Improving Different Aspects in RL - Accelerating Convergence Rate & Enhancing Safety and Robustness

by

Yue Gao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Reinforcement learning (RL) has moved from toy domains to real-world applications, while each of these applications has inherent difficulties which are long-standing challenges in RL, such as: stucking at plateaus, limited training time, costly exploration and safety considerations. I, with my collaborates [19], [35] proposed several RL algorithms to improve different aspects of the performance including **geometry-aware gradient descent (GNGD)**, a policy gradient method (which is also applicable to other non-convex optimizations) which is powerful in terms of theoretical convergence result; and **a family of Q-learning algorithms** enhancing risk-aversion and robustness empirically in trading market.

Not only in RL, **geometry-aware descent methods** could also be applied in any first-order non-uniform optimization and can converge to global optimality faster than the classical $\Omega(1/t^2)$ lower bounds.

e.g, for its application to PG and GLM, it can be shown that normalizing the gradient ascent method can accelerate convergence to $O(e^{-t})$ while incurring less overhead than existing algorithms, which significantly improves the best known results. It can also be shown that the proposed geometry-aware descent methods escape landscape plateaus faster than standard gradient descent. Experimental results are used to illustrate and complement the theoretical findings.

On the empirical side of RL, for the purpose of enhancing robustness and reducing risk, a family of Q-learning algorithm were proposed by taking characteristics such as *risk-awareness*, *robustness to perturbations* and *low learning variance* as building blocks, and they perform well in trading market and balance theoretical guarantees with practical use.

Preface

My thesis is mainly composed of the problems discussed in two of my works - *Leveraging Non-uniformity in First-order Non-convex Optimization* [35], which is accepted to the conference **ICML 2021**, and *Robust Risk-Sensitive Reinforcement Learning Agents for Trading Markets*[19], which is accepted to the conference workshop **ICML 2021 RL4RealLife**.

In Chapter 4, I introduce the algorithm **geometry aware gradient descent (GNGD)**, which is co-proposed by Jincheng Mei*, **Yue Gao***, Bo Dai, Csaba Szepesvari and Dale Schuurmans[35]. In this chapter, I present a categorised discussion of convergence rate of GNGD, where I contributed to the proof of case (1), case (3) in Section 4.2.1, and I raised & implemented the example of (1a), (1b), and (2b) of Example 4. In Chapter 5, I present the application of GNGD on Markov Decision Process, where I observed the non-uniform smoothness of MDP and proved Lemmas 2 and 3, jointly proved Theorems 1 and 2 and Lemmas 6 and 7 with Jincheng Mei[35], and implemented GNGD and GD on the one-state MDP. In Chapter 6, I present the application of GNGD on Generalized Linear Model, where I jointly proved Lemmas 9 and 10 and Theorem 4, and implemented GNGD and GD on GLM.

In Chapter 7, I present the family of Q-learning algorithms RA2-Q, RA2.1-Q, RA3-Q and discuss their convergence guarantees, where all of those algorithms are proposed by myself. In Chapter 8, I present the empirical game theory for risk-averse payoff game, where the theorems and proofs were established by me, and checked & revised by my co-authors Kry Yik Chau Lui, Pablo Hernandez-Leal [19]. In Chapter 9, I present the empirical results of my proposed algorithms on trading market, where the experiments are done by myself.

What's past is prologue.

– William Shakespeare, 1611.

Acknowledgements

Much of this work was done when I was supervised by Csaba Szepesvari and Or Sheffet. Csaba gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC; and Or acknowledges grant # 2017–06701 of the Natural Sciences and Engineering Research Council of Canada (NSERC).

I want to acknowledge **BorealisAI** for funding & supervising my research internship work 'Robust Risk-Sensitive Reinforcement Learning Agents for Trading Markets'[19].

I want to thank my supervisor Csaba Szepesvari for funding and supervising my work. Csaba is the most erudite person in RL that I've ever known, as well as helpful and amicable. I also appreciate my co-supervisor Or Sheffet for his funding and supervising, without whom I would not get a chance to do research in differential privacy. During my research, Or enlightened my thoughts, made his valuable suggestions, and helped me on my writings. It's my good fortune to meet those two patient and professional scholars in my early academic career, I'll keep my enthusiasm for research under their inspiration.

I'd like to to express my gratitude for the help from my knowledgeable and enlightening co-authors Jincheng Mei, Bo Dai, Dale Schuurmans, Kry Yik Chau Lui, and Pablo Hernandez-Leal, without whom I would not be able to finish those works. Also, I'm very honored to be a member of AMII and RLAI, the great people there often inspire my research ideas through our communication, and I can always learn a lot from other people's diversified works.

I also appreciate my colleagues in BorealisAI, during my research internship there, they're all friendly and always supportive for my problems. Especially I'm grateful to Pablo, who enlightened my research topic and is always patient to answer each of my question and correct my paper.

I want to thank my friends living in Hong Kong, U.S., China, and Sweden, who often chat with me to relieve my boredom and loneliness during the COVID lockdown.

Finally, I want to express my deepest gratitude to my parents, who encouraged me to be curious about the world at the beginning of my life, and always respect whatever I'm enthusiastic about. When I encounter any academic or life setbacks, my parents always encourage me to gaze myself in a broader and more inclusive way.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Aspects in Reinforcement Learning - Convergence Rate & Risk-Awareness

Reinforcement learning (RL) has moved from toy domains to real-world applications such as games [8], navigation [7], software engineering [4], industrial design [39], and finance [31]. Each of these applications has inherent difficulties which are long-standing fundamental challenges in RL, such as: stucking at plateaus, limited training time, costly exploration and safety considerations, among others. This work focus on improving different aspects in RL - *Accelerate convergence rate and escaping plateaus faster* for PG methods (tabular case), and *Reducing risk and enhancing robustness* for Q learning algorithms applied in trading market.

Policy gradient (PG) methods is an essential branch in RL, it's a type of reinforcement learning techniques that rely upon optimizing parametrized policies with respect to the expected return. Inspired by Mei *et al.*, Mei *et al.*, I would explore the ways to escape plateaus and accelerate convergence rate of gradient descent. Commonly used techniques to escape plateaus include NGD [41], RMSProp[24], but each of them has its own shortcomings. As shown in Chapter 4, NGD does not converge in some cases, and RMSProp does not accelerate the escape rate significantly when the gradient is extremely close to zero. Inspired by the non-uniform properties of functions, novel gradient-based methods that better exploit local structure could be proposed, and the convergence rate analysis could be improved. Not only in PG, such methods could be applied in different scenarios like Generalized Linear Model, non-convex optimizations and so on.

Q learning is a model-free reinforcement learning method to learn the value of an action in a particular state. It's widely used in cases where the states and actions are limited, a typical example is trading market agents. Trading markets represent a real-world financial application to deploy reinforcement learning agents, however, they often meet challenges such as high variance and costly exploration. Moreover, markets are inherently a multiagent domain composed of many actors taking actions and changing the environment, thus requiring the agents to be robust. To tackle these type of scenarios, agents need to exhibit certain characteristics such as *risk-awareness*, *robustness to perturbations* and *low learning variance*. In this work, I take those three characteristics as building blocks and propose a family of three *Q learning algorithms*. Those algorithms either theoretically or empirically reduce risk or enhance robustness. In order to evaluate the performance of the algorithms, I'll extend and apply empirical game theory.

