal{B}_s(\eta_s)|}{\eta_s} \sum_{a \in \mathcal{B}_s(\eta_s)} (\pi_{s,a})^2 + \sum_{a' \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a'} \frac{\left(\Delta^\pi_{s,a'}\right)^2}{\Delta^\pi_{s,a'} - \bar{\Delta}^\pi_s} \right)$$
$$\geq \left(\sum_{a \in \mathcal{B}_s(\eta_s)} \pi_{s,a} \Delta^\pi_{s,a} + \sum_{a' \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a'} \Delta^\pi_{s,a'} \right)^2 = \left(\max_{a \in \mathcal{A}} A^\pi_{s,a} \right)^2.$$

Therefore, we can obtain

$$f_s(\eta_s) \geq \frac{1}{G} \left(\max_{a \in \mathcal{A}} \mathcal{A}^\pi_{s,a} \right)^2,$$

where

$$G := \underbrace{\frac{|\mathcal{B}_s(\eta_s)|}{\eta_s} \sum_{a \in \mathcal{B}_s(\eta_s)} (\pi_{s,a})^2}_{G_1} + \underbrace{\sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} \frac{\left(\Delta^\pi_{s,a}\right)^2}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s}}_{G_2}.$$

Next we will give an upper bound of G. For the term G_1, it is straightforward to see that

$$G_1 < \frac{|\mathcal{B}_s(\eta_s)|}{\eta_s} \left(\sum_{a \in \mathcal{B}_s(\eta_s)} \pi_{s,a} \right)^2 < \frac{|\mathcal{A}|}{\eta_s}.$$

For the term G_2, a direct computation yields that,

$$\begin{aligned}
G_2 &= \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} \frac{\left(\Delta^\pi_{s,a}\right)^2}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s} \\
&= \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} \frac{\left(\Delta^\pi_{s,a}\right)^2 - \left(\bar{\Delta}^\pi_s\right)^2 + \left(\bar{\Delta}^\pi_s\right)^2}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s} \\
&= \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} \left(\Delta^\pi_{s,a} + \bar{\Delta}^\pi_s + \frac{\left(\bar{\Delta}^\pi_s\right)^2}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s} \right) \\
&\leq \sum_{a \in \mathcal{A}} \pi_{s,a} \Delta^\pi_{s,a} + \bar{\Delta}^\pi_s \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} + \left(\bar{\Delta}^\pi_s\right)^2 \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \frac{\pi_{s,a}}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s} \\
&= \max_{a \in \mathcal{A}} A^\pi_{s,a} + \bar{\Delta}^\pi_s \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} + \left(\bar{\Delta}^\pi_s\right)^2 \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \frac{\pi_{s,a}}{\Delta^\pi_{s,a} - \bar{\Delta}^\pi_s}. \qquad (23)
\end{aligned}$$

Lemma 12 shows that

$$\forall a \in \mathcal{B}_s(\eta_s) : \Delta^\pi_{s,a} \leq \frac{2\pi_s(\mathcal{A} \setminus \mathcal{A}^\pi_s)}{\eta_s} \leq \frac{2}{\eta_s} \quad \Longrightarrow \quad \bar{\Delta}^\pi_s \leq \frac{2}{\eta_s}. \qquad (24)$$

Plugging (22) and (24) into (23) we have

$$G_2 \leq \max_{a \in \mathcal{A}} A^\pi_{s,a} + \frac{2}{\eta_s} \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \pi_{s,a} + \left(\frac{2}{\eta_s}\right)^2 \sum_{a \in (\mathcal{A} \setminus \mathcal{B}_s(\eta_s)) \cap \mathcal{B}_s(0)} \frac{\pi_{s,a}}{\frac{1}{\eta_s} \pi_{s,a}}$$

$$= \max_{a\in\mathcal{A}} A^\pi_{s,a} + \frac{2}{\eta_s} \sum_{a\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)} \pi_{s,a} + \frac{4}{\eta_s} \sum_{a\in(\mathcal{A}\setminus\mathcal{B}_s(\eta_s))\cap\mathcal{B}_s(0)} 1$$
$$\le \max_{a\in\mathcal{A}} A^\pi_{s,a} + \frac{2+4\left|\mathcal{A}\right|}{\eta_s}.$$

Thus we can finally obtain

$$G = G_1 + G_2 \le \max_{a\in\mathcal{A}} A^\pi_{s,a} + \frac{2+5\left|\mathcal{A}\right|}{\eta_s},$$

and

$$f_s(\eta_s) \ge \frac{\left(\max_a A^\pi_{s,a}\right)^2}{G} \ge \frac{\left(\max_a A^\pi_{s,a}\right)^2}{\max_a A^\pi_{s,a} + \frac{2+5|\mathcal{A}|}{\eta_s}}. \tag{25}$$

4. Finite iteration convergence results

4.1 Finite iteration convergence of PPG and PQA

In this section, we show that both PPG and PQA output an optimal policy after a finite iteration k_0 and we will use the sublinear analysis (Theorem 15 for PPG and (32) for PQA) to derive an upper bound of k_0. The overall idea is first sketched as follows. For an arbitrary $s \in \mathcal{S}$, letting $\mathcal{B} = \mathcal{A}^*_s$, $\mathcal{C} = \mathcal{A} \setminus \mathcal{A}^*_s$ in Lemma 9 (recall that $\mathcal{A}^*_s$ is the set of optimal actions, i.e. π^*-optimal actions), we have

$$\forall a' \notin \mathcal{A}^*_s,\ \ \pi^+_{s,a'} = 0 \iff \sum_{a\in\mathcal{A}^*_s} \left(\pi_{s,a} + \eta_s A^\pi_{s,a} - \max_{a'\notin\mathcal{A}^*_s} \left(\pi_{s,a'} + \eta_s A^\pi_{s,a'}\right)\right)_+ \ge 1. \tag{26}$$

By the definition of b^π_s and Δ,

$$b^\pi_s = \pi_s\left(\mathcal{A}\setminus\mathcal{A}^*_s\right), \qquad \Delta = \min_{s\in\tilde{S}, a\notin\mathcal{A}^*_s} \left|A^*(s,a)\right|,$$

when V^π is sufficiently close to V^*, we know that for any $a' \notin \mathcal{A}^*_s$,

$$\sum_{a\in\mathcal{A}^*_s} \pi_{s,a} \approx 1,\ \ A^\pi_{s,a} - A^\pi_{s,a'} \approx A^*_{s,a} - A^*_{s,a'} \ge \Delta.$$

Since $\pi_{s,a'} \le b^\pi_s$, we asymptotically have

$$\sum_{a\in\mathcal{A}^*_s} \left(\pi_{s,a} + \eta_s A^\pi_{s,a} - \max_{a'\notin\mathcal{A}^*_s} \left(\pi_{s,a'} + \eta_s A^\pi_{s,a'}\right)\right)_+ \ge 1 - \mathcal{O}\left(b^\pi_s\right) + \mathcal{O}\left(\Delta\right).$$

This implies that if b^π_s is sufficiently small, the condition in (26) will be met. Then both PPG and PQA will output the optimal policy.

