A note for: " Elementary Analysis of Policy Gradient Methods"

yifan.xu2021

May 2025

Link to the paper: https://arxiv.org/abs/2404.03372.

Outline of the paper:

- Introduction of RL
- General convergence analysis framework on policy gradient methods
- Convergence analysis on Projected Policy Gradient (PPG) methods
- Convergence analysis on Softmax Policy Gradient (PG) methods
- Convergence analysis on Softmax Natural Policy Gradient (NPG) methods
- Convergence analysis on Entropy Softmax Policy Gradient (PG) methods
- Convergence analysis on Entropy Softmax Natural Policy Gradient (NPG) methods

Outline of my note:

1. Basics of Markov Decision Process (MDP)
2. Basics of Reinforcement Learning (RL)
   (a) Definition of value function
   (b) Bellman equations/operators
   (c) Bellman optimal operators/equations
   (d) Policy existence theorem
3. Preliminaries
   (a) Basics of policy gradient methods
   (b) Policy Gradient Theorem
   (c) Performance difference lemma (*)
4. General analysis framework on convergence of policy gradient methods
   (a) Sublinear convergence
   (b) Linear convergence
5. Convergence of Projected Policy Gradient (PPG) (TBD)

As the results of the paper cover multiple algorithms. I'll try to talk about the first three parts this time (which are the introduction of RL, general convergence analysis for policy gradient methods and the convergence analysis for PPG.) If there're new stuff in the rest parts, I'll save them for the next time, or simply move on to the next paper, depending on the content of the paper.

Note: The proof techniques used in this paper are elementary, as suggested by the title. However, the RL-related notations can be confusing to those who are not familiar with the topic. You can refer to the notation tables at the end of this note.

1 Introduction

Motivation:
This section provides a brief introduction to the basics of Markov Decision Process (MDP) and Reinforcement Learning (RL), for those who are not familiar with the concepts.
Learning Outcome:
After the introduction, one should know what is an MDP, what is the optimization goal for RL, what is the role of (state/action) value function in RL, and the Bellman equations for value functions.

1.1 Markov Decision Process

1.1.1 Definition

$\mathcal{M}(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (We only consider the standard MDP definition in this paper.)

- State space: $\mathcal{S}$ (a finite set of states)
- Action space: $\mathcal{A}$ (a finite set of actions)
- State transition probability: $P(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- Reward function: $R(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Discount factor: $\gamma \in [0, 1)$

Markov property refers to

$$P(s_{t+1} \mid s_t) = P(s_{t+1} \mid s_0, \ldots, s_t) \tag{1.1}$$

"The future is independent of the past given the present."

1.1.2 An example

As shown in Figure 1, the circles represent states, the square is the terminating state (the process terminates when reaching this state), and the black dot is the beginning state. The red texts are the actions (Study, Sleep, Facebook, Pub, Quit) taken by the student, the lines with arrows represent state transition relations. (The initial state is sampled from $\rho(s)$.)

1.2 Reinforcement Learning

1.2.1 Definition

Let $\Delta(\mathcal{A})$ be the probability simplex over the action space $\mathcal{A}$,

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

We can define the set of all admissible policies for MDP as,

$$\Pi := \{\pi = (\pi(\cdot \mid s))_{s \in \mathcal{S}} \mid \pi(\cdot \mid s) \in \Delta(\mathcal{A}) \text{ for all } s \in \mathcal{S}\}$$

Example: Student MDP

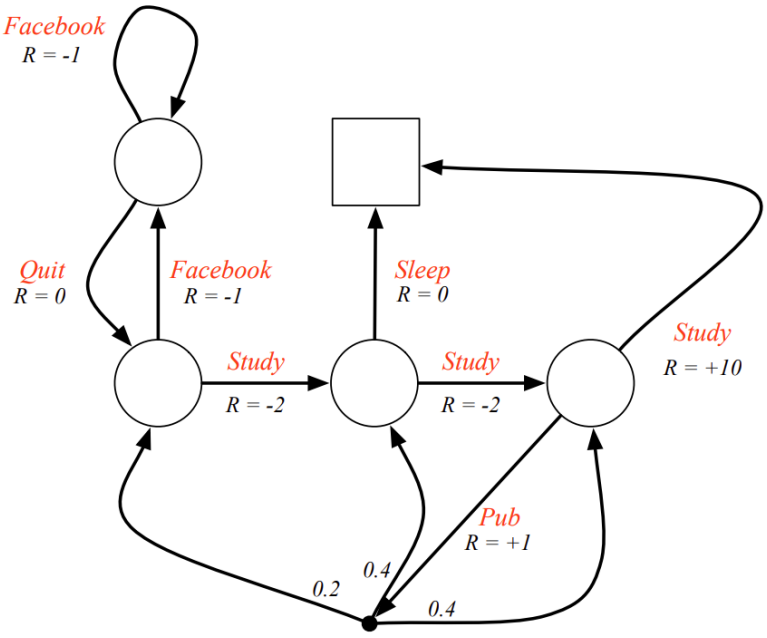


Figure 1: Illustration of MDP (Taken from the lecture slide of David Silver).

Note: The policies are stationary (time-independent), which means,

$$a_t \sim \pi(\cdot \mid s_t), \quad \forall t \geq 0$$

Given a policy $\pi \in \Pi$, one can take actions in the MDP based on the policy and obtain a complete trajectory. For example, at time step 0, we're at the initial state s_0, then we take an action a_0 sampled from $\pi(\mid s_0)$, the MDP then gives the reward based on (s_0, a_0) $r(s_0, a_0)$, the next state then follows $P(\cdot \mid s_0, a_0)$. Therefore, we have a trajectory $\tau = (s_0, a_0, r_0, s_1, \ldots, s_H)$, where H is the horizon (length) of the trajectory (H can be infinite).

we need some metric that quantifies the performance of the specific policy π on the MDP, which is defined as the average discounted cumulative reward over the random trajectory starting from s and induced by π. We have, for the state value function

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, \pi\right]. \tag{1.2}$$

The goal of RL is to seek an optimal policy to maximize state values, which can be expressed as

$$\max_{\pi \in \Pi} V^\pi(\rho) = \mathbb{E}_{s \sim \rho}\left[V^\pi(s)\right],$$

where $\rho \in \Delta(\mathcal{S})$ denotes the initial state distribution.

1.2.2 Bellman Equations

Before we discuss ways to maximize Equation (1.2), we first need to know how to compute $V^\pi(s)$.

We first define the value for taking an action, which is the action value function:

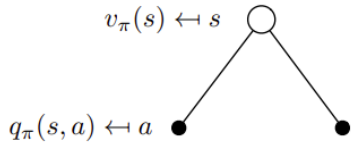
$$Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, \pi\right]. \tag{1.3}$$

Compared with state value function $V^\pi(\cdot)$, $Q^\pi(\cdot, a)$ additionally takes the action a as another input. The rest is the same, i.e., $Q^\pi(s, a)$ is the "value" at state s by taking the action a. Naturally, we can have the "bonus" by taking the action a defined as

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \tag{1.4}$$

The advantage function $A^\pi(s, a)$ quantifies the "advantage" of taking the action a at state s.

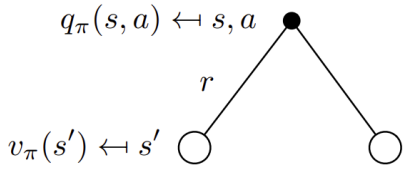
Now let's see what happens at the state s for the MDP:



Apparently, $V^\pi(s)$ can be expressed as taking the expectation of $Q^\pi(s, a)$ w.r.t a,

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^\pi(s, a) = \mathbb{E}_{a \sim \pi(\cdot|s)} Q^\pi(s, a). \tag{1.5}$$

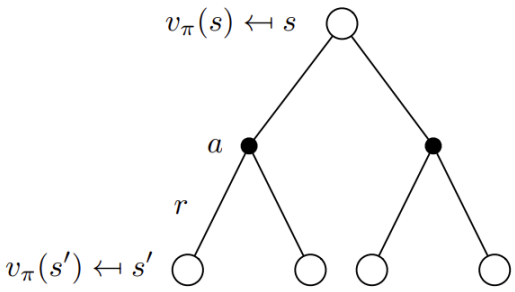
It is similar for $Q^\pi(s, a)$, we can express $Q^\pi(s, a)$ in the form of $V^\pi(s')$,



$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^\pi(s') = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^\pi(s') \right] \tag{1.6}$$

Note that γ is needed here to discount the reward at the subsequent steps.

Taking one step further, we can express $V^\pi(s)$ using $V^\pi(s')$, which is the Bellman equation for the (state) value function.