1.2 Non-Uniformity in First Order Non-convex Optimization

While gradient-based algorithms remain the method of choice in machine learning, the convergence of such algorithms to global minimizers has still only been established in restrictive settings where one can assert two strong assumptions about the objective function: (1) that the objective is smooth, and (2) that the objective satisfies a gradient dominance over sub-optimality such as the Łojasiewicz inequality. I, with my collaborators find it beneficial to recall the definitions of these properties.

In my work, I, together with my collaborators[35], expanded the class of problems for which gradient-based optimization is globally convergent, develop novel gradient-based methods that better exploit local structure, and improve the convergence rate analysis. We achieve these results by first defining then investigating a new set of *non-uniform* smoothness and Łojasiewicz inequalities, which generalize the classical definitions and allow a refined characterization of the space of objectives. Given these refined notions, we then proposed novel gradient-based algorithms that improve previous methods for these new problem classes, and extended the analysis to exploit these new forms of non-uniformity, achieving significantly stronger convergence rates in different cases. Importantly, these improvements are achieved in non-convex optimization problems that arise in relevant machine learning problems.

1.3 Risk and Robustness in Trading Market RL Algorithms

In finance, there are some examples of RL in stochastic control problems such as option pricing [32], market making [50], and optimal execution [44]. However, the most well-known finance application is algorithmic trading, where the goal is to design algorithms capable of automatically making trading decisions based on a set of mathematical rules computed by a machine [52].

In algorithmic trading the environment represents the market (and the rest of the actors). The agent's task is to take actions related to how and how much to trade, and the objective is usually to maximize profit while considering risk. There are diverse challenges in this setting such as partial observability, a large action space, a hard definition of rewards and learning objectives [52]. In this work I focus on two properties for learning agents in realistic scenarios: *risk assessment and robustness.*

Risk assessment is a cornerstone in financial applications. A well-known approach is to consider risk while assessing the performance (profit)[1] of a trad-

[1] Even when the usual financial term for profit is *return*, this could be confused with the usual definition of return in RL (cumulative sum of discounted rewards).

ing strategy. Here, risk is a quantity related to variance of the profit and it is commonly refereed to as “volatility”. In particular, the Sharpe ratio [48] considers both the generated profit and the risk (variance) associated with a trading strategy. Note that this objective function (Sharpe ratio) is different from traditional RL where the goal is to optimize the expected return without risk considerations. There are existing works that proposed risk-sensitive RL algorithms [18], [38] and variance reduction techniques [3]. In a similar spirit my proposed algorithms aim to reduce variance while also having convergence guarantees and improved robustness via adversarial learning.

Deep RL has been shown to be brittle in many scenarios [22]. Therefore, improving robustness is essential for deploying agents in realistic scenarios. A line of work has improved robustness of RL agents via adversarial perturbations [40], [46]. In particular, the framework assumes a learning adversary who is allowed to take over control at regular intervals. This approach has shown good experimental results in robotics [45], and my proposed algorithms extend on this idea.

In trading market, the state & actions are limited, so Q learning method is a wise choice. First, I contribute with two algorithms that use risk-averse objective functions and variance reduction techniques. Then, I augment the framework to multi-agent learning and assume an adversary which can take over and perturb the learning process. My third algorithm perform well under this setting and balance theoretical guarantees with practical use.

Since the motivation is to use RL agents in trading markets (which can be seen as multi-agent interactions) I also evaluate these agents from the perspective of game theory. However, it may be too difficult to analyze in standard game theoretic framework since there is no normal form representation (commonly used to analyze games). Fortunately, empirical game theory [56], [59] overcomes this limitation by using the information of several rounds of repeated interactions and assuming a higher level of strategies (agents’ policies).

These modifications have made possible the analysis of multi-agent interactions in complex scenarios such as markets [11], and multi-agent games [53]. However, these works have not studied the interactions under risk-aware metrics as my work.

Chapter 2

Preliminaries

2.1 Single Agent RL Setting - A Tabular Case

For a finite set $\mathcal{N}$, let $\Delta(\mathcal{N})$ denote the set of all probability distributions on $\mathcal{N}$. A finite MDP $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ is determined by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, a scalar reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1)$.

In policy-based RL, an agent interacts with the environment, i.e., the MDP $\mathcal{M}$, using a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Given a state s_t, the agent takes an action $a_t \sim \pi(\cdot|s_t)$, receives a one-step scalar reward $r(s_t, a_t)$ and a next-state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. The long-term expected reward, also known as the value function of π under s, is defined as

$$V^\pi(s) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0=s, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \tag{2.1}$$

The state distribution of π is defined as,

$$d^\pi_{s_0}(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi, \mathcal{P}). \tag{2.2}$$

Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, we denote $V^\pi(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$ and $d^\pi_\rho(s) \coloneqq \mathbb{E}_{s_0\sim\rho}\left[d^\pi_{s_0}(s)\right]$. There exists an optimal policy π^* such that $V^{\pi^*}(\rho) = \sup_{\pi:\mathcal{S}\to\Delta(\mathcal{A})} V^\pi(\rho)$. For convenience, we denote $V^* \coloneqq V^{\pi^*}$. Consider a tabular representation, i.e., $\theta(s, a) \in \mathbb{R}$ for all (s, a), so that the policy π_θ can be parameterized by θ as $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$; that is, for all (s, a),

$$\pi_\theta(a|s) = \frac{\exp\{\theta(s, a)\}}{\sum_{a'\in\mathcal{A}} \exp\{\theta(s, a')\}}. \tag{2.3}$$

When there is only one state the policy $\pi_\theta = \text{softmax}(\theta)$ is defined as $\pi_\theta(a) = \exp\{\theta(a)\}/\sum_{a'\in\mathcal{A}}\exp\{\theta(a')$. The problem of policy-based RL is then to find a policy π_θ that maximizes the value function, i.e.,

$$\sup_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} V^{\pi_\theta}(\rho). \tag{2.4}$$

For convenience, and without loss of generality, we assume $r(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

2.2 Multi Agent RL Setting

In RL, each agent i aims to maximize its own total expected return, e.g., for a Markov game with two agents, for a given initial state distribution d_0, the discounted returns are respectively :

$$J^1(d_0, \pi^1, \pi^2) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[r_t^1 | \pi^1, \pi^2, d_0\right] \tag{2.5}$$

$$J^2(d_0, \pi^1, \pi^2) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[r_t^2 | \pi^1, \pi^2, d_0\right] \tag{2.6}$$

where γ is a discount factor, $r_t^1, r_t^2,\ t = 1, 2, ...$ are respectively immediate rewards for agent 1 & 2. And a Nash equilibrium for Markov game (with two agents) is defined as following

Definition 1. *[23] A Nash equilibrium point of game (J^1, J^2) is a pair of strategies (π_*^1, π_*^2) such that for $\forall s \in \mathcal{S}$,*

$$J^1(s, \pi_*^1, \pi_*^2) \geq J^1(s, \pi^1, \pi_*^2) \quad \forall \pi^1 \tag{2.7}$$

$$J^2(s, \pi_*^1, \pi_*^2) \geq J^2(s, \pi_*^1, \pi^2) \quad \forall \pi^2 \tag{2.8}$$