Lemma 19 (Optimality condition) *Consider the prototype update in* (12)*. Define*

$$\varepsilon^\pi_{s,a} := \eta_s\left(A^\pi_{s,a} - A^*_{s,a}\right) \quad \textit{and} \quad \varepsilon^\pi_s := [\varepsilon^\pi_{s,a}]_{a\in\mathcal{A}}.$$

If the input policy π satisfies,

$$b_s^\pi + \|\varepsilon_s^\pi\|_\infty \leq \frac{\eta_s \Delta}{2}, \quad \forall s \in \mathcal{S}, \tag{27}$$

then π^+ is an optimal policy.

Proof For any $s \in \mathcal{S}$, a direction computation yields that

$$\begin{aligned}
&\sum_{a \in \mathcal{A}_s^*} \left(\pi_{s,a} + \eta_s A_{s,a}^\pi - \max_{a' \notin \mathcal{A}_s^*} \left(\pi_{s,a'} + \eta_s A_{s,a'}^\pi \right) \right)_+ \\
&\geq \sum_{a \in \mathcal{A}_s^*} \left(\pi_{s,a} + \eta_s A_{s,a}^\pi - \max_{a' \notin \mathcal{A}_s^*} \left(\pi_{s,a'} + \eta_s A_{s,a'}^\pi \right) \right) \\
&\overset{(a)}{=} \sum_{a \in \mathcal{A}_s^*} \left(\pi_{s,a} + \eta_s A_{s,a}^\pi - \left(\pi_{s,\tilde{a}} + \eta_s A_{s,\tilde{a}}^\pi \right) \right) \\
&= \sum_{a \in \mathcal{A}_s^*} \left[\left(\pi_{s,a} - \eta_s \left| A_{s,a}^* \right| + \varepsilon_{s,a}^k \right) - \left(\pi_{s,\tilde{a}} - \eta_s \left| A_{s,\tilde{a}}^* \right| + \varepsilon_{s,\tilde{a}}^\pi \right) \right] \\
&\overset{(b)}{=} \sum_{a \in \mathcal{A}_s^*} \left[\left(\pi_{s,a} + \varepsilon_{s,a}^k \right) - \left(\pi_{s,\tilde{a}} - \eta_s \left| A_{s,\tilde{a}}^* \right| + \varepsilon_{s,\tilde{a}}^\pi \right) \right] \\
&\overset{(c)}{\geq} \sum_{a \in \mathcal{A}_s^*} \left[\left(\pi_{s,a} - \pi_{s,\tilde{a}} \right) + \eta_s \Delta + \left(\varepsilon_{s,a}^\pi - \varepsilon_{s,\tilde{a}}^\pi \right) \right] \geq \sum_{a \in \mathcal{A}_s^*} \left[\left(\pi_{s,a} - b_s^\pi \right) + \eta_s \Delta - 2 \left\| \varepsilon_s^\pi \right\|_\infty \right] \\
&= \sum_{a \in \mathcal{A}_s^*} \pi_{s,a} + \left| \mathcal{A}_s^* \right| \left(\eta_s \Delta - b_s^\pi - 2 \left\| \varepsilon_s^\pi \right\|_\infty \right) = 1 - b_s^\pi + \left| \mathcal{A}_s^* \right| \left(\eta_s \Delta - b_s^\pi - 2 \left\| \varepsilon_s^\pi \right\|_\infty \right) \\
&\geq 1 - \left| \mathcal{A}_s^* \right| b_s^\pi + \left| \mathcal{A}_s^* \right| \left(\eta_s \Delta - b_s^\pi - 2 \left\| \varepsilon_s^\pi \right\|_\infty \right) = 1 + \left| \mathcal{A}_s^* \right| \left[\eta_s \Delta - 2 \left(b_s^\pi + \left\| \varepsilon_s^\pi \right\|_\infty \right) \right] \geq 1,
\end{aligned}$$

where $\tilde{a} := \operatorname{argmax}_{a' \notin \mathcal{A}_s^*} \left\{ \pi_{s,a'}^k + \eta_s A_{s,a'}^k \right\}$ in (a), (b) is due to $A_{s,a}^* = 0$ for all $a \in \mathcal{A}_s^*$, and (c) follows from the definition of Δ. Combining this result with Lemma 9 we obtain that

$$\pi_{s,a'}^+ = 0, \quad \forall\, a' \notin \mathcal{A}_s^*.$$

which means π^+ is an optimal policy. ■

Next we will show that the LHS of (27) is actually of order $\mathcal{O}\left(\left\|V^* - V^k\right\|_\infty\right)$. Thus the condition (27) can be satisfied provided the value error converges to zero and step sizes are constant (in this case the RHS of (27) is $\mathcal{O}(\Delta)$).

Lemma 20 (Optimality condition continued in terms of state values) *Consider the prototype update in* (12)*. If the state values of the input policy π satisfies,*

$$\|V^* - V^\pi\|_\infty \leq \frac{\Delta}{2} \frac{\eta_s \Delta}{1 + \eta_s \Delta}, \quad \forall s \in \mathcal{S}, \tag{28}$$

then π^+ is an optimal policy.

Proof For any $s \in \mathcal{S}$, setting $\rho(\cdot) = \mathbb{I}(\cdot = s)$ in Lemma 6, where $\mathbb{I}$ is the indicator function, yields that

$$b_s^\pi \leq \frac{V^*(s) - V^\pi(s)}{\Delta} \leq \frac{\|V^* - V^\pi\|_\infty}{\Delta}.$$

Combining this result with Lemma 3 we have

$$\max_{s\in\mathcal{S}} b_s^\pi + \|\varepsilon_s^\pi\|_\infty \leq \frac{1}{\Delta}\|V^* - V^\pi\|_\infty + \eta_s \|V^* - V^\pi\|_\infty = \left(\frac{1}{\Delta} + \eta_s\right)\|V^* - V^\pi\|_\infty. \quad (29)$$

The proof is completed by noting the assumption and Lemma 19. ■

Since the sublinear convergence of PPG (Theorem 15) and PQA ((32)) has already been established, there must exist an iteration k_0 such that $\left\|V^* - V^k\right\|_\infty$ is smaller than the threshold given in Lemma 20.