$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^\pi(s') \right] \right]. \tag{1.7}$$

Again, similar for $Q^\pi(s, a)$, we have

$$Q^\pi(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')} \left[Q^\pi(s', a') \right]. \tag{1.8}$$

Bellman equations (Equation (1.7) and Equation (1.8)) are the foundations for RL algorithms. We can therefore use recursions to compute the value functions and iteratively update our policy.

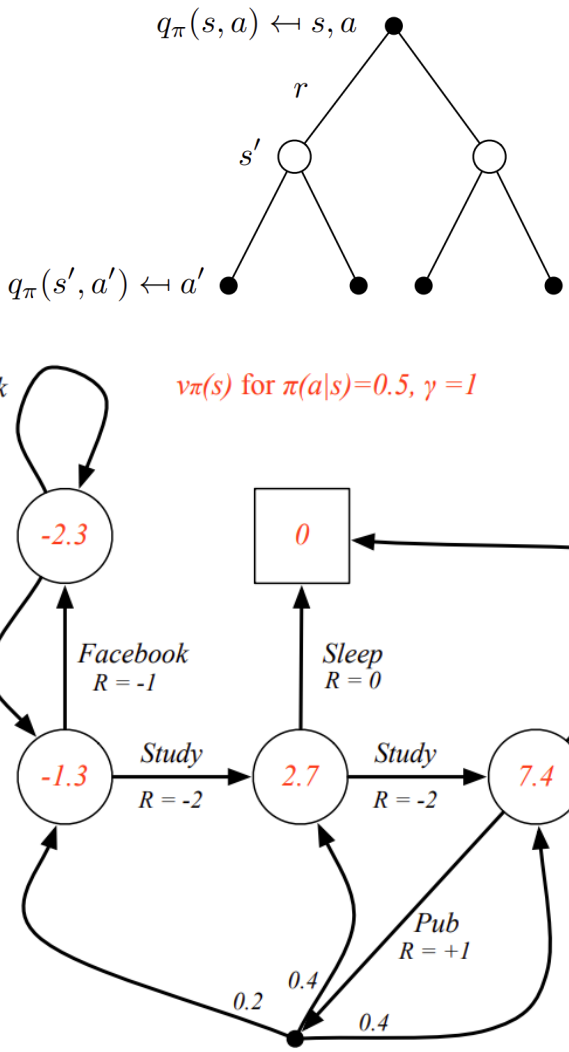


Figure 2: An example showing $V^\pi(\cdot)$ for a policy π

Bellman Operator for Fixed Policy

We've seen that value functions satisfy recursive Bellman equations. These equations can be viewed as mathematical operations that transform value functions. We formalize this as *Bellman operators*.

Given a policy π and for an arbitrary vector $V \in \mathbb{R}^{|\mathcal{S}|}$, define the Bellman operator $\mathcal{T}^\pi$:

$$\mathcal{T}^\pi V(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a) + \gamma V(s')\right]$$

- Properties:
  1. Linearity: $\mathcal{T}^\pi$ is affine in V
  2. Contraction mapping: $\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_\infty \le \gamma \|V_1 - V_2\|_\infty$
  3. Fixed point: $\mathcal{T}^\pi V^\pi = V^\pi$ where V^π is the true value function
- Geometric intuition:

Value function space
$\downarrow$
$v \xrightarrow{\mathcal{T}^\pi}$ new v
$\downarrow$
Contraction toward fixed point v^π

- Action-value version:

$$(\mathcal{T}^\pi q)(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{a' \in A} \pi(a'|s') q(s',a')$$

While $\mathcal{T}^\pi$ computes values for a fixed policy, what if we want to find the optimal policy directly? This requires a new operator that embeds optimization.

Proof. **First we show that the Bellman operator is monotonically increasing with the value function** $V(\cdot)$, which means that given any two value functions $U(\cdot)$ and $V(\cdot)$, if $U(s) \leq V(s), \forall s$, then $\mathcal{T}^\pi U(s) \leq \mathcal{T}^\pi V(s), \forall s$.

Recall the definition of Bellman operator for any policy π:

$$\mathcal{T}^\pi V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(s'|s,a)} [R(s,a) + \gamma V(s')].$$

As $U(s) \leq V(s)$ for all s, then

$$\mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(s'|s,a)} [R(s,a) + \gamma V(s')] \leq \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(s'|s,a)} [R(s,a) + \gamma U(s')],$$

which means that $\mathcal{T}^\pi U(s) \leq \mathcal{T}^\pi V(s)$.

Then we show that the Bellman operator is γ-contraction under the infinity norm, which means that

$$\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

$$\begin{aligned}
|(\mathcal{T}^\pi V_1)(s) - (\mathcal{T}^\pi V_2)(s)| &= \left| \sum_a \pi(a|s) \gamma \sum_{s'} P(s'|s,a) \left[V_1(s') - V_2(s') \right] \right| \\
&\leq \sum_a \pi(a|s) \gamma \sum_{s'} P(s'|s,a) \left| V_1(s') - V_2(s') \right| \\
&\leq \gamma \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) \|V_1 - V_2\|_\infty \\
&= \gamma \|V_1 - V_2\|_\infty.
\end{aligned}$$

Last, we show that V^π is the only fixed point for $\mathcal{T}^\pi$. As $\mathcal{T}^\pi$ is a contraction mapping, by applying the Banach fixed-point theorem, we know that there exists the only fixed point for $\mathcal{T}^\pi$, we list the Banach fixed-point theorem here for reference.

Theorem 1 (Banach Fixed-Point Theorem)**.** *Let (X, d) be a* ***complete metric space****, and let $T \colon X \to X$ be a* ***contraction mapping*** *on X, i.e., there exists a constant $0 \leq \gamma < 1$ such that for all $x, y \in X$,*

$$d(Tx, Ty) \leq \gamma \, d(x, y).$$

Then:

1. *(**Existence and Uniqueness**) T has a* ***unique fixed point*** *$x^* \in X$, i.e., $Tx^* = x^*$.*
2. *(**Convergence**) For any initial $x_0 \in X$, the sequence (x_n) defined by $x_{n+1} = Tx_n$ converges to x^*.*
3. *(**Error Bound**) The convergence rate satisfies $d(x_n, x^*) \leq \frac{\gamma^n}{1-\gamma} d(x_0, x_1)$.*

Then it suffices to show that V^π is the fixed point for $\mathcal{T}^\pi$.

$$(\mathcal{T}^\pi V^\pi)(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s \right] = V^\pi(s) \quad \text{(by the Bellman equation for } V^\pi\text{)}.$$

□

Bellman Optimal Operator

Define the Bellman optimal operator $\mathcal{T}$:

$$\mathcal{T} V(s) := \max_{\pi \in \Pi} \mathcal{T}^\pi V(s) = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a) + \gamma V(s') \right] \right\}$$

Key Properties

1. Nonlinearity: Due to max operator
2. Contraction: $\|\mathcal{T}^* v_1 - \mathcal{T}^* v_2\|_\infty \leq \gamma \|v_1 - v_2\|_\infty$
3. Fixed point: $\mathcal{T}^* v^* = v^*$ (Bellman optimal equation)

Bellman Optimality Equations

$$\begin{aligned}
\text{State-value:} \quad & v^*(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v^*(s') \right] \\
\text{Action-value:} \quad & q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \max_{a' \in A} q^*(s',a') \\
\text{Relationship:} \quad & v^*(s) = \max_{a \in A} q^*(s,a)
\end{aligned}$$

Optimal Policy Existence Theorem

With these tools, we can now answer fundamental questions: Does an optimal policy exist? Is it unique? How do we characterize it?

Theorem 2 (Policy Existence Theorem)**.** *Consider any finite Markov Decision Process $\mathcal{M} = (S, A, P, R, \gamma)$ with discount factor $\gamma < 1$. Then the following hold:*

(Existence of Optimal Policy) *There exists a policy π^* that dominates all other policies:*

$$\forall \pi, \forall s \in S: \quad V^{\pi^*}(s) \geq V^{\pi}(s)$$

(Uniqueness of Value Functions) *For any two optimal policies π_1^* and π_2^*, they share identical value functions:*

$$\begin{aligned}
V^{\pi_1^*}(s) &= V^{\pi_2^*}(s) \equiv V^*(s) \\
Q^{\pi_1^*}(s,a) &= Q^{\pi_2^*}(s,a) \equiv Q^*(s,a)
\end{aligned}$$

for all states $s \in S$ and actions $a \in A$.

(Deterministic Implementation) *There exists a deterministic optimal policy π^* given by:*

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s,a) \quad \textit{for all } s \in S$$

with the convention that ties are broken arbitrarily when multiple actions maximize the value.