A Markov game for N agents is defined by a set of states $\mathcal{S}$ describing the possible configurations of all agents, a set of actions $\mathcal{A}_1, ..., \mathcal{A}_N$ and a set of observations $\mathcal{O}_1, ..., \mathcal{O}_N$ for each agent. To choose actions, each agent i uses a stochastic policy $\pi_{\theta_i} : \mathcal{O}_i \times \mathcal{A}_i \to [0,1]$ parameterized by θ_i, which produces the next state according to the state transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \to \mathcal{S}$. Each agent i obtains rewards as a function of the state and agents' action

$r_i : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \rightarrow \mathbb{R}$, and receives a private observation correlated with the state $\mathbf{o}_i : \mathcal{S} \rightarrow \mathcal{O}_i$. The initial states are determined by a distribution $d_0 : \mathcal{S} \rightarrow [0,1]^{|\mathcal{S}|}$. In multi-agent Q learning, the Q tables are defined over joint actions for each of the agents. Each agent receives rewards according to its reward function, with transitions dependent on the actions chosen jointly by the set of agents.

2.3 Empirical Game Theory

The multi-agent behaviours in a trading market could be analyzed using empirical game theory, where a *player* corresponds to an agent, and a *strategy* corresponds to a learning algorithm. Then, in a p-player game, players are involved in a single round strategic interaction. Each player i chooses a strategy π^i from a set of k strategy $S^i = \{\pi^i_1, ..., \pi^i_k\}$ and receives a stochastic payoff $R^i(\pi^1, ..., \pi^p) : S^1 \times S^2 \times ... \times S^p \rightarrow \mathbb{R}$. The underlying game that is usually studied is $r^i(\pi^i, ..., \pi^p) = \mathbb{E}[R^i(\pi^1, ..., \pi^p)]$. In general, we denote the payoff of player i as μ^i and $\mathbf{x}^{-i}$ as the joint strategy of all players except for player i.

Definition 2. *A joint strategy* $\mathbf{x} = (x^1, ..., x^p) = (x^i, \mathbf{x}^{-i})$ *is a Nash equilibrium if for all* i *:*

$$\mathbb{E}_{\pi \sim \mathbf{x}} \left[\mu^i(\pi)\right] = \max_{\pi^i} \mathbb{E}_{\pi^{-i} \sim \mathbf{x}^{-i}} \left[\mu^i(\pi^i, \pi^{-i})\right] \tag{2.9}$$

Definition 3. *A joint strategy* $\mathbf{x} = (x^1, ..., x^p) = (x^i, \mathbf{x}^{-i})$ *is an* ϵ*-Nash equilibrium if for all* i*:*

$$\max_{\pi^i} \mathbb{E}_{\pi^{-i} \sim \mathbf{x}^{-i}} \left[\mu^i(\pi^i, \pi^{-i})\right] - \mathbb{E}_{\pi \sim \mathbf{x}} \left[\mu^i(\pi)\right] \leq \epsilon \tag{2.10}$$

Evolutionary dynamics have been used to analyze multi-agent interactions. A well-known model is replicator dynamics (RD) [57] which describes how a population evolves through time under evolutionary pressure (in our analysis, a population is composed by learning algorithms). RD assumes that the reproductive success is determined by interactions and their outcomes. For example, the population of a certain type increases if they have a higher *fitness* (in our case this means the expected return in certain interaction) than the population average; otherwise that population share will decrease.

To view the dominance of different strategies, it is common to plot the directional field of the payoff tables using the replicator dynamics for a number of strategy profiles $\mathbf{x}$ in the simplex strategy space [53]. In Section 9.1 I'll present results in this format evaluating my proposed algorithms.

Chapter 3

Non-uniform Properties

In this chapter, two core non-uniform properties, Non-uniform Smoothness (NS) and Non-uniform Łojasiewicz (NŁ) inequality are presented, and the key contribution is to show that the *combination* of those two non-uniform concepts could be applied to important non-convex objectives in machine learning, and allows the development of improved algorithms and analysis. The combination of NS and NŁ benefits a number of optimization problems in terms of *generality*, *better convergence results*, and *practical implications.*

3.1 Non-uniform Smoothness

Here's the definition of uniform smoothness :

Definition 4 (Smoothness). *The function $f : \Theta \to \mathbb{R}$ is β-smooth ($\beta > 0$) if it is differentiable and for all $\theta, \theta' \in \Theta$,*

$$\left| f(\theta') - f(\theta) - \left\langle \tfrac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \le \tfrac{\beta}{2} \cdot \|\theta' - \theta\|_2^2. \tag{3.1}$$

Based on Definition 4, the notion of smoothness can be generalized, where the non-uniform parameter depends on the function parameters non-uniformly.

Definition 5 (Non-uniform Smoothness (NS)). *The function $f : \Theta \to \mathbb{R}$ satisfies $\beta(\theta)$ non-uniform smoothness if f is differentiable and for all $\theta, \theta' \in \Theta$,*

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2,$$

where β is a positive valued function: $\beta : \Theta \to (0, \infty)$.

Here's a simple example of non-uniform smooth function :

Example 1. *Define the function $f : \mathbb{R} \to \mathbb{R}$ as*

$$f(\theta) = \theta^4$$

By Taylor Expansion, we have that

$$\left| f(\theta') - f(\theta) - \left\langle 4 \cdot \theta^3, \theta' - \theta \right\rangle \right| \leq \frac{12 \cdot \theta^2}{2} \cdot \|\theta' - \theta\|_2^2$$

When $\theta \approx 0$, $\frac{12 \cdot \theta^2}{2} \approx 0$; And when $\theta \approx \infty$, $\frac{12 \cdot \theta^2}{2} \approx \infty$, so unlike in Definition 4, we cannot use a constant β to denote the smoothness of function f.

In the later context, I'll refer to $\beta(\theta)$ in Definition 5 as the *NS coefficient*.

Zhang *et al.* raised the notion of (L_0, L_1) smoothness, where $\beta(\theta) = L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$. NS also generalizes the notion of (L_0, L_1) smoothness.

Wilson *et al.* proposed the notion of strong smoothness of order p, where $\beta(\theta) = c \cdot \|\nabla f(\theta)\|_2^{\frac{p-2}{p-1}}$, NS also reduces to this notion of strong smoothness of order p.

Finally, with $\beta(\theta) = c / \|\theta\|_p^2$, NS reduces to a special form of non-uniform smoothness considered in Mei *et al.*

I will show later that NS also covers other previously unstudied smoothness variants.

3.2 Non-uniform Łojasiewicz Inequality

Here's the definition of Łojasiewicz Inequality given by [29], [34], [47]

Definition 6. *[29], [34], [47] The differentiable function $f : \Theta \to \mathbb{R}$ satisfies the (C, ξ)-Łojasiewicz inequality if for all $\theta \in \Theta$,*

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C \cdot \left(f(\theta) - \inf_{\theta \in \Theta} f(\theta) \right)^{1-\xi}, \tag{3.2}$$

where $C > 0$ and $\xi \in [0, 1]$.

We leverage a generalized Łojasiewicz inequality introduced by [37].