Theorem 21 (Finite iteration convergence of PPG) *With any constant step size $\eta_k = \eta > 0$, PPG terminates after at most*

$$k_0 := \left\lceil \frac{2}{\Delta}\left(1 + \frac{1}{\eta\tilde{\mu}\Delta}\right)\frac{1}{\tilde{\mu}(1-\gamma)^2}\left\|\frac{d_\mu^*}{\mu}\right\|_\infty \left(1 + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}\right)\right\rceil$$

iterations.

Proof Since $\eta_s^k = \frac{\eta d_\mu^k(s)}{1-\gamma} > \eta\tilde{\mu}$ for PPG, the RHS of (28) satisfies

$$\frac{\Delta}{2}\frac{\eta_s^k\Delta}{1+\eta_s^k\Delta} \geq \frac{\Delta}{2}\frac{\eta\tilde{\mu}\Delta}{1+\eta\tilde{\mu}\Delta}. \quad (30)$$

According to Lemma 3 and Theorem 15,

$$\begin{aligned}\left\|V^* - V^k\right\|_\infty &\leq \frac{V^*(\mu) - V^k(\mu)}{\tilde{\mu}} \leq \frac{1}{k}\frac{1}{(1-\gamma)^2}\left\|\frac{d_\mu^*}{\mu}\right\|_\infty \frac{1}{\tilde{\mu}}\left(1 + \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}\right) \\ &\leq \frac{\Delta}{2}\frac{\eta_s\Delta}{1+\eta_s\Delta}, \end{aligned} \quad (31)$$

where the last inequality follows from (30) and the expression of k_0. ■

Theorem 22 (Finite iteration convergence of PQA) *With any constant step size $\eta_k = \eta > 0$, PQA terminates after at most*

$$k_0 := \left\lceil \frac{2}{\Delta}\left(1 + \frac{1}{\eta\Delta}\right)\left(\frac{1}{\eta(1-\gamma)} + \frac{1}{(1-\gamma)^2}\right) - 1\right\rceil$$

iterations.

Proof Note that the following sublinear convergence of PQA has been established in Xiao (2022) for any constant step size,

$$V^*(\rho) - V^k(\rho) \leq \frac{1}{k+1}\left(\frac{\mathbb{E}_{s\sim d_\rho^*}\left[\left\|\pi_s^* - \pi_s^0\right\|_2^2\right]}{2\eta(1-\gamma)} + \frac{1}{(1-\gamma)^2}\right). \tag{32}$$

Plugging $\rho_s(\cdot) := \mathbb{I}\{\cdot = s\}$ into (32) yields that

$$\begin{aligned} V^*(s) - V^k(s) &\leq \frac{1}{k+1}\left(\frac{\mathbb{E}_{s\sim d_{\rho_s}^*}\left[\left\|\pi_s^* - \pi_s^0\right\|_2^2\right]}{2\eta(1-\gamma)} + \frac{1}{(1-\gamma)^2}\right) \\ &\leq \frac{1}{k+1}\left(\frac{1}{\eta(1-\gamma)} + \frac{1}{(1-\gamma)^2}\right). \end{aligned}$$

Since this bound holds for any s, it also holds for $\|V^* - V^k\|_\infty$. Then it can be easily verified that the condition in Lemma 20 is satisfied given the expression of k_0. ■

Before proceeding, we give two short discussions on the finite iteration convergence of PPG and PQA. Firstly, it will be shown that a condition similar to that in Lemma 19 can be obtained based on the optimality condition of the optimization problem. Secondly, though the finite iteration convergence for the homotopic PQA is discussed in Li et al. (2023), it does not imply the finite iteration convergence of PQA for any constant step size. A simple bandit example is used to illustrate that the homotopic PQA requires sufficiently large step size to converge (in fact, the finite iteration of the homotopic PQA is established for exponentially increasing step sizes in Li et al. (2023)).

4.1.1 Short discussion I

Recall that the update (12) corresponds to the following optimization:

$$\pi_s^+ = \arg\max_{p\in\Delta(\mathcal{A})}\left\{\eta_s\langle Q^\pi(s,\cdot), p\rangle - \frac{1}{2}\|p-\pi_s\|_2^2\right\} = \arg\max_{p\in\Delta(\mathcal{A})}\left\{\eta_s\langle A^\pi(s,\cdot), p\rangle - \frac{1}{2}\|p-\pi_s\|_2^2\right\}.$$

The optimality condition for this problem is given by (see for example Rockafellar, 1996)

$$\langle \eta_s A^\pi(s,\cdot) - \pi_s^+ + \pi_s,\ p' - \pi_s^+\rangle \leq 0, \quad \forall\, p' \in \Delta(\mathcal{A}). \tag{33}$$

Define $N_\Delta(p)$ as the normal cone of $\Delta(\mathcal{A})$ at p,

$$N_\Delta(p) = \{g \mid g^T(p'-p) \leq 0,\ \forall\, p' \in \Delta(\mathcal{A})\}.$$

The condition in (33) can be equivalently expressed as

$$\eta_s A^\pi(s,\cdot) - \pi_s^+ + \pi_s \in N_\Delta(\pi_s^+).$$

Moreover, note that (see for example Beck, 2017)

$$N_\Delta(\pi_s^+) = \left\{(g_1, \cdots, g_{|\mathcal{A}|}) \mid g_i \leq g_j = g_\ell,\ \forall\, i \notin \mathrm{supp}(\pi_s^+) \text{ and } \forall\, j,\ell \in \mathrm{supp}(\pi_s^+)\right\}.$$

Therefore, if $\forall\, s \in \mathcal{S}$, it can be shown that there exists $g^{\pi}_{s,\cdot} \in N_{\Delta}(\pi^{+}_{s})$, such that $\forall\, a \in \mathcal{A}^{*}_{s}$ and $a' \notin \mathcal{A}^{*}_{s}$,

$$g^{\pi}_{s,a} - g^{\pi}_{s,a'} = \left(\eta_s A^{\pi}_{s,a} - \pi^{+}_{s,a} + \pi_{s,a}\right) - \left(\eta_s A^{\pi}_{s,a'} - \pi^{+}_{s,a'} + \pi_{s,a'}\right) > 0, \tag{34}$$

we can conclude that

$$\forall s \in \mathcal{S},\, a' \notin \mathcal{A}^{*}_{s}: \quad a' \notin \operatorname{supp}(\pi^{+}_{s}),$$

which implies π^{+} is an optimal policy.