Proof. **Part 1: Existence and uniqueness of V^***

Define the Bellman optimality operator:

$$(\mathcal{T}^* V)(s) \coloneqq \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V(s') \right]$$

Lemma 1.1. *$\mathcal{T}^*$ is a γ-contraction in $\|\cdot\|_\infty$ norm:*

$$\|\mathcal{T}^* V - \mathcal{T}^* U\|_\infty \leq \gamma \|V - U\|_\infty \quad \textit{for all } V, U : S \to \mathbb{R}$$

Proof. For any state $s \in S$:

$$\begin{aligned}
&|(\mathcal{T}^* V)(s) - (\mathcal{T}^* U)(s)| \\
&= \left| \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right] - \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) U(s') \right] \right| \\
&\leq \max_a \left| \gamma \sum_{s'} P(s'|s,a)(V(s') - U(s')) \right| \\
&\leq \gamma \max_a \sum_{s'} P(s'|s,a) |V(s') - U(s')| \\
&\leq \gamma \|V - U\|_\infty
\end{aligned}$$

Taking supremum over s gives the result. □

Since $(\mathbb{R}^{|S|}, \|\cdot\|_\infty)$ is a complete metric space and $0 \leq \gamma < 1$, by the Banach Fixed-Point Theorem, there exists a unique $V^* : S \to \mathbb{R}$ such that:

$$\mathcal{T}^* V^* = V^*$$

Part 2: Construction of optimal policy

Define the deterministic policy $\pi^* : S \to A$ by:

$$\pi^*(s) \coloneqq \arg\max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right]$$

where ties are broken arbitrarily when multiple actions achieve the maximum.

Claim: $V^{\pi^*} = V^*$

Proof of Claim. The value function of π^* satisfies the Bellman equation:

$$\begin{aligned}
V^{\pi^*}(s) &= \mathbb{E}_{\pi^*} \left[R_t + \gamma V^{\pi^*}(S_{t+1}) \mid S_t = s \right] \\
&= R\left(s, \pi^*(s)\right) + \gamma \sum_{s' \in S} P\left(s'|s, \pi^*(s)\right) V^{\pi^*}(s') \quad \text{(by MDP dynamics)} \\
&= \max_{a \in A} \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right] \quad \text{(by definition of } \pi^*\text{)} \\
&= (\mathcal{T}^* V^*)(s) = V^*(s) \quad \text{(fixed point property)}
\end{aligned}$$

□

Part 3: Optimality

For any policy π and state $s \in S$, consider the k-step truncated value:

$$V_k^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t R_{t+1} \mid S_0 = s \right]$$

By induction on k:

$$\begin{aligned}
V_1^\pi(s) &= \mathbb{E}_\pi[R_1 | S_0 = s] \leq \max_a R(s,a) \leq (\mathcal{T}^* \mathbf{0})(s) \\
V_{k+1}^\pi(s) &= \mathbb{E}_\pi[R_1 + \gamma V_k^\pi(S_1) | S_0 = s] \\
&\leq \mathbb{E}_\pi \left[R_1 + \gamma (\mathcal{T}^*)^k \mathbf{0}(S_1) \mid S_0 = s \right] \quad \text{(by induction hypothesis)} \\
&\leq \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) (\mathcal{T}^*)^k \mathbf{0}(s') \right] \\
&= (\mathcal{T}^*)^{k+1} \mathbf{0}(s)
\end{aligned}$$

Taking limits as $k \to \infty$:

$$V^\pi(s) = \lim_{k \to \infty} V_k^\pi(s) \leq \lim_{k \to \infty} (\mathcal{T}^*)^k \mathbf{0}(s) = V^*(s)$$

where the last equality follows from the fixed-point convergence.

For π^*, we have $V^{\pi^*}(s) = V^*(s)$, so:

$$V^{\pi^*}(s) \geq V^\pi(s) \quad \text{for all } \pi \text{ and } s \in S$$

□

Note: Because Bellman optimal equation is non-linear, and in general has no closed-form solutions, so we generally use iteration methods to solve the MDP.

2 General Analysis Framework

2.1 Preliminaries

Therefore, this paper mainly focuses on the convergence analysis of a class of methods in RL, which is the policy gradient methods. Given a parameterized policy class $\{\pi_\theta : \mathbb{R}^d \to \Pi\}$, the state value function $V^\pi(\cdot)$ becomes a function of parameters, denoted as $V^{\pi_\theta}(\mu)$. **Thus, the problem of seeking an optimal policy over the parameterized policy class turns to be a finite dimensional optimization problem**. Among all the optimization methods, our focus in this paper is on the policy gradient method,

$$\theta^+ \leftarrow \theta + \eta \nabla_\theta V^{\pi_\theta}(\mu), \tag{2.1}$$

where $\nabla_\theta V^{\pi_\theta}(\mu)$ is the policy gradient of $V^{\pi_\theta}(\mu)$. Different parameterization methods lead to different policy gradient methods. This paper mainly discusses the convergence analysis of the following policy gradient methods:

- Projected policy gradient (PPG)
- Softmax Policy gradient (Softmax PG)
- Natural policy gradient (NPG)

Theorem 3 (Policy Gradient Theorem)**.** *(Informal)(TBD) For a differentiable policy $\pi_\theta(a|s)$, the gradient of the expected return $J(\theta)$ is:*

$$\nabla_\theta J(\theta) = \nabla_\theta V^{\pi_\theta}(\rho) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \ln \pi_\theta(a_t|s_t)\right].$$

Proof. To simplify notation, we omit θ in π_θ and V^{π_θ}. We start with the derivation of the state value function:

$$\begin{aligned}
\nabla_\theta V^\pi(s) &\overset{(a)}{=} \nabla_\theta \left(\sum_{a\in\mathcal{A}} \pi(a \mid s) Q^\pi(s,a)\right) \\
&= \sum_{a\in\mathcal{A}} \left(\nabla_\theta \pi(a \mid s) Q^\pi(s,a) + \pi(a \mid s) \nabla_\theta Q^\pi(s,a)\right) \\
&\overset{(b)}{=} \sum_{a\in\mathcal{A}} \left(\nabla_\theta \pi(a \mid s) Q^\pi(s,a) + \pi(a \mid s) \nabla_\theta \left(R(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a) V^\pi(s')\right)\right) \\
&\overset{(c)}{=} \sum_{a\in\mathcal{A}} \left(\nabla_\theta \pi(a \mid s) Q^\pi(s,a) + \pi(a \mid s) \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \nabla_\theta V^\pi(s')\right).
\end{aligned}$$

Let $\phi(s) = \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a \mid s) Q^\pi(s,a)$, we can rewrite the above equation as:

$$\nabla_\theta V^\pi(s) = \phi(s) + \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \nabla_\theta V^\pi(s'). \tag{2.2}$$

An interesting observation is that this equation has a recursive form (thanks to the Bellman equation). Therefore, we can continue unrolling the future states for s'.

$$
\begin{aligned}
\nabla_\theta V^\pi(s) &= \phi(s) + \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi(a \mid s) P(s' \mid s,a) \nabla_\theta V^\pi(s') \\
&= \phi(s) + \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi(a \mid s) P(s' \mid s,a) \left(\phi(s') + \sum_{s''\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \pi(a' \mid s') P(s'' \mid s',a') \nabla_\theta V^\pi(s'') \right) \\
&= \sum \sum \Pr(s_t \mid s) \phi(s_t)
\end{aligned}
\tag{2.3}
$$

□

Lemma 2.1 (Performance Difference Lemma)**.** *Given two policies π_1, π_2, there holds,*

$$
\begin{aligned}
V^{\pi_1}(\rho) - V^{\pi_2}(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi_1}} \left[\mathcal{T}^{\pi_1} V^{\pi_2}(s) - V^{\pi_2}(s) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi_1}} \mathbb{E}_{a\sim\pi_1(\cdot|s)} \left[A^{\pi_2}(s,a) \right].
\end{aligned}
\tag{2.4}
$$

Note: This equation is core in RL, which quantifies the performance gap between two policies using the advantage function of the reference policy (π_2) with actions and states sampled by the updated policy (π_1).

Proof. Recall the Bellman equations for state value function and action value function are:

$$
V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} \left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\pi(s') \right]. \tag{2.5}
$$

$$
Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \mathbb{E}_{a'\sim\pi(\cdot|s')} \left[Q^\pi(s',a') \right]. \tag{2.6}
$$

$$
V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} \left[r(s,a) + \gamma Q^\pi(s,a) \right]. \tag{2.7}
$$

$$
Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} [V^\pi(s')]. \tag{2.8}
$$

The definition for the advantage function $A^\pi(s,a)$ is

$$
A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{2.9}
$$

We first prove that $\mathcal{T}^{\pi_1} V^{\pi_2}(s) - V^{\pi_2}(s) = \mathbb{E}_{a\sim\pi_1(\cdot|s)}[A^{\pi_2}(s,a)]$.

$$
\begin{aligned}
\mathcal{T}^{\pi_1} V^{\pi_2}(s) - V^{\pi_2}(s) &= \mathbb{E}_{a\sim\pi_1(\cdot|s)} \left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} [V^{\pi_2}(s')] \right] - V^{\pi_2}(s) \\
&\overset{(a)}{=} \mathbb{E}_{a\sim\pi_1(\cdot|s)} \left[Q^{\pi_2}(s,a) - V^{\pi_2}(s) \right] \\
&\overset{(b)}{=} \mathbb{E}_{a\sim\pi_1(\cdot|s)} \left[A^{\pi_2}(s,a) \right],
\end{aligned}
$$

where (a) holds because $V^{\pi_2}(s)$ is not related to the action distribution π and can be taken into the expectation, and further apply Equation (2.8); (b) holds because of Equation (2.9).