Definition 7. *[37] The differentiable function $f : \Theta \to \mathbb{R}$ satisfies the $(C(\theta), \xi)$ non-uniform Łojasiewicz inequality if for all $\theta \in \Theta$,*

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi}, \tag{3.3}$$

where $\xi \in (-\infty, 1]$, and $C(\theta) : \Theta \to \mathbb{R} > 0$ holds for all $\theta \in \Theta$. In this definition, either $\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$, or $f(\theta^*)$ is replaced with $\inf_\theta f(\theta)$ if the global optimum is not achieved within the domain Θ.

ξ is the *NŁ degree* and $C(\theta)$ is the *NŁ coefficient.* Generally speaking, a larger NŁ degree ξ and NŁ coefficient $C(\theta)$ indicate faster convergence for gradient based algorithms.

Here're some examples of remarkable non-convex functions that satisfy the NŁ inequality for various ξ and $C(\theta)$.

Example 2. ***Expected reward, softmax parameterization.*** *As shown in Mei* et al.*, Lemma 3,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{3.4}$$

Expected reward, escort parameterization. *As shown in Mei* et al.*, Lemma 3,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r. \tag{3.5}$$

Value function, softmax parameterization. *As shown in Mei* et al.*, Lemma 8,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right]. \tag{3.6}$$

Value function, escort parameterization. *As shown in Mei* et al.*, Lemma 7,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{p}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s \|\theta(s, \cdot)\|_p} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right]. \tag{3.7}$$

Entropy regularized expected reward, softmax parameterization. *As shown in Mei* et al.*, Proposition 5,*

$$\left\| \frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} \right\|_2 \ge \sqrt{2\tau} \cdot \min_a \pi_\theta(a) \cdot \left[{\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - \pi_\theta^\top (r - \tau \log \pi_\theta) \right]^{\frac{1}{2}}. \tag{3.8}$$

Entropy regularized value function, softmax parameterization. *As shown in Mei* et al.*, Lemma 15,*

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}. \tag{3.9}$$

Entropy regularized value function, escort parameterization. *As shown in Mei* et al.*, Lemma 12,*

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{p \cdot \sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \frac{\min_{s,a} \pi_\theta(a|s)^{1-1/p}}{\max_s \|\theta(s,\cdot)\|_p} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}. \tag{3.10}$$

Cross entropy, escort parameterization. *As shown in Mei* et al.*, Lemma 17,*

$$\left\| \frac{d\{D_{\mathrm{KL}}(y \| \pi_\theta)\}}{d\theta} \right\|_2 \ge \frac{p}{\|\theta\|_p} \cdot \min_a \pi_\theta(a)^{\frac{1}{2} - \frac{1}{p}} \cdot D_{\mathrm{KL}}(y \| \pi_\theta)^{\frac{1}{2}}. \tag{3.11}$$

Generalized linear models, sigmoid activation, mean squared error.

Denote $u(\theta) := \min_i \{\pi_i \cdot (1 - \pi_i)\}$, *and* $v := \min_i \{\pi_i^* \cdot (1 - \pi_i^*)\}$. *We have, for all* $i \in [N]$,

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \ge 8 \cdot u(\theta) \cdot \min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^N (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}. \tag{3.12}$$

where λ_ϕ *is the smallest positive eigenvalue of* $\frac{1}{N} \cdot \sum_{i=1}^N \phi_i \phi_i^\top$.

Proof. Denote $\pi_i' \coloneqq \sigma(z_i')$, where $z_i' \coloneqq \phi_i^\top \theta + \zeta \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right)$ for some $\zeta \in [0,1]$. We have,

$$
\begin{aligned}
(\pi_i - \pi_i^*)^2 &= (\pi_i - \pi_i^*) \cdot \frac{d\sigma(z_i')}{dz_i'} \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \qquad \text{(by the mean value theorem)} && (3.13)\\
&= \pi_i' \cdot (1 - \pi_i') \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) && (3.14)\\
&\le \frac{1}{4} \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right). \qquad \left(x \cdot (1-x) \le \frac{1}{4},\ \forall x \in [0,1];\ (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \ge 0\right) && (3.15)
\end{aligned}
$$

Therefore we have,

$$
\begin{aligned}
\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 &\le \frac{1}{4N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \qquad \text{(by Eq. (A.224))} && (3.16)\\
&= \frac{1}{4N} \cdot \sum_{i=1}^{N} \frac{1}{\pi_i \cdot (1-\pi_i)} \cdot \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) && (3.17)\\
&\le \frac{1}{4N} \cdot \frac{1}{\min_i \pi_i \cdot (1-\pi_i)} \cdot \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \qquad \left((\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \ge 0\right) && (3.18)\\
&= \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1-\pi_i)} \cdot \left(\frac{2}{N} \cdot \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i\right)^\top (\theta - \theta^* - c \cdot v_{\phi,\perp}) && (3.19)\\
&= \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1-\pi_i)} \cdot \left(\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right)^\top (\theta - \theta^* - c \cdot v_{\phi,\perp}) \qquad \left(\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{2}{N} \cdot \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i\right) && (3.20)\\
&\le \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1-\pi_i)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \left\| \theta - \theta^* - c \cdot v_{\phi,\perp} \right\|_2 \qquad \text{(by Cauchy-Schwarz)} && (3.21)\\
&= \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \left\| \theta - \theta^* - c \cdot v_{\phi,\perp} \right\|_2, \qquad \left(u(\theta) \coloneqq \min_i \left\{\pi_i \cdot (1-\pi_i)\right\}\right) && (3.22)
\end{aligned}
$$

where $v_{\phi,\perp}$ is orthogonal to the space $\text{Span}\left\{\phi_1, \phi_2, \ldots, \phi_N\right\}$, and $\theta - \theta^* - c \cdot v_{\phi,\perp}$ refers to the vector after cutting off all the components $v_{\phi,\perp}$ from $\theta - \theta^*$, such that $\theta - \theta^* - c \cdot v_{\phi,\perp} \in \text{Span}\left\{\phi_1, \phi_2, \ldots, \phi_N\right\}$.

Next, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{d\sigma(z_i')}{dz_i'} \right)^2 \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \qquad \text{(by the mean value theorem)} \tag{3.23}$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i')^2 \cdot (1 - \pi_i')^2 \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \qquad \text{(by Eq. (A.224))} \tag{3.24}$$

$$\geq \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \tag{3.25}$$

$$= \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \cdot (\theta - \theta^*)^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^*) \tag{3.26}$$

$$= \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{3.27}$$

$$\geq \min \left\{ u(\theta)^2, v^2 \right\} \cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \qquad \left(v \coloneqq \min_i \left\{ \pi_i^* \cdot (1 - \pi_i^*) \right\} \right) \tag{3.28}$$

$$\geq \min \left\{ u(\theta)^2, v^2 \right\} \cdot \lambda_\phi \cdot \| \theta - \theta^* - c \cdot v_{\phi,\perp} \|_2^2, \tag{3.29}$$

where λ_ϕ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$. Therefore, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \| \theta - \theta^* - c \cdot v_{\phi,\perp} \|_2 \qquad \text{(by Eq. (A.228))} \tag{3.30}$$

$$\leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \frac{1}{\min \left\{ u(\theta), v \right\}} \cdot \frac{1}{\sqrt{\lambda_\phi}} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}, \qquad \text{(by Eq. (A.237))} \tag{3.31}$$

which implies,

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq 8 \cdot u(\theta) \cdot \min \left\{ u(\theta), v \right\} \cdot \sqrt{\lambda_\phi} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}. \qquad \square$$

Chapter 4

Geometry Normalized Gradient Descent (GNGD)

In optimization, gradient descent (GD) is a commonly used first-order iterative algorithm.