Recalling the definition of $\varepsilon^{\pi}_{s,a} = \eta_s \left(A^{\pi}_{s,a} - A^{*}_{s,a}\right)$ in Lemma 19, one has

$$\begin{aligned} g^{\pi}_{s,a} - g^{\pi}_{s,a'} &= \left(\eta_s A^{*}_{s,a} + \varepsilon^{\pi}_{s,a} - \pi^{+}_{s,a} + \pi_{s,a}\right) - \left(\eta_s A^{*}_{s,a'} + \varepsilon^{\pi}_{s,a'} - \pi^{+}_{s,a'} + \pi_{s,a'}\right) \\ &= \eta_s \left(A^{*}_{s,a} - A^{*}_{s,a'}\right) + (\varepsilon^{\pi}_{s,a} - \varepsilon^{\pi}_{s,a'}) - (\pi^{+}_{s,a} - \pi_{s,a}) + (\pi^{+}_{s,a'} - \pi_{s,a'}) \\ &\geq \eta_s \Delta - 2\|\varepsilon^{\pi}_{s}\|_{\infty} - \|\pi^{+}_{s} - \pi_{s}\|_{\infty} - b^{\pi}_{s}. \end{aligned} \tag{35}$$

In addition, setting $p' = \pi_s$ in (33) yields

$$\begin{aligned} \|\pi^{+}_{s} - \pi_{s}\|_2^2 &\leq \eta_s \sum_{a} \pi^{+}_{s,a} A^{\pi}_{s,a} \leq \eta_s \sum_{s'} \sum_{a} \pi^{+}_{s',a} A^{\pi}_{s',a} = \eta_s \sum_{s'} \frac{d^{\pi^{+}}_{\mu}(s')}{d^{\pi^{+}}_{\mu}(s')} \sum_{a} \pi^{+}_{s',a} A^{\pi}_{s',a} \\ &\leq \frac{\eta_s}{(1-\gamma)\tilde{\mu}} (V^{\pi^{+}}(\mu) - V^{\pi}(\mu)) \leq \frac{\eta_s}{(1-\gamma)\tilde{\mu}} (V^{*}(\mu) - V^{\pi}(\mu)). \end{aligned}$$

Together with (35), one has $g^{\pi}_{s,a} - g^{\pi}_{s,a'} > 0$ provided

$$b^{\pi}_{s} + 2\|\varepsilon^{\pi}_{s}\|_{\infty} + \min\left\{\sqrt{\frac{\eta_s}{(1-\gamma)\tilde{\mu}}(V^{*}(\mu) - V^{\pi}(\mu))},\ 1\right\} < \eta_s \Delta.$$

It is clear that this condition (but not as concise as the one presented in Lemma 19) can also be used to derive the finite iteration convergence of PPG and PQA for any constant step size.

4.1.2 Short discussion II

In Li et al. (2023), the finite iteration convergence of homotopic policy mirror ascent methods under certain Bregman divergence is investigated. When considering the case where the Bregman divergence is generated by the squared Euclidean distance, it reduces to the following homotopic PQA method:

$$\begin{aligned} \pi^{k+1}_{s} &= \arg\max_{p \in \Delta}\ \eta_k \left[\langle Q^{k}(s,\cdot), p\rangle - \frac{\tau_k}{2}\|p - \pi^{0}_{s}\|_2^2\right] - \frac{1}{2}\|p - \pi^{k}_{s}\|_2^2 \\ &= \arg\min_{p \in \Delta}\ \frac{1}{2}\left\|p - \frac{1}{1+\eta_k\tau_k}\pi^{k}_{s} - \frac{\eta_k}{1+\eta_k\tau_k} Q^{k}(s,\cdot)\right\|_2^2 \\ &= \operatorname{Proj}_{\Delta}\left(\frac{1}{1+\eta_k\tau_k}\pi^{k}_{s} + \frac{\eta_k}{1+\eta_k\tau_k} Q^{k}(s,\cdot)\right) \end{aligned}$$

$$= \mathrm{Proj}_{\Delta}\left(\frac{1}{1+\eta_k\tau_k}\pi_s^k + \frac{\eta_k}{1+\eta_k\tau_k}A^k(s,\cdot)\right),$$

where π_s^0 is a uniform policy, τ_k is the regularization parameter, and the last line follows from the fact that $\frac{\eta_k}{1+\eta_k\tau_k}V^k(s)\cdot 1$ is a vector with all the same entries. It follows that there exists λ_s^k such that[1]

$$\pi_{s,a}^{k+1} = \frac{1}{1+\eta_k\tau_k}\left(\pi_{s,a}^k + \eta_k A^k(s,a) - \lambda_s^k\right)_+ \quad \text{and} \quad \sum_a \pi_{s,a}^{k+1} = 1.$$

Consider the case where $\eta_k\tau_k$ is fixed, for example $1+\eta_k\tau_k = 1/\gamma$ with $0<\gamma<1$ as considered in Li et al. (2023). Then the update reduces to

$$\pi_{s,a}^{k+1} = \frac{1}{1/\gamma}\left(\pi_{s,a}^k + \eta_k A^k(s,a) - \lambda_s^k\right)_+ \quad \text{and} \quad \sum_a \left(\pi_{s,a}^k + \eta_k A^k(s,a) - \lambda_s^k\right)_+ = \frac{1}{\gamma}. \quad (36)$$

Note that this update is overall similar to the update of PQA, differing only in the extra factor $\frac{1}{1/\gamma}$. However, next we will use a very simple example to show that it requires η_k to be sufficiently large for (36) to be able to convergence. Therefore, even the finite iteration convergence of (36) holds, it does not leads to the finite iteration convergence of PQA for any constant step size.

More precisely, consider the bandit case where there are only two actions a_1 and a_2. Assume a_1 is the single optimal action. Suppose π^k is already optimal, i.e., $\pi_{a_1}^k = 1$ and $\pi_{a_2}^k = 0$. Then $A_{a_1}^k = 0$ and $A_{a_2}^k < 0$. Letting $\Delta = |A_{a_2}^k|$, there exists a λ^k such that

$$\pi_{a_1}^{k+1} = \gamma(1-\lambda^k)_+, \quad \pi_{a_2}^{k+1} = \gamma(-\eta_k\Delta_k - \lambda^k)_+.$$

Moreover,

$$(1-\lambda^k)_+ + (-\eta_k\Delta - \lambda^k)_+ = \frac{1}{\gamma} > 1. \quad (37)$$

First note that there must hold $\lambda^k < 0$; otherwise the above equality cannot hold since $\eta_k\Delta > 0$. Assume $\eta_k\Delta < \frac{1}{\gamma} - 1$. Then it is easy to verify by contradiction that one should have $-\lambda^k > \eta_k\Delta$ in order to satisfy (37). It follows that

$$\lambda^k = \frac{1}{2}(1 - 1/\gamma - \eta_k\Delta) > 1 - \frac{1}{\gamma}.$$

Therefore, when $\eta_k\Delta < \frac{1}{\gamma} - 1$, one has $\pi_{a_1}^{k+1} = \gamma(1-\lambda^k)_+ < 1$. That is, π^{k+1} is not optimal anymore. In other words, in order for π^{k+1} still to be optimal, one must have $\eta_k\Delta \geq 1/\gamma - 1$, that is, $\eta_k \geq (1/\gamma - 1)/\Delta$ which can very large when Δ is small.