Note: This formula quantifies how much performance gap would happen by taking one step update of π_2 using actions taken in π_1, in the form of the advantage function for π_2.

We then prove the following equation:

$$
\mathbb{E}_{\tau\sim\Pr^\pi} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^\pi} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[f(s,a) \right], \tag{2.10}
$$

where $|f(s_t, a_t)| < \infty$, τ is the trajectory and $\Pr^\pi$ indicates the trajectory distribution following π. The visitation measure $d^\pi_{s_0}$ is defined as:

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid s_0, \pi). \tag{2.11}$$

Note: This equation shows that we can transform the expectation over trajectories into the expectation over the state space, by utilizing the visitation measure (long-term state visiting distribution) defined on the state space. This is helpful in the subsequent analysis of algorithm convergence.

We first explain the normalization term $1-\gamma$. Apparently, $d^\pi_{s_0}$ is a distribution over the state space $\mathcal{S}$, and $\Pr(\cdot \mid s_0, \pi)$ at time step t is also a distribution over the state space $\mathcal{S}$. Therefore, we have

$$\sum_s d^\pi_{s_0}(s) = 1$$

$$\sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid s_0, \pi) = \sum_{t=0}^{\infty}\gamma^t \sum_{s\in\mathcal{S}} \Pr(s_t = s \mid s_0, \pi) = \frac{1}{1-\gamma}.$$

This again highlights the role of discount factor $\gamma \in [0, 1)$ in RL, which controls the convergence of series, like in the value function.

To prove Equation (2.10), we first notice that in Equation (2.11), the RHS is about the expectation over time, and is not related to action. Therefore, we first expand the LHS of Equation (2.10).

$$\begin{aligned}
\mathbb{E}_{\tau\sim\Pr^\pi}\left[\sum_{t=0}^{\infty}\gamma^t f\left(s_t, a_t\right)\right] &= \mathbb{E}_{s_t\sim\Pr(\cdot|s_0,\pi)}\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}\left[\sum_{t=0}^{\infty}\gamma^t f\left(s_t, a_t\right)\right] \\
&= \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s_t\sim\Pr(\cdot|s_0,\pi)}\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}\left[f\left(s_t, a_t\right)\right] \\
&= \sum_{t=0}^{\infty}\gamma^t \sum_{s\in\mathcal{S}}\Pr(s_t = s \mid s_0, \pi)\sum_{a\in\mathcal{A}}\pi(a \mid s_t = s) f(s, a) \\
&\overset{(a)}{=} \sum_{t=0}^{\infty}\gamma^t \sum_{s\in\mathcal{S}}\Pr(s_t = s \mid s_0, \pi) g(s) \\
&= \sum_{s\in\mathcal{S}} g(s)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid s_0, \pi) \\
&\overset{(b)}{=} \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} g(s) d^\pi_{s_0}(s) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_{s_0}}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[f(s, a)\right],
\end{aligned}$$

where (a) uses $g(s)$ to replace the expectation term $\sum_{a\in\mathcal{A}}\pi(a \mid s_t = s) f(s, a)$; (b) holds because the definition for visitation measure defined in Equation (2.11).

Note: During the derivation, we transform expectation over trajectory probability into expectation over state, and eliminate the time factor by using visitation measure.

We then expand the value difference on a single state s,

$$
\begin{aligned}
V^{\pi_1}(s) - V^{\pi_2}(s) &= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] - V^{\pi_2}(s) \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) + V^{\pi_2}(s_t) - V^{\pi_2}(s_t)\right)\right] - V^{\pi_2}(s) \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) + V^{\pi_2}(s_t) - V^{\pi_2}(s_t)\right) - \gamma^0 V^{\pi_2}(s_0)\right] \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) - V^{\pi_2}(s_t)\right) + \sum_{t=0}^{\infty}\left(\gamma^t V^{\pi_2}(s_t)\right) - \gamma^0 V^{\pi_2}(s_0)\right] \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) - V^{\pi_2}(s_t)\right) + \sum_{t=1}^{\infty}\gamma^t V^{\pi_2}(s_t)\right] \\
&\overset{(a)}{=} \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) - V^{\pi_2}(s_t)\right) + \sum_{t=0}^{\infty}\gamma^{t+1} V^{\pi_2}(s_{t+1})\right] \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) + \gamma V^{\pi_2}(s_{t+1}) - V^{\pi_2}(s_t)\right)\right] \\
&\overset{(b)}{=} \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t,a_t) + \gamma \mathbb{E}[V^{\pi_2}(s_{t+1})|s_t,a_t] - V^{\pi_2}(s_t)\right)\right] \\
&\overset{(c)}{=} \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t \left(Q^{\pi_2}(s_t,a_t) - V^{\pi_2}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau\sim \Pr^{\pi_1}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_2}(s_t,a_t)\right] \\
&\overset{(d)}{=} \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}A^{\pi_2}(s',a),
\end{aligned}
$$

where (a) holds because of the index shift invariance of convergent series, (b) holds because the law of iterated expectations ($\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$); (c) is based on Equation (2.8), and (d) holds because of Equation (2.10).

□

Note: Two techniques worth noting in the above derivation.

1. Construct a telescoping series by adding and substracting $V^{\pi_2}(s_t)$;
2. Use of the law of iterated expectation to construct Q function.

Lemma 2.2. *Assume $r(s,a) \in [0,1]$. For any policy π, one has*

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

Proof. We first expand the value function.

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t) \mid s_0=s,\pi\right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\right] \\ &\leq \sum_{t=0}^{\infty}\gamma^t \\ &= \frac{1}{1-\gamma}, \end{aligned}$$

where (a) holds because $r(s,a)\in[0,1]$. In practice, $R(\cdot)$ is also a bounded function.

Then for the action value function

$$\begin{aligned} Q^{\pi}(s,a) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t) \mid s_0=s,a_0=a,\pi\right] \\ &\leq \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\right] \\ &\leq \sum_{t=0}^{\infty}\gamma^t \\ &= \frac{1}{1-\gamma}. \end{aligned}$$

The proof is similar to the state value function.

Then for the advantage function, we know that $A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s)$. As both $V^{\pi}(s)$ and $Q^{\pi}(s)$ are bounded, we can get the desired result.

□

Setting. We consider only optimization on a fixed initial state distribution $\mu\in\Delta(\mathcal{S})$ with $\tilde{\mu}=\min_s\mu(s)>0$, which means that the support of μ is the entire state space $\mathcal{S}$. This assumption allows us to analyze value error $V^*(\rho)-V^k(\rho)$ over a fixed distribution μ instead of all the distributions defined on $\mathcal{S}$. To justify this assumption, we observe that

$$\begin{aligned} V^*(\rho)-V^k(\rho) &= \sum_s \rho(s)\left(V^*(s)-V^k(s)\right) \\ &= \sum_s \frac{\rho(s)}{\mu(s)}\mu(s)\left(V^*(s)-V^k(s)\right) \\ &\leq \left\|\frac{\rho}{\mu}\right\|_{\infty}\left(V^*(\mu)-V^k(\mu)\right). \end{aligned}$$

As $V^*(\rho)-V^k(\rho)$ can be bounded by $V^*(\mu)-V^k(\mu)$, we can simply study the properties of $V^*(\mu)-V^k(\mu)$.

> **Note:** However, we observe that the paper still uses notation ρ in the subsequent derivations. Maybe an abuse of notation. I inherit the same notation ρ as used in the paper, so just assume ρ is a fixed distribution on $\mathcal{S}$ with support being the entire $\mathcal{S}$.

The subsequent analysis centers around the following two quantities:

$$\begin{aligned} \mathcal{L}_k^{k+1} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\rho}^{\pi^{k,*}}}\left[\mathcal{T}^{k+1}V^k(s)-V^k(s)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\rho}^{\pi^{k,*}}}\left[\sum_{a\in\mathcal{A}}\pi^{k+1}(a\mid s)A^k(s,a)\right], \\ \mathcal{L}_k^{*} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\rho}^{\pi^{k,*}}}\left[\mathcal{T}^{*}V^k(s)-V^k(s)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\rho}^{\pi^{k,*}}}\left[\sum_{a\in\mathcal{A}}\pi^{k,*}(a\mid s)A^k(s,a)\right], \end{aligned} \tag{2.12}$$

where $\pi^{k,*}$ is the optimal policy chosen at k-th iteration, and $\mathcal{T}^*$ is the Bellman operator associated with $\pi^{k,*}$.