Definition 8 (Gradient Descent (GD))**.**

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla f(\theta_t). \tag{4.1}$$

The key challenge with deploying GD is choosing the step size η; if η is too large, instability ensues, if too small, progress becomes slow. Especially, GD can take arbitrarily long to escape from plateaus. There're many methods accelerating escaping plateaus, including RMSProp, normalized gradient descent, etc.

Definition 9 (Normalized Gradient Descent (NGD))**.**

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla f(\theta_t) \Big/ \left\| \nabla f(\theta_t) \right\|. \tag{4.2}$$

Although NGD can help escaping plateaus faster, in many cases, NGD looses the convergence guarantee.

Example 3. *Define the function $f : \mathbb{R} \to \mathbb{R}$ as*

$$f(\theta) = \theta^4$$

then the gradient of f is

$$\nabla f(\theta) = 4 \cdot \theta^3$$

Hence the NGD iteration rule is

$$\theta_{t+1} \leftarrow \theta_t - \eta. \tag{4.3}$$

By setting $\eta = 0.1$, at $\theta_{t_0} = 0.05$, Eq. (4.3) *yields $\theta_{t_0+1} = -0.05$, and in the following $t \geq t_0$, θ_t oscillate among 0.05 and -0.05.*

In the presence of *non-uniform* smoothness $\beta(\theta)$ given in NS, the step-size should be adapted to $1/\beta(\theta)$. This leads to a new variant of normalized gradient descent - Geometry-aware Normalized GD (GNGD).

4.1 GNGD Algorithm

Definition 10 (Geometry-aware Normalized GD (GNGD))**.**

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \frac{\nabla f(\theta_t)}{\beta(\theta_t)}. \tag{4.4}$$

For function f, if there's an efficient way to compute $\beta(\theta)$, GNGD is practical.

4.2 Convergence Rate of GD and GNGD in Different Function Classes

4.2.1 Convergence Rate of Functions

Here's an analysis for GD and GNGD based on non-uniform smoothness and non-uniform NŁ.

For function f s.t. $\inf_{\theta} f(\theta) > -\infty$ (for minimizing problem), it can be classified into different categories according to two non-uniform properties :

non-uniform smoothness and non-uniform NŁ.

Let $\beta(\theta)$ denote the non-uniform smoothness parameter, $\delta(\theta) := f(\theta) - f(\theta^*)$ be the sub-optimality gap, $(C(\theta), \xi)$ denote the non-uniform NŁ parameter. Note that we assume $\inf\limits_{t\geq 1} C(\theta_t) > 0$. Then the functions can be classified into the following categories and GD & GNGD respectively achieves convergence rates as stated below :

(1a) If $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (-\infty, 1/2)$, then GD with $\eta \in O(1)$ achieves $\delta(\theta_t) \in \Theta(1/t^{\frac{1}{1-2\xi}})$, and GNGD achieves $\delta(\theta_t) \in O(e^{-t})$;

(1b) If $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (-\infty, 1/2)$, then GD with $\eta \in O(1)$ achieves $\delta(\theta_t) \in \Theta(1/t^{\frac{1}{1-2\xi}})$, and GNGD achieves $\delta(\theta_t) \in O(e^{-t})$.

(2a) if $\beta(\theta) \leq L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$, then GD and GNGD both achieve $\delta(\theta_t) \in O(1/t^{\frac{1}{1-2\xi}})$ when $\xi \in (-\infty, 1/2)$, and $O(e^{-t})$ when $\xi = 1/2$. GNGD has strictly better constant than GD $(1 \geq C \geq C^2)$.

(2b) if $\beta(\theta) \leq L_0 \cdot \frac{\|\nabla f(\theta)\|^2}{\delta(\theta)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta)\|_2$, then GD and GNGD both achieve $\delta(\theta_t) \in O(1/t^{\frac{1}{1-2\xi}})$ when $\xi \in (-\infty, 1/2)$, and $O(e^{-t})$ when $\xi = 1/2$. GNGD has strictly better constant than GD $(1 \geq C \geq C^2)$.

(3a) if $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (1/2, 1)$, then GD with $\eta \in \Theta(1)$ does not converge, while GNGD achieves $\delta(\theta_t) \in O(e^{-t})$;

(3b) if $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (1/2, 1)$, then GD with $\eta \in \Theta(1)$ does not converge, while GNGD achieves $\delta(\theta_t) \in O(e^{-t})$.

For the detailed proof of those convergence rates, please check Mei *et al.*, Appendix A.1.

Remark 1. *The above cases cover all possibilities of the non-uniform smoothness parameter $\beta(\theta^*)$, where θ^* is the global minimum. Since $\nabla^2 f(\theta^*)$ is positive semi-definite if it exists.*

Here're examples of functions satisfying those non-uniform properties respectively, and experimental results are used to illustrate and complement the theoretical findings :

Example 4. **(1a)** *The convex function $f : x \mapsto |x|^p$ with $p > 1$ satisfies the NŁ inequality with $\xi = 1/p$ and the NS property with $\beta(x) \le c_1 \cdot \delta(x)^{1-2\xi}$.*

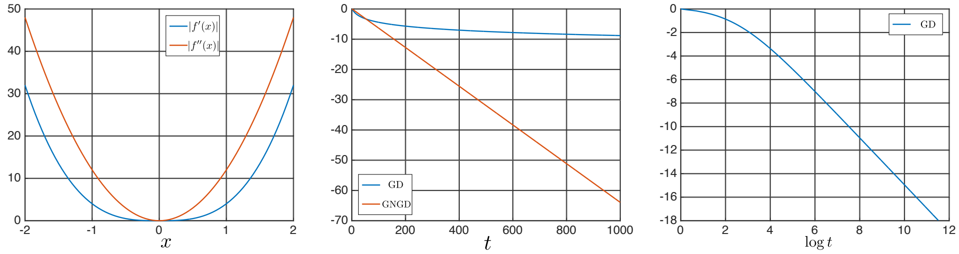
For $p > 1$, f is differentiable, and we have,

$$|f'(x)| = \left|p \cdot |x|^{p-1} \cdot \operatorname{sign}\{x\}\right| = p \cdot \left(|x|^p\right)^{\frac{p-1}{p}} = p \cdot (f(x) - f(0))^{1-\frac{1}{p}}, \tag{4.5}$$

which means f satisfies NŁ inequality with $\xi = 1/p$. On the other hand, the Hessian of f is,

$$|f''(x)| = \left|p \cdot (p-1) \cdot |x|^{p-2}\right| = p \cdot (p-1) \cdot \left(|x|^p\right)^{\frac{p-2}{p}} = p \cdot (p-1) \cdot (f(x) - f(0))^{1-\frac{2}{p}}.$$

*Hence for $p > 2$, the function $f : x \mapsto |x|^p$ with $p > 2$ satisfies the conditions in **(1a)**. Here in Fig. 4.1 is the simulation result of GD and GNGD on the function $f : x \mapsto |x|^4$.*



(a) gradient and Hessian (b) $\log \delta(x_t)$ (c) $\log \delta(x_t)$, slope ≈ -1.9996

Figure 4.1: GD and GNGD on $f : x \mapsto |x|^p$, $p = 4$.