1. Note that in Li et al. (2023), a slightly different version is indeed considered. That is, if $\pi_{s,a}^k = 0$, the starting point can be negative due to the requirement for the careful selection of the subgradient in order to establish the finite iteration convergence of the algorithm for exponentially increasing step sizes.

4.2 Finite iteration convergence of PI and VI

As a by-product, we will derive a new dimension-free bound for the finite iteration convergence of policy iteration (PI) and value iteration (VI) in terms of Δ in this section. The following lemma demonstrates that once a vector is sufficiently close to the optimal value vector, then the policy retrieved from that vector is an optimal policy.

Lemma 23 *For any $V \in \mathbb{R}^{|\mathcal{S}|}$ (not necessarily associated with a policy), define $Q^V \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as follows:*

$$Q^V(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[r(s,a,s') + \gamma V(s')].$$

If $\gamma\|V^* - V\|_\infty \le \frac{\Delta}{3}$, then $\arg\max_a Q^V(s,) \subset \mathcal{A}_s^*$. That is, the greedy policy supported on $\arg\max_a Q^V(s,a)$ is an optimal policy.

Proof First, it is easy to see that $\forall s, a$,

$$|Q^*(s,a) - Q^V(s,a)| = \gamma|\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s') - V(s')]| \le \gamma\|V^* - V\|_\infty \le \frac{\Delta}{3}.$$

It follows that for s having non-optimal actions, $a \in \mathcal{A}_s^*$ and $a' \notin \mathcal{A}_s^*$, we have

$$Q^V(s,a) \ge Q^*(s,a) - \frac{\Delta}{3} \ge Q^*(s,a') + \frac{2\Delta}{3} \ge Q^V(s,a') + \frac{\Delta}{3} > Q^V(s,a'),$$

which concludes the proof. ■

Theorem 24 (Finite iteration convergence of PI) *PI terminates after at most*

$$k_0 = \left\lceil \frac{1}{1-\gamma}\log\left(\frac{3}{(1-\gamma)\Delta}\right)\right\rceil$$

iterations.

Proof Notice that the value error generated by PI satisfies (see for example Bertsekas, 2019 for a proof)

$$\left\|V^* - V^k\right\|_\infty \le \gamma^k \left\|V^* - V^0\right\|_\infty \le \frac{\gamma^k}{1-\gamma},$$

According to Lemma 23, when

$$\frac{\gamma^{k+1}}{1-\gamma} \le \frac{\Delta}{3}, \tag{38}$$

we have $\mathcal{A}_s^k \subset \mathcal{A}_s^*$ after that. It's trivial to verify that (38) holds for π^k when $k \ge k_0$. Since PI puts all the probabilities on the action set $\mathcal{A}_s^k$ in each iteration, we have $\mathcal{A}_s^k \subset \mathcal{A}_s^*$ when $k \ge k_0$, which implies PI outputs an optimal policy after k_0. ■

Remark 25 *It is well-known that PI is a strong polynomial algorithm (see for example* Scherrer, 2013*), which means PI outputs an optimal policy after* $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}\log\frac{1}{1-\gamma}\right)$ *iterations. Compared with this strong polynomial bound, the bound in Theorem 24 is dimension-free but relies on the parameter* Δ *that depends on the particular MDP problem. The dimension-free bound is better in the case* $\frac{1}{\Delta} = o\left(\frac{1}{(1-\gamma)^{|\mathcal{S}||\mathcal{A}|}}\right)$.

Theorem 26 *Let* π^k *be the sequence of greedy policy generated by* V^k *in VI (Note that* V^k *is not necessarily a value function of* π^k*). Then after at most*

$$k_0 := \left\lceil \frac{1}{1-\gamma}\log\left(\frac{3\|V^* - V^0\|_\infty}{\Delta}\right)\right\rceil$$

iterations, π^k *is an optimal policy.*

Proof The value error generated by VI satisfies

$$\left\|V^* - V^k\right\|_\infty \leq \gamma^k\|V^* - V^0\|_\infty \leq \frac{\Delta}{3},$$

where the second inequality follows from the assumption. Then the application of Lemma 23 concludes the proof. ■

Remark 27 *It is worth noting that since VI does not evaluate the value function of* π^k *in each iteration, Theorem 26 does not really mean the algorithm terminates in a finite number of iterations.*

5. Linear convergence and equivalence to PI

5.1 Linear convergence of PPG under non-adaptive increasing step sizes

In Theorem 15, we have established the sublinear convergence of PPG for constant step sizes. In this section, we further show that with increasing step sizes $\eta_k \geq \mathcal{O}\left(\frac{1}{\gamma^{2k}}\right)$, the classical γ-rate linear convergence of PPG can be achieved globally. Note that this result can indeed be obtained based on a similar argument for PQA (see Johnson et al., 2023). Here, for the sake of self-completeness, we present a different proof based on Lemma 12 instead of the three point descent lemma used in Johnson et al. (2023).

Theorem 28 *Consider the prototype update in* (12)*. Suppose the step size in the* k*-th iteration satisfies*

$$\eta_s^k \geq \frac{1}{\gamma^{2k+1}c_0}\cdot 2\pi_s^k\left(\mathcal{A}\setminus\mathcal{A}_s^k\right), \quad \forall s\in\mathcal{S}, \tag{39}$$

for a given constant $c_0 > 0$*. Then the value errors satisfy*

$$\left\|V^* - V^k\right\|_\infty < \gamma^k\left(\left\|V^* - V^0\right\|_\infty + \frac{c_0}{1-\gamma}\right).$$

Proof For simplicity of notation, let $\tilde{\eta}_s^k := \frac{2\pi_s^k\left(\mathcal{A}\backslash\mathcal{A}_s^k\right)}{\eta_s^k}$. According to Lemma 12, for any $k > 0$ and $s \in \mathcal{S}$,

$$\sum_{a\in\mathcal{A}} \pi_{s,a}^{k+1} Q_{s,a}^k \geq \sum_{a\in\mathcal{A}} \pi_{s,a}^{k+1} \left(\max_{\tilde{a}\in\mathcal{A}} Q_{s,\tilde{a}}^k - \tilde{\eta}_s^k\right) = \max_{\tilde{a}\in\mathcal{A}} Q_{s,\tilde{a}}^k - \tilde{\eta}_s^k.$$