Based on the performance difference lemma, we can see that

$$\mathcal{L}_k^* = V^*(\rho) - V^k(\rho).$$

Note: $\mathcal{L}_k^*$ is irrelevant to the choice of the optimal policy $\pi^{k,*}$, because $V^*(\rho)$ is related to the MDP itself. Although there might exist multiple optimal policies that achieve the same optimal value, chosing which optimal policy does not affect the value of $V^*(\rho)$.

2.2 Sublinear Convergence Analysis

Note: Convergence Rate
Let $\{\boldsymbol{x}_k\}$ be a sequence in $\mathbb{R}^n$, and let $\boldsymbol{x}^* \in \mathbb{R}^n$. Then we say

- $\{\boldsymbol{x}_k\}$ converge to $\boldsymbol{x}^*$ if $\lim \|\boldsymbol{x}_k - \boldsymbol{x}^*\| = 0$
- $\{\boldsymbol{x}_k\}$ q-linearly converge to x^* if there exists $c \in [0, 1)$ such that

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = c$$

- $\{\boldsymbol{x}_k\}$ q-sublinearly converge to $\boldsymbol{x}^*$ if

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = 1$$

- $\{\boldsymbol{x}_k\}$ q-superlinearly converge to $\boldsymbol{x}^*$ if there exists $c_k \to 0$ such that

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = 0$$

- $\{\boldsymbol{x}_k\}$ converge to $\boldsymbol{x}^*$ with order p if there exists constants $p > 1, c \geq 0$ such that

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^q} = c$$

Theorem 2.3 (Sublinear Convergence)**.** *Let $\{\pi^k\}_{k\geq 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1}V^k \geq V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s)A^k(s,a) \geq 0, \forall s$). Suppose there exists a sequence of optimal policies $\{\pi^{k,*}\}_{k\geq 0}$ and a positive constant sequence $\{C_k\}_{k\geq 0}$ such that*

$$\mathcal{L}_k^{k+1} \geq C_k(\mathcal{L}_k^*)^2,$$

where both $\mathcal{L}_k^{k+1}$ and $\mathcal{L}_k^*$ are defined w.r.t $d_\rho^{\pi^{k,*}}$. Then

$$V^*(\rho) - V^k(\rho) \leq \frac{1}{k}\frac{1}{C(1-\gamma)\vartheta},$$

where $C = \inf_k C_k$ and $\vartheta = \inf_k \left\| \frac{d_\rho^{\pi^{k,*}}}{\rho} \right\|_\infty \geq \tilde{\rho}$.

Proof. As $\mathcal{L}_k^* = V^*(\rho) - V^k(\rho)$ is independent of $d_\rho^{\pi^{k,*}}$. We have

$$
\begin{aligned}
\mathcal{L}_k^* - \mathcal{L}_{k+1}^* &= V^{k+1}(\rho) - V^k(\rho) \\
&\overset{(a)}{=} \frac{1}{1-\gamma} \sum_s d_\rho^{k+1}(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{k+1}(s)}{d_\rho^{\pi^{k,*}}(s)} d_\rho^{\pi^{k,*}}(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&\overset{(b)}{\geq} \sum_s \frac{\rho(s)}{d_\rho^{\pi^{k,*}}(s)} d_\rho^{\pi^{k,*}}(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&\geq \left\| \frac{d_\rho^{\pi^{k,*}}}{\rho} \right\|_\infty^{-1} \sum_s d_\rho^{\pi^{k,*}}(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&\overset{(c)}{\geq} (1-\gamma) \left\| \frac{d_\rho^{\pi^{k,*}}}{\rho} \right\|_\infty^{-1} \mathcal{L}_k^{k+1} \\
&\geq (1-\gamma) \left\| \frac{d_\rho^{\pi^{k,*}}}{\rho} \right\|_\infty^{-1} C_k (\mathcal{L}_k^*)^2 \\
&\geq C(1-\gamma)\vartheta (\mathcal{L}_k^*)^2,
\end{aligned}
$$

where (a) is the direct application of the performance difference lemma, (b) holds because $\sum_a \pi^{k+1}(a \mid s) A^k(s,a) \geq 0$, (c) is the definition of $\mathcal{L}_k^{k+1}$.

As $\mathcal{L}_k^* \geq \mathcal{L}_{k+1}^*$,

$$
\frac{1}{\mathcal{L}_{k+1}^*} - \frac{1}{\mathcal{L}_k^*} = \frac{\mathcal{L}_k^* - \mathcal{L}_{k+1}^*}{\mathcal{L}_k^* \mathcal{L}_{k+1}^*} \geq \frac{\mathcal{L}_k^* - \mathcal{L}_{k+1}^*}{(\mathcal{L}_k^*)^2} \geq C(1-\gamma)\vartheta.
$$

It follows that

$$
\frac{1}{\mathcal{L}_k^*} \geq \sum_{j=1}^{k-1} \left(\frac{1}{\mathcal{L}_{j+1}^*} - \frac{1}{\mathcal{L}_j^*} \right) \geq C(1-\gamma)\vartheta k,
$$

which means that

$$
V^*(\rho) - V^k(\rho) \leq \frac{1}{k} \frac{1}{C(1-\gamma)\vartheta}.
$$

□

Theorem 2.4 (Sublinear Convergence by Controlling Error Terms)**.** *Let $\{\pi^k\}_{k \geq 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1} V^k \geq V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s) A^k(s,a) \geq 0, \forall s$). Assume that*

$$
\mathcal{L}_k^{k+1} \geq C \mathcal{L}_k^* - \epsilon_k, \quad \text{where } C > 0 \text{ and } \sum_{k=0}^{\infty} \epsilon_k < \infty.
$$

Then

$$
V^*(\rho) - V^k(\rho) \leq \frac{1}{kC} \left(\frac{1}{(1-\gamma)^2} + \sum_{t=0}^{k-1} \epsilon_t \right).
$$

Lemma 1 (Policy Improvement Lemma)**.** *(Informal) Let $\{\pi^k\}_{k \geq 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1} V^k \geq V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s) A^k(s,a) \geq 0, \forall s$). Then*

$$
V^{k+1}(s) \geq V^k(s), \forall s.
$$

Proof. The proof mainly utilizes the properties of the Bellman operator.

By iteratively apply this property, we have

$$
(\mathcal{T}^{k+1})^n V^k \geq V^k
$$

Then we want to show that $\lim_{n\to\infty}(\mathcal{T}^{k+1})^n V^k = V^{k+1}$, to do this, we further use the **γ-contraction property of the Bellman operator**, which is

$$\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_\infty \le \gamma \|V_1 - V_2\|_\infty.$$

As we already know, V^{k+1} is the fixed point for $\mathcal{T}^{k+1}$. Let V_n denote $(\mathcal{T}^{k+1})^n V^k$, we have

$$\begin{aligned}\|V_n - V^{k+1}\|_\infty &= \|\mathcal{T}^{k+1}V_{n-1} - \mathcal{T}^{k+1}V^{k+1}\|_\infty \\ &\le \gamma \|V_{n-1} - V^{k+1}\|_\infty \\ &\le \gamma^n \|V_0 - V^{k+1}\|_\infty\end{aligned}$$

As $\gamma \in [0,1)$, and $\|V_0 - V^{k+1}\|$ is bounded, $\lim_{n\to\infty}\|V_n - V^{k+1}\|_\infty = 0$, which completes the proof. □

Note: *This lemma indicates that* ***as long as the one-step policy improvement is non-negative, the value function sequence would be non-decreasing.***

Proof of Theorem 2.4. We first observe that

$$d^{k+1}_{d^*_\rho}(s) = (1-\gamma)\sum_{t=0}^{\infty}\Pr(s_t = s \mid d^*_\rho, \pi) \ge (1-\gamma) d^*_\rho(s),$$

the inequality holds by discarding terms in the summation with $t \ge 1$.