*Subfigure (c) shows that the standard GD with constant learning rate $\eta = 0.01$ achieves sublinear rate about $O(1/t^2)$, while subfigure (b) shows that GNGD with $\eta = 0.01$ enjoys linear rate $O(e^{-c \cdot t})$, verifying the result in **(1a)**.*

(1b) *Consider maximizing the expected reward,*

$$f(\theta) = \pi_\theta^\top r, \tag{4.6}$$

where $\pi_\theta = \operatorname{softmax}(\theta)$ and $\theta \in \mathbb{R}^K$. According to Lemma 1, we have,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r, \tag{4.7}$$

which means f satisfies NŁ inequality with $\xi = 0$. As shown in Lemma 2, we have $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$. Therefore, $\beta(\theta) \le 3 \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$.

(2a) *Consider minimizing the function* $f : \mathbb{R} \to \mathbb{R}$,

$$f(\theta) = \begin{cases} 2 \cdot (\pi_\theta - \pi_{\theta^*})^2, & \text{if } |\pi_\theta - \pi_{\theta^*}| \le 0.2, \\ 25 \cdot (\pi_\theta - \pi_{\theta^*})^4 + 0.04, & \text{otherwise} \end{cases} \tag{4.8}$$

where $\theta \in \mathbb{R}$, $\theta^* = 0$, *and* π_θ *is defined as,*

$$\pi_\theta = \sigma(\theta) = \frac{1}{1 + e^{-\theta}}, \tag{4.9}$$

where $\sigma : \mathbb{R} \to (0, 1)$ *is the sigmoid activation. Fig. 4.2 shows the image of* f, *indicating that* f *is a non-convex function. Since* $\theta^* = 0$, *we have*

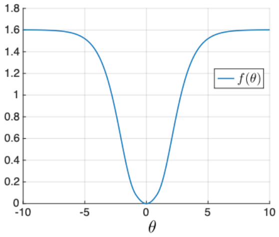


Figure 4.2: The image of f.

$\pi_{\theta^*} = 1/2$, *and for all* $|\pi_\theta - \pi_{\theta^*}| > 0.2$,

$$\begin{aligned} \left|\frac{df(\theta)}{d\theta}\right| &= \left|\frac{d\pi_\theta}{d\theta} \cdot \frac{df(\theta)}{d\pi_\theta}\right| && (4.10) \\ &= \left|\pi_\theta \cdot (1 - \pi_\theta) \cdot 100 \cdot (\pi_\theta - \pi_{\theta^*})^3\right| && (4.11) \\ &= 100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot \left[(\pi_\theta - \pi_{\theta^*})^4\right]^{\frac{3}{4}} && (4.12) \\ &= 100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot \left[f(\theta) - f(\theta^*)\right]^{1 - \frac{1}{4}}, && (4.13) \end{aligned}$$

which means f *satisfies NŁ inequality with* $\xi = 1/4 < 1/2$. *For all* $|\pi_\theta - \pi_{\theta^*}| \le 1$, *the Hessian of* f *is,*

$$\begin{aligned} \left|\frac{d^2 f(\theta)}{d\theta^2}\right| &= \left|\frac{d}{d\theta}\left\{100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*})\right\}\right| && (4.14) \\ &= \left|100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*}) \cdot (1 - 2\pi_\theta) + 100 \cdot \pi_\theta^2 \cdot (1 - \pi_\theta)^2\right| && (4.15) \\ &\le \left|100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*}) \cdot (1 - 2\pi_\theta)\right| + \frac{25}{4} && (4.16) \\ &\le \left|\frac{df(\theta)}{d\theta}\right| + \frac{25}{4}. && (4.17) \end{aligned}$$

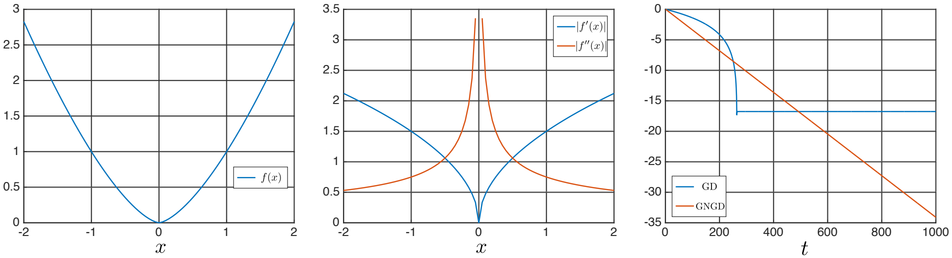
(2b) *GLM with mean-squared-error* $\mathcal{L}(\theta)$ *satisfies* $\beta(\theta)$ *NS with*

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left(\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right).$$

and (C, ξ) *NL where* $\xi = \frac{1}{2}$. *For detail, please check Chapter 6.*

(3a, 3b) *Similar to **(1a)**, for function* $f : x \mapsto |x|^{1.5}$ *satisfies the NL inequality with* $\xi = \frac{2}{3}$ *and the NS property with* $\beta(x) \leq c_1 \cdot \delta(x)^{1-2\xi}$. *Also, since* $|f'(x)| = 1.5 \cdot x^{0.5}$ *and* $|f''(x)| = 1.5 \cdot 0.5 \cdot x^{-0.5}$, f *also satisfies the condition in (3a).*

The function f *is differentiable, and the Hessian* $|f''(x)| = 1.5 \cdot (1.5 - 1) \cdot |x|^{1.5-2} \to \infty$, *as* $x \to 0$, *which indicates GD with* $\eta \in \Theta(1)$ *does not converge. Fig. 4.3(a) shows the image of* $f : x \mapsto |x|^{1.5}$. *As shown in subfigure (b), the gradient of* f *exists at* $x = 0$, *and the Hessian* $|f''(x)| \to \infty$ *as* $x \to 0$. *The results of GD with* $\eta = 0.005$ *and GNGD are presented in subfigure (c). The sub-optimality of GD update decreased for some time, and then it increased later. This is due to the Hessian is unbounded near* $x = 0$, *and thus constant learning rates cannot guarantee monotonic progresses for GD. On the other hand, GNGD with* $\eta = 0.01$ *enjoys* $O(e^{-c \cdot t})$ *convergence rate, verifying the results in the case (3a,3b).*



(a) $f : x \mapsto |x|^{1.5}$

(b) gradient and Hessian

(c) $\log \delta(x_t)$

Figure 4.3: GD and GNGD on $f : x \mapsto |x|^p$, $p = 1.5$.