Then

$$V^*(s) - V^{k+1}(s) = V^*(s) - \mathbb{E}_{a\sim\pi_s^{k+1}}\left[Q_{s,a}^{k+1}\right] \leq V^*(s) - \mathbb{E}_{a\sim\pi_s^{k+1}}\left[Q_{s,a}^k\right]$$
$$\leq V^*(s) - \left(\max_{\tilde{a}\in\mathcal{A}} Q_{s,\tilde{a}}^k - \tilde{\eta}_s^k\right) = \max_{a\in\mathcal{A}} Q_{s,a}^* - \max_{\tilde{a}\in\mathcal{A}} Q_{s,\tilde{a}}^k + \tilde{\eta}_s^k \leq \gamma \left\|V^* - V^k\right\|_\infty + \tilde{\eta}_s^k,$$

where in the first inequality we have used the fact $Q_{s,a}^{k+1} \geq Q_{s,a}^k$ due to the improvement. It follows that

$$\begin{aligned}\left\|V^* - V^k\right\|_\infty &\leq \gamma \left\|V^* - V^{k-1}\right\|_\infty + \max_{s\in\mathcal{S}} \tilde{\eta}_s^k \\ &\leq \gamma^2 \left\|V^* - V^{k-2}\right\|_\infty + \gamma \max_{s\in\mathcal{S}} \tilde{\eta}_s^{k-1} + \max_{s\in\mathcal{S}} \tilde{\eta}_s^k \leq ... \\ &\leq \gamma^k \left\|V^* - V^0\right\|_\infty + \sum_{i=0}^{k-1} \left(\max_{s\in\mathcal{S}} \tilde{\eta}_s^i\right) \gamma^{k-1-i}. \end{aligned} \tag{40}$$

Notice that the condition (39) is equivalent to $\max_{s\in\mathcal{S}} \tilde{\eta}_s^i \leq c_0\gamma^{2i+1}$. Plugging it into (40) yields

$$\left\|V^* - V^k\right\|_\infty \leq \gamma^k \left\|V^* - V^0\right\|_\infty + c_0 \sum_{i=0}^{k-1} \gamma^{2i+1}\gamma^{k-1-i} < \gamma^k \left(\left\|V^* - V^0\right\|_\infty + \frac{c_0}{1-\gamma}\right),$$

which completes the proof. ■

The γ-rate linear convergence of PPG follows immediately by noting that $\eta_s^k = \eta_k \frac{d_\mu^k(s)}{1-\gamma}$ in PPG and $d_\mu^k(s) \geq (1-\gamma)\tilde{\mu}$.

Proposition 29 *For PPG, if* $\eta_k \geq \frac{1}{\tilde{\mu}} \frac{1}{c_0} \frac{2}{\gamma^{2k+1}}$, *then*

$$\left\|V^* - V^k\right\|_\infty < \gamma^k \left(\left\|V^* - V^0\right\|_\infty + \frac{c_0}{1-\gamma}\right). \tag{41}$$

Remark 30 *Recalling from Lemma 20 that when the value error satisfies*

$$\left\|V^* - V^k\right\|_\infty \leq \frac{\Delta}{2} \frac{\eta_s^k \Delta}{1 + \eta_s^k \Delta}, \tag{42}$$

the prototype update in (12) *outputs an optimal policy. Using the step sizes in Proposition 29 for PPG, it is easy to see that the RHS of* (42) *satisfies*

$$\frac{\Delta}{2} \frac{\eta_s^k \Delta}{1 + \eta_s^k \Delta} = \frac{\Delta}{2}\left(1 - \frac{1}{1 + \eta_s^k \Delta}\right) \geq \frac{\Delta}{2}\left(1 - \frac{1}{1 + \eta_k \tilde{\mu} \Delta}\right) \geq \frac{\Delta}{2}\left(\frac{2}{c_0 + 2}\right). \tag{43}$$

Combining (41), (42) *and* (43) *together implies that, after at most*

$$k_0 := \left\lceil \frac{1}{1-\gamma} \log\left(\frac{(c_0+1)(c_0+2)}{(1-\gamma)\Delta} \right) \right\rceil$$

iterations, PPG with the non-adaptive increasing step sizes achieves exact convergence.

5.1.1 Empirically Verify the Convergence of PPG

Here we conduct numerical experiments to verify the convergence of PPG under the constant and geometrically increasing step sizes on a random MDP problem. The sizes of the state and action spaces are set to $|\mathcal{S}| = 50$, $|\mathcal{A}| = 15$, and the discount factor is set to $\gamma = 0.98$. The reward $r(s,a)$ and transition model $P(s'|s,a)$ are uniformly generated from $[0,1]$ (P is further normalized to be a probability matrix), and μ is chosen to be the uniform distribution over $\mathcal{S}$ (so $\tilde{\mu} = 1/|\mathcal{S}|$). PPG is tested with a constant step size $\eta_k = 2$ and the geometrically increasing step size of the form $\eta_k = 0.05/(\tilde{\mu}\gamma^{2k+1})$, and we plot the log value error $\log\left(V^*(\mu) - V^k(\mu)\right)$ versus the number of iterations in Figure 1. It is evident that PPG with increasing step size exhibits faster convergence than that with constant step size. Moreover, in both cases, PPG can find an optimal policy within a finite number of iterations.

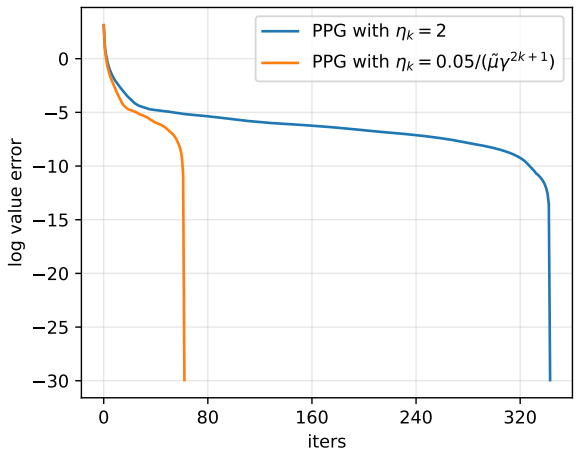

Figure 1: Numerical tests of PPG on a random MDP problem.

5.1.2 Discussion about Inexact PPG

As in Xiao (2022), we can study the convergence of inexact PPG under non-adaptive increasing step sizes, as well as the sample complexity under a generative model. The key is still the advantage lower bound for support actions in Lemma 12, but in an inexact form.