One has that

$$\begin{aligned}\mathcal{L}^{k+1}_k &= \frac{1}{1-\gamma}\sum_s d^*_\rho(s)(\mathcal{T}^{k+1}V^k(s) - V^k(s)) \\ &\le \frac{1}{(1-\gamma)^2}\sum_s d^{k+1}_{d^*_\rho}(s)(\mathcal{T}^{k+1}V^k(s) - V^k(s)) \\ &\overset{(a)}{=} \frac{1}{1-\gamma}(V^{k+1}(d^*_\rho) - V^k(d^*_\rho)),\end{aligned}$$

where (a) is a direct application of the performance difference lemma. As in Lemma 1, $V^{k+1} \ge V^k$, which means that $\mathcal{L}^*_{k+1} \le \mathcal{L}^*_k$, the sequence $\{\mathcal{L}^*_k\}_{k\ge 0}$ is non-increasing. So we have

$$\begin{aligned}\mathcal{L}^*_k &\le \frac{1}{k}\sum_{t=0}^{k-1}\mathcal{L}^*_t \\ &\le \frac{1}{kC}\sum_{t=0}^{k-1}(\mathcal{L}^{t+1}_k + \epsilon_t) \\ &\le \frac{1}{kC}\left(\frac{1}{1-\gamma}(V^k(d^*_\rho) - V^0(d^*_\rho)) + \sum_{t=0}^{k-1}\epsilon_t\right) \\ &\le \frac{1}{kC}\left(\frac{1}{(1-\gamma)^2} + \sum_{t=0}^{k-1}\epsilon\right).\end{aligned} \tag{2.13}$$

□

Theorem 2.5 (Sublinear Lower Bound)**.** *Let $\{\pi^k\}_{k\ge 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1}V^k \ge V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s)A^k(a \mid s) \ge 0, \forall s$). Assume that*

$$0 \le \mathcal{L}^{k+1}_k \le C(\mathcal{L}^*_k)^2.$$

Then for any fixed $\alpha \in (0,1)$, there exists a time $T(\sigma)$ such that

$$\forall k \ge T(\sigma): \quad V^*(\rho) - V^k(\rho) \ge \frac{1}{k}\frac{1-\sigma}{2C}\left\|\frac{1}{d^*_\rho}\right\|^{-1}.$$

Proof. Similar to Theorem 2.3, we first note that

$$\begin{aligned}
\mathcal{L}_k^* - \mathcal{L}_{k+1}^* &= V^{k+1}(\rho) - V^k(\rho) \\
&\overset{(a)}{=} \frac{1}{1-\gamma} \sum_s d_\rho^{k+1}(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{k+1}(s)}{d_\rho^*(s)} d_\rho^*(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&\leq \max_s \frac{d_\rho^{k+1}(s)}{d_\rho^*(s)} \cdot \frac{1}{1-\gamma} \sum_s d_\rho^*(s) \sum_a \pi^{k+1}(a \mid s) A^k(s,a) \\
&= \left\| \frac{1}{d_\rho^*} \right\|_\infty \mathcal{L}_k^{k+1} \\
&\leq C \left\| \frac{1}{d_\rho^*} \right\|_\infty (\mathcal{L}_k^*)^2.
\end{aligned}$$

As we have proven in **??**, the sequence $\{V^k(\rho)\}$ is non-decreasing, so $\lim_{k\to\infty} V^k(\rho)$ exists. If $\lim_{k\to\infty} V^k(\rho) < V^*(\rho)$, then we can always find a $T(\sigma)$ big enough for all $\sigma \in (0,1)$ such that the result holds. Then we consider the case where $\lim_{k\to\infty} V^k(\rho) = V^*(\rho)$. Under this condition, there exists a $T_1(\sigma)$ such that $\mathcal{L}_k^* \leq \frac{\sigma}{C} \left\| \frac{1}{d_\rho^*} \right\|^{-1}$ for $k \geq T_1(\sigma)$ and

$$\mathcal{L}_{k+1}^* \geq \mathcal{L}_k^* - C \left\| \frac{1}{d_\rho^*} \right\|_\infty (\mathcal{L}_k^*)^2 \geq (1-\sigma)\mathcal{L}_k^*.$$

Substituting this result into the above inequality gives

$$\frac{1}{\mathcal{L}_{k+1}^*} - \frac{1}{\mathcal{L}_k^*} \leq C \left\| \frac{1}{d_\rho^*} \right\|_\infty (1-\sigma)^{-1}.$$

Consequently, for $k \geq T_1(\sigma)$,

$$\begin{aligned}
\frac{1}{\mathcal{L}_k^*} &= \sum_{t=T_1(\sigma)}^{k-1} \left(\frac{1}{\mathcal{L}_{t+1}^*} - \frac{1}{\mathcal{L}_t^*} \right) + \frac{1}{\mathcal{L}_{T_1(\sigma)}^*} \\
&\leq (k - T_1(\sigma))\, C \left\| \frac{1}{d_\rho^*} \right\|_\infty (1-\sigma)^{-1} + \frac{1}{\mathcal{L}_{T_1(\sigma)}^*} \\
&\leq kC \left\| \frac{1}{d_\rho^*} \right\|_\infty (1-\sigma)^{-1} + \frac{1}{\mathcal{L}_{T_1(\sigma)}^*}
\end{aligned}$$

As $\frac{1}{\mathcal{L}_{T_1(\sigma)}^*}$ is fixed, there exists a $T(\sigma) \geq T_1(\sigma)$ such that $\frac{1}{\mathcal{L}_{T_1(\sigma)}^*} \leq kC \left\| \frac{1}{d_\rho^*} \right\|_\infty (1-\sigma)^{-1}$ for $k \geq T(\sigma)$. Therefore, in this range,

$$\frac{1}{\mathcal{L}_k^*} \leq 2kC \left\| \frac{1}{d_\rho^*} \right\|_\infty (1-\sigma)^{-1}.$$

□

2.3 Linear Convergence Analysis

Theorem 2.6 (Linear Convergence under Weighted State Value Error)**.** *Let $\{\pi^k\}_{k\geq 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1}V^k \geq V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s)A^k(a \mid s) \geq 0, \forall s$). Assume that*

$$\mathcal{L}_k^{k+1} \geq C_k \mathcal{L}_k^*,$$

where $\mathcal{L}_k^{k+1}$ and $\mathcal{L}_k^*$ are defined under certain optimal policy $\pi^{k,*}$. One has

$$\mathcal{L}_k^* \leq \left(1 - (1-\gamma) \left\| \frac{d^{\pi^{k,*}}}{\rho} \right\|_\infty^{-1} C_k \right) \mathcal{L}_k^*.$$

Proof. The techniques used in this proof is the same with Theorem 2.3, except that the condition given in the theorem is different, please refer to the paper for details. □

Theorem 2.7 (Linear Convergence under Infinite Norm I)**.** *Assume for some $C_k \in (0,1)$,*

$$\forall s: \quad \sum_a \pi_{s,a}^{k+1} A_{s,a}^k \geq C_k \max_a A_{s,a}^k.$$

Then,

$$\|V^* - V^{k+1}\|_\infty \leq (1-(1-\gamma)C_k)\|V^* - V^k\|_\infty.$$

Proof. First the assumption can be rewritten using the performance difference lemma.

$$\mathcal{T}^{k+1}V^k(s) - V^k(s) \geq C_k(\mathcal{T}V^k(s) - V^k(s)).$$

It follows that $\forall s$,

$$\begin{aligned}
V^*(s) - V^{k+1}(s) &\overset{(a)}{=} V^*(s) - \mathcal{T}^{k+1}V^{k+1}(s) \\
&\overset{(b)}{\leq} V^*(s) - \mathcal{T}^{k+1}V^k(s) \\
&\leq V^*(s) - V^k(s) - C_k(\mathcal{T}V^k(s) - V^k(s)) \\
&\overset{(c)}{=} V^*(s) - V^k(s) + C_k\mathcal{T}V^*(s) - C_k\mathcal{T}V^k(s) + C_kV^k(s) - C_kV^*(s) \\
&= C_k(\mathcal{T}V^*(s) - \mathcal{T}V^k(s)) + (1-C_k)(V^*(s) - V^k(s)).
\end{aligned}$$

The contraction property of Bellman optimality operator gives

$$\begin{aligned}
\|V^* - V^{k+1}\|_\infty &\leq \|C_k(\mathcal{T}V^* - \mathcal{T}V^k)\|_\infty + (1-C_k)\|V^* - V^k\|_\infty \\
&\leq \gamma C_k\|V^* - V^k\|_\infty + (1-C_k)\|V^* - V^k\|_\infty \\
&= (1-(1-\gamma)C_k)\|V^* - V^k\|_\infty.
\end{aligned}$$

□

Theorem 2.8 (Linear Convergence under Infinite Norm by Controlling Error Terms)**.** *Let $\{\pi^k\}_{k\geq 0}$ be a policy sequence which satisfies $\mathcal{T}^{k+1}V^k \geq V^k$ for all k (or equivalently, $\sum_a \pi^{k+1}(a \mid s)A^k(s,a) \geq 0, \forall s$). Assume there exists a constant $c_0 > 0$ such that*

$$\forall s: \sum_a \pi_{s,a}^{k+1} A_{s,a}^k \geq C \max_a A_{s,a}^k - \epsilon_k \quad \textit{and} \quad \sum_{t=0}^{k-1} \frac{\epsilon_t}{(1-(1-\gamma)C)^{t+1}} \leq c_0.$$

Then

$$\|V^* - V^k\|_\infty \leq (1-(1-\gamma)C)^k(\|V^* - V^k\|_\infty + c_0).$$

Proof. Repeating the proof of Theorem 2.12 yields

$$\left\|V^* - V^{k+1}\right\|_\infty \leq (1-(1-\gamma)C)\left\|V^* - V^k\right\|_\infty + \varepsilon_k.$$

Iterating this procedure gives

$$\left\|V^* - V^k\right\|_\infty \leq (1-(1-\gamma)C)^k\left\|V^* - V^0\right\|_\infty + \sum_{t=0}^{k-1} \varepsilon_t(1-(1-\gamma)C)^{k-1-t}$$

□

Theorem 4 (Policy Iteration Convergence)**.** *(Informal)*

1. $Q^{k+1} \geq \mathcal{T}Q^k \geq Q^k$
2. $\|Q^{k+1} - Q^*\|_\infty \geq \gamma\|Q^k - Q^*\|_\infty$

The proof is built on the properties of optimal Bellman operator and contraction mapping.