4.2.2 $\Omega(1/t^2)$ Lower Bound for Convexity-Smootheness

Note that GNGD satisfies $x_{t+1} = x_1 - \sum_{i=1}^{t} \frac{\eta}{\beta(x_i)} \cdot \nabla f(x_i) \in \text{Span}\left\{x_1, \nabla f(x_1), \ldots, \nabla f(x_t)\right\}$, which is a first-order oracle Nesterov. Thus there exists a worst-case objective in the convex-smooth class that forces $\delta(x_t) \in \Omega(1/t^2)$ for $t \in O(n)$, where n

is the parameter dimension [13], [42], [43]. This is not a contradiction, since the lower bound is established by constructing a convex smooth function with a *constant* $\beta > 0$ [13], and $\beta(x) \to \beta > 0$ as $x, x' \to x^*$, e.g., Example 4 (2a). Hence, the $\Omega(1/t^2)$ result covers *some* functions.

Meanwhile, as shown in Example 4 (1a), convex functions $f : x \mapsto |x|^p$ with $p > 2$ ($\beta(x) \to \beta = 0$ as $x, x' \to x^*$) achieves linear convergence rate using GNGD, which implies that the standard convex-smooth class consists of two subclasses. One subclass admits first-order sub-linear lower bounds, while the other allows linear convergence using first-order methods. Here, note that the NS property is also divided into two sub classes, $\beta(x) \to \beta > 0$ as $x, x' \to x^*$; $\beta(x) \to \beta = 0$ as $x, x' \to x^*$. This partition also inspires geometry-aware gradient descent.

4.2.3 $\Omega(1/\sqrt{t})$ lower bound for (L_0, L_1)-smoothness.

For (2a) in Section 4.2.1, i.e., $\beta(\theta) = L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$ with $L_0, L_1 \geq 1$, standard normalized GD is subject to a $\Omega(1/\sqrt{t})$ lower bound (Zhang *et al.*) However, in Chapter 5, we will show that normalized policy gradient (PG) method achieves a linear rate of $O(e^{-c \cdot t})$.

Again, this is not a contradiction for similar reasons, the lower bound is a worst-case, and we can also construct some objective function such that GNGD meets the lower bound. With $L_0 \geq 1$, $\beta(\theta) \to L_0 > 0$ as $\theta, \theta' \to \theta^*$, the $\Omega(1/\sqrt{t})$ lower bound will hold for *some* functions in (2a) in Example 4 (where $\xi = -\frac{1}{2}$). While in Chapter 5 the objective satisfies $L_0 = 0$ and $L_1 > 0$, hence it's not in the case discussed in Zhang *et al.* This implies that a similar separation NS conditions lead to separation of convergence rates for first-order methods. Note that we're not sure whether GNGD achieves optimal worst-case convergence rate.

Chapter 5

GNGD Applied to Markov Decision Process - Geometry Aware Normalized Policy Gradient Method

As mentioned in Section 4.2.3, it can be shown that the expected return objective considered in direct policy optimization in RL achieves linear convergence rate using GNPG, and GNPG escapes plateaus faster than PG. In this chapter, the PG problem is classified into two categories - one-state MDP PG and multi-state MDP PG, and the convergence rates are shown respectively. Note that in this chapter, it can be shown that geometry-aware normalized PG is equivalent to normalized PG. For the proofs of Lemmas & Theorems in this section, please refer to Section A.1.

5.1 One-State MDP

In this section, some key insights for one-state MDPs with K actions and $\gamma = 0$ are illustrated (The problem is equivalent to K-arm bandit). The value function Eq. (2.1) reduces to expected reward $\pi_\theta^\top r$, where $r \in [0,1]^K$, $\theta \in \mathbb{R}^K$, and $\pi_\theta = \text{softmax}(\theta)$. Mei *et al.* have shown that even though $\max_\theta \pi_\theta^\top r$ is a non-concave maximization, global convergence can be achieved with a $O(1/t)$ rate using uniform smoothness and the NŁ inequality:

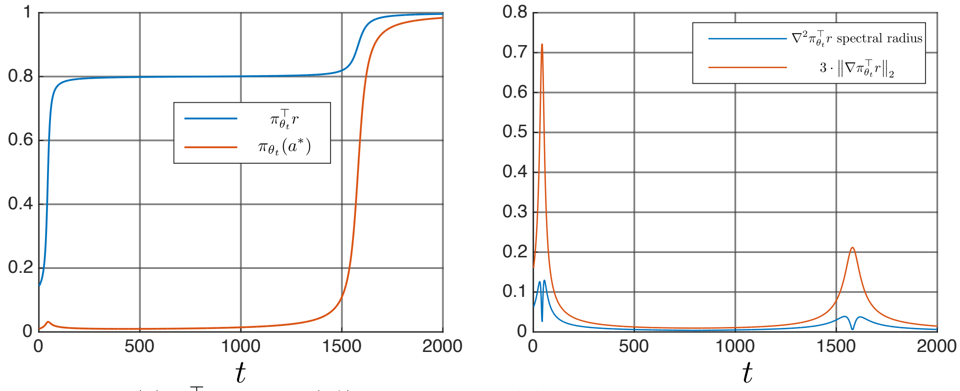
Lemma 1 (NŁ, Mei *et al.*, Lemma 3)**.** *Let a^* be the optimal action. Denote*

$\pi^* = \arg\max_{\pi\in\Delta} \pi^\top r$. *Then,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{5.1}$$

Note that Lemma 1 is not improvable in terms of the coefficients $C(\theta) = \pi_\theta(a^*)$ and $\xi = 0$ (Mei *et al.*, Remark 1 and Lemma 17).

However, the $O(1/t)$ convergence result is based on only using a *uniform* smoothness coefficient $\beta = 5/2$ (Mei *et al.*, Lemma 2), which can be significantly refined by our *uniform* smoothness Definition 5. Empirically, Mei *et al.* showed that it takes a long time for PG to escaping plateaus. To illustrate, a standard policy gradient (PG) on a 3-action one-state MDP is executed.



(a) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ (b) Hessian spectral radius and PG norm

Figure 5.1: PG results on $r = (1.0, 0.8, 0.1)^\top$.

As shown in Fig. 5.1(a), PG first goes through a long suboptimal plateau, and then eventually escapes to approach π^*. Fig. 5.1(b) presents the spectral radius of the Hessian and the PG norm $3 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$ as functions of time t.

Fig. 5.1(b) indicates that the smoothness parameter is non-uniform: it is close to zero at the suboptimal plateau and near π^*, highly aligned with the PG norm. Hence, instead of universal constant β, the PG norm characterizes the non-uniform landscape information far more precisely. This observation is formalized by proving the following key result:

Lemma 2 (NS)**.** *Denote* $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0,1]$. *For any* $r \in [0,1]^K$, $\theta \mapsto \pi_\theta^\top r$ *satisfies* $\beta(\theta_\xi)$ *non-uniform smoothness with* $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$.

Note that Lemma 2 satisfies the condition (1b) in Section 4.2.1, the NŁ parameter $\xi = 0$, and GNGD requires normalizing $\beta(\theta_\zeta)$, which is the PG norm of θ_ζ. However, ζ is unknown. Fortunately, it can be shown that, if we normalize the PG norm of θ in each iteration, the $\beta(\theta_\zeta)$ in Lemma 2 can be upper bounded by $\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$, given the learning rate η is small enough:

Lemma 3. *Let* $\theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$. *Denote* $\theta_\zeta := \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0, 1]$. *We have, for all* $\eta \in (0, 1/3)$,

$$\left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \leq \frac{1}{1 - 3\eta} \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2. \tag{5.2}$$

To finish building the convergence guarantee, it's also necessary to show that the NŁ coefficient $\pi_\theta(a^*)$ is bounded away from 0, which provides constants in the convergence rate results.