Consider the following inexact PPG:

$$\pi_s^{k+1} = \mathrm{Proj}_\Delta(\pi_s^k + \eta_s^k \hat{Q}_s^k) = \mathrm{Proj}_\Delta(\pi_s^k + \eta_s^k \hat{A}_s^k),$$

where

$$\hat{Q}^k_{s,a} = Q^k_{s,a} + w^k_{s,a}, \quad \hat{A}^k_{s,a} = \hat{Q}^k_{s,a} - V^k_s = A^k_{s,a} + w^k_{s,a}, \quad k = 0, 1, \cdots,$$

with $w^k_{s,a}$ being the estimation error. We will assume $\|w^k\|_\infty \le \omega$. Noticing that the proof of Lemma 12 does not utilize any particular property of $A^\pi_{s,a}$, there still holds

$$\hat{A}^k_{s,a} \ge \max_{\tilde{a}\in\mathcal{A}} \hat{A}^k_{s,\tilde{a}} - \frac{2\pi^k_s(\mathcal{A}\setminus\hat{\mathcal{A}}^k_s)}{\eta^k_s} \ge \max_{\tilde{a}\in\mathcal{A}} \hat{A}^\pi_{s,\tilde{a}} - \frac{2}{\eta^k_s} \ge \max_{\tilde{a}\in\mathcal{A}} \hat{A}^\pi_{s,\tilde{a}} - \frac{2}{\eta^k\tilde{\mu}}, \quad \forall a \in \mathcal{B}_s(\eta^k_s),$$

where $\mathcal{B}_s(\eta^k_s)$ is the support of π^{k+1}_s, $\hat{\mathcal{A}}^k_s = \arg\max_{\tilde{a}} \hat{A}^k_{s,\tilde{a}}$, and the last inequality follows from $\eta^k_s = \eta^k d^k_\mu(s)/(1-\gamma) \ge \eta^k\tilde{\mu}$. From this, a simple algebra yields that

$$\sum_a \pi^{k+1}_{s,a} A^k_{s,a} = \sum_{a\in\mathcal{B}_s(\eta^k_s)} \pi^{k+1}_{s,a}(\hat{A}^k_{s,a} - w^k_{s,a}) \ge \max_{a\in\mathcal{A}} \hat{A}^k_{s,a} - \frac{2}{\eta^k\tilde{\mu}} - \omega \ge \max_{a\in\mathcal{A}} A^k_{s,a} - \frac{2}{\eta^k\tilde{\mu}} - 2\omega.$$

By the performance difference lemma, one has

$$\begin{aligned} V^{k+1}(\mu) - V^k(\mu) &\ge \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{k+1}_\mu}\mathbb{E}_{a\sim\pi^{k+1}_s}\left[\max_{a\in\mathcal{A}} A^k_{s,a} - \frac{2}{\eta^k\tilde{\mu}} - 2\omega\right] \\ &\ge \frac{1}{1-\gamma}(1-\gamma)\tilde{\mu}\sum_s d^*_\mu(s)\max_a A^k_{s,a} - \frac{2}{1-\gamma}\left(\frac{1}{\eta^k\tilde{\mu}} + \omega\right) \\ &\ge (1-\gamma)\tilde{\mu}\left(V^*(\mu) - V^k(\mu)\right) - \frac{2}{1-\gamma}\left(\frac{1}{\eta^k\tilde{\mu}} + \omega\right). \end{aligned}$$

It follows that

$$V^*(\mu) - V^{k+1}(\mu) \le (1-(1-\gamma)\tilde{\mu})(V^*(\mu) - V^k(\mu)) + \frac{2}{1-\gamma}\left(\frac{1}{\eta^k\tilde{\mu}} + \omega\right).$$

Therefore, under the condition

$$\eta^k \ge (1-\tilde{\mu}(1-\gamma))^{-(2k+1)},\ k = 0, \cdots, T-1, \tag{44}$$

it is not hard to obtain the following result:

$$V^*(\mu) - V^T(\mu) \le (1-(1-\gamma)\tilde{\mu})^T\left[(V^*(\mu) - V^0(\mu)) + \frac{2}{\tilde{\mu}^2(1-\gamma)^2}\right] + \frac{2\omega}{\tilde{\mu}(1-\gamma)^2}. \tag{45}$$

Consider the Monte Carlo estimation under the generative model adopted in Xiao (2022). For each iteration $k = 0, \ldots, T-1$ and every state-action pair (s, a), the model generates M independent trajectories, truncated by horizon H,

$$\tau^{(k,i)}_{s,a} = \left\{(s^{(i)}_0, a^{(i)}_0), \ldots, (s^{(i)}_{H-1}, a^{(i)}_{H-1}) \,\middle|\, s^{(i)}_0 = s,\, a^{(i)}_0 = a\right\}_{\pi^k}, \quad i = 1, \cdots, M,$$

where the subscript π^k means that the trajectories are generated by policy π^k. Then the estimated action value is given by

$$\hat{Q}^k_{s,a} = \frac{1}{M}\sum_{i=1}^M \sum_{t=0}^{H-1} \gamma^t r(s^{(i)}_t, a^{(i)}_t).$$

It has been shown in Xiao (2022, Theorem 16) that if

$$M \geq \frac{\gamma^{-2H}}{2} \log\left(\frac{2|\mathcal{S}||\mathcal{A}| \cdot T}{\delta}\right),$$

then with probability at least $1-\delta$,

$$\|\hat{Q}^k - Q^k\|_\infty \leq \frac{2\gamma^H}{1-\gamma} \triangleq \omega, \quad \forall\, k = 0, \ldots, T-1.$$

Plugging this result into (45), it can be directly verified that when

$$H \geq \frac{1}{1-\gamma} \log \frac{8}{\varepsilon(1-\gamma)^3\tilde{\mu}} \geq (\log \gamma^{-1})^{-1} \log \frac{8}{\varepsilon(1-\gamma)^3\tilde{\mu}}, \tag{46}$$

$$T \geq \frac{1}{(1-\gamma)\tilde{\mu}} \log \frac{6}{\varepsilon(1-\gamma)^2\tilde{\mu}^2} \geq \left[\log(1-(1-\gamma)\tilde{\mu})^{-1}\right]^{-1} \log \frac{6}{\varepsilon(1-\gamma)^2\tilde{\mu}^2}, \tag{47}$$

there holds

$$\omega \leq \frac{\varepsilon(1-\gamma)^2\tilde{\mu}}{4} \quad \text{and} \quad V^*(\mu) - V^T(\mu) \leq \varepsilon.$$

Then the total sample number of state-action samples is given by $|\mathcal{S}||\mathcal{A}| \cdot T \cdot M \cdot H \geq \tilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^8\tilde{\mu}^3}\right)$, where a poly-logarithmic factor is hidden in $\tilde{\mathcal{O}}(\cdot)$.