Proof. We first show that $\mathcal{T}Q^k \geq Q^k$.

$$\begin{aligned}\mathcal{T}Q^{\pi_k}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^{\pi_k}(s',a')\right] \\ &\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_k}(s', \pi_k(s'))\right] = Q^{\pi_k}(s,a).\end{aligned}$$

Then we show that $Q^{k+1} \geq \mathcal{T}Q^k$. Let us first show that $Q^{k+1} \geq Q^k$:

$$Q^{\pi_k} = r + \gamma P^{\pi_k} Q^{\pi_k} \leq r + \gamma P^{\pi_{k+1}} Q^{\pi_k} \leq \sum_{t=0}^{\infty} \gamma^t (P^{\pi_{k+1}})^t r = Q^{\pi_{k+1}},$$

from this result, we know that policy iteration yields non-negative policy improvement for each iterate.

Then we show that $Q^{k+1} \geq \mathcal{T}Q^k$:

$$\begin{aligned}Q^{\pi_{k+1}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_{k+1}}(s', \pi_{k+1}(s'))\right] \\ &\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[Q^{\pi_k}(s', \pi_{k+1}(s'))\right] \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^{\pi_k}(s',a')\right] = \mathcal{T}Q^{\pi_k}(s,a)\end{aligned}$$

Then we prove the linear convergence under the infinite norm by using the first claim:

$$\left\|Q^\star - Q^{\pi_{k+1}}\right\|_\infty \leq \left\|Q^\star - \mathcal{T}Q^{\pi_k}\right\|_\infty = \left\|\mathcal{T}Q^\star - \mathcal{T}Q^{\pi_k}\right\|_\infty \leq \gamma \left\|Q^\star - Q^{\pi_k}\right\|_\infty.$$

□

In policy iteration, we have that

$$\mathcal{T}Q^k \geq Q^k,$$

and from this inequality, we can derive that the value function at each iterate is always improving, therefore leading to convergence.

So the authors want to use this result as a bridge, because some policy gradient methods are similar to the PI under certain conditions, for example, PPG under infinite step size.

So I think the authors are trying to bound this term, which represents the value function improvement at the k-th iteration.

$$\mathcal{T}^{k+1}V^k - V^k$$

with the following term that represents the value function improvement in the policy iteration:

$$\mathcal{T}V^k \geq V^k.$$

So this is a brief answer to our second question, and I believe it also explains the meanings of $\mathcal{L}_k^*$ and $\mathcal{L}_k^{k+1}$.

Now we can get back to our first question, which is how can we establish such a bound. I give an example of a specific algorithm, which is also the content of section 3 in this paper.

3 Projected Policy Gradient

By parameterizing the policy π in the form of π_θ, we can use gradient methods.

$$\theta^+ \leftarrow \theta + \eta \nabla_\theta V^{\pi_\theta}(\mu).$$

Perhaps the simplest form of the parameterization is the direct/simplex parameterization, recall that the policy class for each state is initially defined as all the distributions on the probability simplex over the action space:

$$\Pi := \{\pi = (\pi(\cdot \mid s))_{s \in \mathcal{S}} \mid \pi(\cdot \mid s) \in \Delta(\mathcal{A}) \text{ for all } s \in \mathcal{S}\}.$$

Direct parameterization means that the parameterized policy class equals the original admissible policy class. We direct search in the probability simplex for each state.

$$\pi = \left\{ (\pi_s)_{s \in \mathcal{S}} \mid \pi_s \in \Delta(\mathcal{A}) \right\}.$$

Now consider that if we directly apply the gradient ascent method, which is

$$\pi^{k+1} \leftarrow \pi^k + \eta \nabla_\pi V^\pi(s) \mid_{\pi=\pi^k}.$$

The new policy π^{k+1} may not satisfy the probability constraint, which is

$$\sum_a \pi^{k+1}(a \mid s) = 1.$$

So we project the π^{k+1} onto the probability simplex $\Delta(\mathcal{A})$,

$$\pi^{k+1} = \arg\min_\pi \left\| \pi - \left(\pi^k + \eta_k \nabla_{\pi^k} V^{\pi^k}(s) \right) \right\|^2.$$

$$\begin{aligned}
\text{LHS} &= \arg\min_\pi \left\| \pi - \left(\pi^k + \eta_k \nabla_{\pi^k} V^{\pi^k}(s) \right) \right\|^2 \\
&= \arg\min_\pi \|\pi - \pi^k\|^2 - 2\eta_k < \pi - \pi^k, \nabla_{\pi^k} V^{\pi^k}(s) > - \|\eta_k \nabla_{\pi^k} V^{\pi^k}(s)\|^2 \\
&= \arg\min_\pi \|\pi - \pi^k\|^2 - 2\eta_k < \pi - \pi^k, \nabla_{\pi^k} V^{\pi^k}(s) > \\
&= \arg\max_\pi \left\{ \eta_k < \pi - \pi^k, \nabla_{\pi^k} V^{\pi^k}(s) > - \frac{1}{2} \|\pi - \pi^k\|^2 \right\} \\
&\overset{(a)}{=} \arg\max_\pi \left\{ \eta_k \left\langle \pi - \pi^k, \frac{d_\mu^\pi(s)}{1-\gamma} Q^\pi(s, \cdot) \right\rangle - \frac{1}{2} \|\pi - \pi^k\|^2 \right\} \\
&= \text{Proj}_{\Delta(\mathcal{A})}(\pi^k + \eta^k Q^k(s, \cdot)),
\end{aligned}$$

(a) holds because of the policy gradient theorem, we obtain the policy gradient:

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

where $\eta_s^k = \frac{\eta_k}{1-\gamma} d_\mu^k(s)$, and $\text{Proj}_{\Delta(\mathcal{A})}$ means the projection onto the probability simplex.

Lemma 3.1 (Improvement Lower Bound)**.** *Let $\eta_k > 0$ be step size in the k-th iteration of PPG. Then,*

$$\sum_a \pi_{s,a}^{k+1} A_{s,a}^k \geq \frac{(\max_a A_{s,a}^k)^2}{\max_a A_{s,a}^k + \frac{2+5|\mathcal{A}|}{\eta_k}}$$

Proof. □

Well, the proof utilizes the properties of the projection methods, it is another very long story, so I won't give the detailed proof this time. Just know that for PPG algorithm, the two terms $\mathcal{T}^{k+1}V^k - V^k$ and $\mathcal{T}V^k - V^k$ can be linked in this inequality.

Then it takes little effort to transform the inequality into our assumption.

$$
\begin{aligned}
\mathcal{L}_k^{k+1} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^*}\left[\sum_a \pi_{s,a}^{k+1} A_{s,a}^k\right] \\
&\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^*}\left[\frac{(\max_a A_{s,a}^k)^2}{\max_a A_{s,a}^k + \frac{2+5|\mathcal{A}|}{\eta_s^k}}\right] \\
&\overset{(a)}{\geq} \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^*}\left[\frac{(\max_a A_{s,a}^k)^2}{\max_a A_{s,a}^k + \frac{2+5|\mathcal{A}|}{\eta_k\tilde{\mu}}}\right] \\
&\overset{(b)}{\geq} \frac{1}{1-\gamma}\left[\frac{\left(\mathbb{E}_{s\sim d_\rho^*}\left[\max_a A_{s,a}^k\right]\right)^2}{\mathbb{E}_{s\sim d_\rho^*}\left[\max_a A_{s,a}^k\right] + \frac{2+5|\mathcal{A}|}{\eta_k\tilde{\mu}}}\right] \\
&\overset{(c)}{\geq} \frac{(1-\gamma)\left(\mathcal{L}_k^*\right)^2}{(1-\gamma)\mathcal{L}_k^* + C_1(\eta_k)} \\
&\overset{(d)}{\geq} \frac{(1-\gamma)\left(\mathcal{L}_k^*\right)^2}{1 + C_1(\eta_k)},
\end{aligned}
\tag{3.1}
$$

where (a) is based on the definition of η_s^k.

$$
\eta_s^k = \frac{\eta_k}{1-\gamma} d_\mu^k(s) = \eta_k \frac{1}{1-\gamma}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid \mu, \pi^k) \geq \eta_k\tilde{\mu},
$$

where $\tilde{\mu} = \min_s \mu(s) > 0$.

(b) holds because of Jensen's inequality.

(c) is based on the fact that

$$
\mathbb{E}_{s\sim d_\rho^*}\left[\max_a A_{s,a}^k\right] \geq (1-\gamma)\mathcal{L}_k^*.
$$

(d) holds because

$$
0 \leq \mathcal{L}_k^* = V^*(\rho) - V^k(\rho) \leq \frac{1}{1-\gamma}
$$

Letting $\eta_k = \eta$ and invoking Theorem 2.3 yields the $O(1/k)$ sublinear convergence of PPG.

Lemma 3.2. *Consider PPG with constant step size $\eta_k = \eta$ (in this case η_s^k is simplified to η_s). If the state value of π^k satisfies*

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

then π^{k+1} is an optimal policy.

Theorem 3.3 (Linear Convergence for any Constant Step Size)**.** *Consider PPG with any constant step size $\eta > 0$. One has*

$$
V^*(\rho) - V^k(\rho) \leq \left(1 - (1-\gamma)\left\|\frac{d_\rho^*}{\rho}\right\|_\infty^{-1} \cdot \frac{(1-\gamma)C_2(\eta)}{(1-\gamma)C_2(\eta) + C_1(\eta)}\right)^k (V^*(\rho) - V^0(\rho)),
$$

where $C_1(\eta) \coloneqq \frac{2+5|\mathcal{A}|}{\eta\tilde{\mu}}$ is defined as above and $C_2(\eta) \coloneqq \frac{\tilde{\rho}\Delta}{2}\frac{\eta\tilde{\mu}\Delta}{1+\eta\tilde{\mu}\Delta}$.

Proof. Let T be the iteration at which PPG terminates with the optimal policy (i.e., π^T is an output optimal policy). Then by Lemma 3.2,

$$
\|V^* - V^k\|_\infty > \frac{\Delta}{2}\frac{\eta_s\Delta}{1+\eta_s\Delta}, \quad \text{if } k \leq T-2.
$$

It follows that $\forall k \leq T-2$,

$$
\mathcal{L}_k^* = V^*(\rho) - V^k(\rho) \geq \tilde{\rho}\|V^* - V^k\|_\infty > \frac{\Delta}{2}\frac{\eta_s\Delta}{1+\eta_s\Delta}\tilde{\rho} \geq \frac{\Delta}{2}\frac{\eta\tilde{\mu}\Delta}{1+\eta\tilde{\mu}\Delta}\tilde{\rho} \coloneqq C_2(\eta).
$$

Combining with Equation (3.1), we have

$$\begin{aligned}\mathcal{L}_k^{k+1} &\geq \frac{(1-\gamma)\left(\mathcal{L}_k^*\right)^2}{(1-\gamma)\mathcal{L}_k^* + C_1(\eta)} \\ &= \left(1 - \frac{C_1(\eta)}{(1-\gamma)\mathcal{L}_k^* + C_1(\eta)}\right) \cdot \mathcal{L}_k^* \\ &\geq \left(1 - \frac{C_1(\eta)}{(1-\gamma)C_2(\eta) + C_1(\eta)}\right) \cdot \mathcal{L}_k^* \\ &= \frac{(1-\gamma)C_2(\eta)}{(1-\gamma)C_2(\eta) + C_1(\eta)} \cdot \mathcal{L}_k^*.\end{aligned}$$

Then the linear convergence result follows from Theorem 2.6.

□

Theorem 3.4 (Linear Convergence for Non-Adaptive Increasing Step Size)**.** *Letting $C_3 > 0$ be any constant, assume the step size of PPG satisfies*

$$\eta_k \geq \frac{2 + 5|\mathcal{A}|}{\tilde{\mu}} \frac{1-\gamma}{C_3} \left(1 + \frac{1-\gamma}{C_3}\right)^{k+1}$$

Then one has

$$V^*(\rho) - V^k(\rho) \leq \frac{1}{1-\gamma} \left(1 - (1-\gamma) \left\| \frac{d_\rho^*}{\rho} \right\|_\infty^{-1} \cdot \frac{1-\gamma}{1-\gamma+C_3}\right)^k$$

Proof. The proof is by induction. First consider $k = 0$, we have the following inequality holds,

$$V^*(\rho) - V^k(\rho) \leq \frac{1}{1-\gamma}.$$

Assume it is true for $k \leq t$. First, the bound on η_t implies that

$$\begin{aligned}C_1(\eta_t) &\leq \frac{C_3}{1-\gamma} \left(1 - \frac{(1-\gamma)}{1-\gamma+C_3}\right)^{t+1} \\ &\leq \frac{C_3}{1-\gamma} \left(1 - (1-\gamma) \left\| \frac{d_\rho^*}{\rho} \right\|_\infty^{-1} \frac{(1-\gamma)}{1-\gamma+C_3}\right)^{t+1}.\end{aligned}$$

Assume

$$V^*(\rho) - V^t(\rho) \geq \frac{1}{1-\gamma} \left(1 - (1-\gamma) \left\| \frac{d_\rho^*}{\rho} \right\|_\infty^{-1} \frac{(1-\gamma)}{1-\gamma+C_3}\right)^{t+1}.$$

□

Notations

Table 1: Notations

Symbol	Definition/Domain	Meaning
$\mathcal{M}$		MDP
$\mathcal{S}$		State Space
$\mathcal{A}$		Action Space
s		state
s_t		state at time t
a		action
a_t		action at time t
$R(s,a)$	$\mathcal{S} \times \mathcal{A} \to \mathbb{R}$	Reward Function
r		reward signal
r_t		reward signal at time t
$P(s' \mid s,a)$	$\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	State Transition Probability
γ	$[0,1)$	Discount Factor
$\Delta(\mathcal{A})$	$\left\{\theta \in \mathbb{R}^{\lvert\mathcal{A}\rvert} \mid \theta_i \geq 0, \sum_{i=1}^{\lvert\mathcal{A}\rvert} \theta_i = 1\right\}$	Policy Space
$\pi(a \mid s)$	$\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$	Policy
μ/ρ	$\mathcal{S} \to \Delta(\mathcal{S})$	Initial State Distribution
$\tilde{\mu}/\tilde{\rho}$	$\min_s \mu(s)/\min_s \mu(s)$	/
Π		Set of Admissible Policies
π_θ	$\mathbb{R}^d \to \Pi$	Parameterized Policy

Table 2: Notations

Symbol	**Definition/Domain**	**Meaning**
$V^{\pi}(s)$	$\mathcal{S} \to \mathbb{R}$	`Value Function over state` s
$V^{\pi}(\mu)$		`Value Function over` μ
$V^{\pi_\theta}(s/\mu)$		`Value Function w.r.t` π^{θ}
$Q^{\pi}(s,a)$	$\mathcal{S} \times \mathcal{A} \to \mathbb{R}$	`Action Value Function`
$A^{\pi}(s,a)$	$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$	`Advantage Function`
$\mathcal{T}^{\pi}$	$\mathcal{T}^{\pi}V(s) :=$ $\mathbb{E}_{a\sim\pi(\cdot\|s)}\mathbb{E}_{s'\sim P(\cdot\|s,a)}\big[r(s,a)+\gamma V(s')\big]$.	`Bellman Operator over` π
$\mathcal{T}$	$\mathcal{T}V(s) := \max_{\pi\in\Pi}\mathcal{T}^{\pi}V(s) =$ $\max_{a\in\mathcal{A}}\big\{\mathbb{E}_{s'\sim P(\cdot\|s,a)}\big[r(s,a)+\gamma V(s')\big]\big\}$	`Optimal Bellman Operator`
π^*		`Optimal Policy`
$V^*(\cdot)$		`Optimal State Value Function`
$Q^*(\cdot)$		`Optimal Action Value Function`
$A^*(s,a)$		`Optimal Advantage Function`
π^k		`Policy at k-th Iteration`
V^k		`State Value Function for` π^k
Q^k		`Action Value Function for` π^k
A^k		`Advantage Value Function for` π^k
$\mathcal{T}^{k+1}$		`Bellman Operator over` π^{k+1}
$d_{\rho}^{\pi}(s)$	$\mathcal{S} \to \Delta(\mathcal{S})$	`Visitation Measure` for ρ and π
$\mathcal{L}_k^{k+1}$		
$\mathcal{L}_k^*$		
$\mathcal{L}_{k+1}^*$		