In fact, when the number of samples are sufficiently large, ω can be enough small so that we can also obtain the sample complexity for inexact PPG to achieve the finite iteration convergence under the non-adaptive increasing step sizes specified in (44). To this end, first note that Lemma 19 still holds if we replace $A^\pi_{s,a}$ with the inexact $\hat{A}^\pi_{s,a}$. This implies that if

$$\|V^* - V^{T-1}\|_\infty + \omega \leq \frac{\Delta}{2} \frac{\eta_s^{T-1}\Delta}{1+\eta_s^{T-1}\Delta},$$

then the exact PPG will output an optimal policy at the T-th iteration (cf. condition in (28) for the exact case). Noticing that $\eta^{T-1} \geq 1$ and $\Delta \leq 1/(1-\gamma)$, this condition can be relaxed to

$$\|V^* - V^{T-1}\|_\infty + \omega \leq \frac{(1-\gamma)\tilde{\mu}\Delta^2}{4}. \tag{48}$$

Thus, setting $\varepsilon = \frac{(1-\gamma)\tilde{\mu}^2\Delta^2}{8}$ in (46) and (47), one has

$$\|V^* - V^T\|_\infty \leq \frac{1}{\tilde{\mu}}(V^*(\mu) - V^T(\mu)) \leq \frac{(1-\gamma)\tilde{\mu}\Delta^2}{8} \quad \text{and} \quad \omega \leq \frac{\varepsilon(1-\gamma)^2\tilde{\mu}}{4} \leq \frac{(1-\gamma)\tilde{\mu}\Delta^2}{8}.$$

Therefore, (48) can be satisfied with the probability at least $1-\delta$ under the sample complexity $\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^{10}\Delta^4\tilde{\mu}^7}\right)$.

5.2 Equivalence of PPG to PI under adaptive step sizes

As already mentioned, it is easy to see PPG should converge to a PI update when $\eta_s \to \infty$. In this section, we study the convergence of PPG with adaptive step sizes and identify the non-asymptotic step size threshold beyond which PPG is equivalent to PI. The analysis of this section is similar to that for the finite iteration convergence. We utilize the gap property (Lemma 9) again to show that once the step size is large enough, then the action set $\mathcal{A} \setminus \mathcal{A}_s^k$ will be eliminated from the support set of the new policy.

Theorem 31 *Consider the prototype update in* (12) *and suppose the step size* η *satisfies*

$$\min_{s\in\mathcal{S}} \eta_s > \mathcal{F}^\pi := \frac{2}{\Delta^\pi} \cdot \max_{s\in\mathcal{S}} \left\{ \pi_s \left(\mathcal{A} \setminus \mathcal{A}_s^\pi \right) \right\}, \tag{49}$$

where $\Delta^\pi := \min_{s\in\mathcal{S}} \left| \max_{a'\in\mathcal{A}} A_{s,a'}^\pi - \max_{a'\notin\mathcal{A}_s^\pi} A_{s,a'}^\pi \right|$. *Then the new policy at state* s *(i.e.,* π_s^+*) is supported on the action set* $\mathcal{A}_s^\pi$*, which implies that the prototype update is equivalent to PI.*

Proof Notice that for each state $s \in \mathcal{S}$ and $a \notin \mathcal{A}_s^\pi$, $A_{s,a}^\pi \le \max_{a'\notin\mathcal{A}_s^\pi} A_{s,a'}^\pi$. By Lemma 12, when the step size satisfies

$$\frac{2\pi_s \left(\mathcal{A} \setminus \mathcal{A}_s^\pi \right)}{\eta_s} < \max_{a'\in\mathcal{A}} A_{s,a'}^\pi - \max_{a'\notin\mathcal{A}_s^\pi} A_{s,a'}^\pi,$$

or equivalently

$$\eta_s < \frac{2\pi_s \left(\mathcal{A} \setminus \mathcal{A}_s^\pi \right)}{\max_{a'\in\mathcal{A}} A_{s,a'}^\pi - \max_{a'\notin\mathcal{A}_s^\pi} A_{s,a'}^\pi}, \tag{50}$$

all the $a' \notin \mathcal{A}_s^\pi$ are not in $\mathcal{B}_s(\eta_s)$. That is, the new policy π_s^+ is supported on $\mathcal{A}_s^\pi$. It's trivial to see that the condition (49) implies (50) for every $s \in \mathcal{S}$, thus the proof is completed. ■

The equivalence of PPG to PI follows immediately from this theorem, which is similarly applicable for PQA.

Corollary 32 *If the step size satisfies* $\eta_k \ge \frac{1}{\tilde{\mu}} \mathcal{F}^{\pi^k}$*, PPG is equivalent to PI.*

Corollary 33 *If the step size satisfies* $\eta_k \ge \mathcal{F}^{\pi^k}$*, PQA is equivalent to PI.*

Remark 34 *It is worth noting that the step size threshold in the above two corollaries only relies on the current policy* π^k*.*

Acknowledgments

This work was partially supported by the National Key R&D Program of China (No. 2023YFA1009300), National Natural Science Foundation of China (No. 92370105, No. 12271011 and No. 12350001), and the MOE Project of Key Research Institute of Humanities and Social Sciences (No. 22JJD110001).

References

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Daniel R. Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.

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

Amir Beck. *First-Order Methods in Optimization*. SIAM and MOR, 2017.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Jozefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique-PondedeOliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv:1912.06680*, 2019.

Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

Jalj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

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.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2160–2169, 2019.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning (ICML '02)*, pages 267–274, 2002.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799, 2021.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198 (1):1059–1106, 2021.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

Yan Li, Guanghui Lan, and Tuo Zhao. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *Mathematical Programming*, 2023.

Jiacai Liu, Jinchi Chen, and Ke Wei. On the linear convergence of policy gradient under Hadamard parameterization. *arXiv:2305.19575*, 2023.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6820–6829, 2020.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62), 2022.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nazi, Jiwoo Pak, Andy Tong, Kavya Srinivasa, William Hang, Emre Tuncer, Quoc V. Le, James Laudon, Richard Ho, Roger Carpenter, and Jeff Dean. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

Bruno Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. In *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*, volume 26, 2013.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5668–5675, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, JohnP. Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hadovan Hasselt, David Silver, TimothyP. Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. Starcraft II: A new challenge for reinforcement learning. *arXiv:1708.04782*, 2017.

Weiran Wang and Miguel A. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv:1309.1541*, 2013.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, volume 33, pages 4572–4583, 2020.