Lemma 4 (Non-vanishing NŁ coefficient)**.** *Using normalized policy gradient method, we have* $\inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$.

Hence, the non-concave function $\pi_\theta^\top r$ satisfies the requirements (1b) in Section 4.2.1 with $\xi = 0$ *in each iteration of normalized PG*[1]:

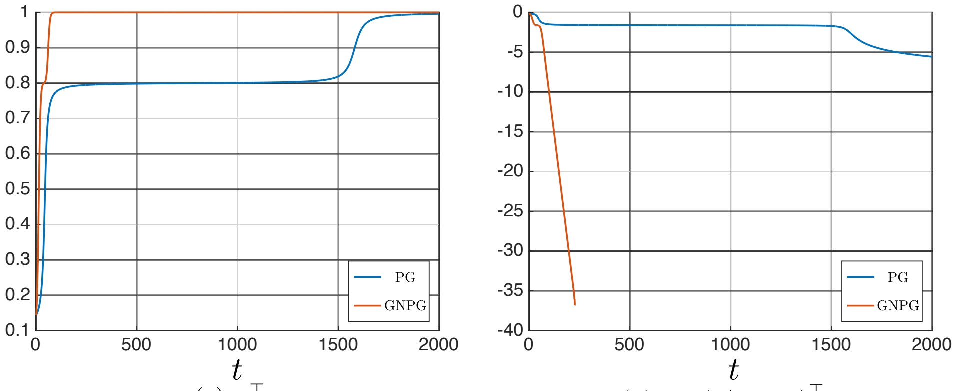
Lemmas 2 and 3 show that in each iteration of normalized PG, the NS coefficient $\beta(\theta_{\zeta_t}) \leq c_1 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$, while Lemmas 1 and 4 guarantee $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \geq c_2 \cdot (\pi^* - \pi_{\theta_t})^\top r$. Combining Lemmas 1 to 4, the global linear convergence rate $O(e^{-c \cdot t})$ of normalized PG could be shown.

Theorem 1. *Using normalized PG updates :* $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$, *with* $\eta = 1/6$, *for all* $t \geq 1$, *we have,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq e^{-\frac{c \cdot (t-1)}{12}} \cdot (\pi^* - \pi_{\theta_1})^\top r, \tag{5.3}$$

where $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ *is from Lemma 4, and c is a constant that depends on* r *and* θ_1*, but not on the time* t.

[1] This essentially means we prove that a uniform Łojasiewicz inequality holds for the entire sequence $\{\theta_t\}_{t \geq 1}$, but this does not imply that the NŁ condition is unnecessary. As shown in (Mei *et al.*, Remark 1), Łojasiewicz-type inequalities with constant $C > 0$ cannot hold. It can only become uniform after specifying an initialization θ_1 and an algorithm (in this case, PG). Otherwise, uniform Łojasiewicz cannot hold since initialization can make the NŁ coefficient $\pi_\theta(a^*)$ arbitrarily close to 0.



(a) $\pi_{\theta_t}^{\top} r$

(b) $\log\left(\pi^* - \pi_{\theta_t}\right)^{\top} r$

Figure 5.2: PG and GNPG on $r = (1.0, 0.8, 0.1)^{\top}$.

PG and GNPG are compared on the one-state MDP problem as shown in Fig. 5.2. Fig. 5.2(a) shows that GNPG escapes from the sub-optimal plateau significantly faster than PG, while Fig. 5.2(b) shows that GNPG follows linear convergence $O(e^{-c \cdot t})$ of sub-optmality, verifying the theoretical results.

5.2 Geometry-aware Normalized PG (GNPG)

GNPG could be generalized from one-state to finite MDPs on value function, and it can be shown that GNPG is equivalent to NPG in this case.[2] Here's the algorithm :

Algorithm 1 Geometry-aware Normalized Policy Gradient

```
Input: Learning rate $\eta > 0$.
Initialize parameter $\theta_1(s, a)$ for all $(s, a)$.
while $t \geq 1$ do
    $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2.$
end while
```

5.3 Multi-State MDP

For general finite MDPs, "sufficient exploration" for the initial state distribution μ is a widely used assumption which is also adapted in literature [Agarwal *et al.*, Mei *et al.*].

[2] We use GNPG as the name of Algorithm 1, since NPG is usually used to refer to the natural PG algorithm in RL literature.

Assumption 1 (Sufficient exploration)**.** *The initial state distribution satisfies* $\min_s \mu(s) > 0$.

Given Assumption 1, Agarwal *et al.* proved asymptotic global convergence for PG on the non-concave $\max_\theta V^{\pi_\theta}(\rho)$ problem, while Mei *et al.* improved this to a $O(1/t)$ rate using a combination of uniform smoothness and the following NŁ inequality that generalizes Lemma 1.

Lemma 5 (NŁ, Mei *et al.*, Lemma 8)**.** *Denote* $S \coloneqq |\mathcal{S}|$ *as the total number of states. We have,* $\forall\ \theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$*,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S}\cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left(V^*(\rho) - V^{\pi_\theta}(\rho)\right),$$

where $a^*(s)$ *is the action that* π^* *selects in state* s*.*

Note that here the NŁ degree $\xi = 0$ is not improvable [Mei *et al.*, Lemma 28], but the convergence rate could still be improved from the *non-uniform* smoothness side.

In one-state MDPs with $S = 1$, Lemma 5 recovers Lemma 1 with the same NŁ coefficient $C(\theta) = \pi_\theta(a^*)$, indicating that $C(\theta)$ in Lemma 5 might also be unimprovable. On the other hand, the uniform smoothness considered in [Agarwal *et al.*, Mei *et al.*] $\beta = 8/(1-\gamma)^3$ is too conservative, particularly when γ is close to 1. It can also be shown that for multi-state MDP, Lemma 2 could be generalized, i.e., the policy value also satisfies a stronger NS property, with the NS coefficient being the PG norm.

Lemma 6 (NS)**.** *Let Assumption 1 hold and denote* $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0, 1]$*.* $\theta \mapsto V^{\pi_\theta}(\mu)$ *satisfies* $\beta(\theta_\zeta)$ *non-uniform smoothness with*

$$\beta(\theta_\zeta) = \left[3 + \frac{2\cdot (C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta} \right\|_2,$$

where $C_\infty \coloneqq \max_\pi \left\| \frac{d_\mu^\pi}{\mu} \right\|_\infty \leq \frac{1}{\min_s \mu(s)} < \infty$*.*

In one-state MDPs with $\gamma = 0$ and $S = 1$, we have $C_\infty = 1 - \gamma$. Thus Lemma 6 reduces to Lemma 2 with the same NS coefficient $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$. Similar to Lemma 3, if we use Algorithm 1 with small enough learning rate, then $\beta(\theta_\zeta)$ in Lemma 6 is upper bounded by the PG norm of θ:

Lemma 7. *Let* $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma+4\cdot(C_\infty-(1-\gamma))}\cdot\frac{1}{\sqrt{S}}$ *and* $\theta' = \theta+\eta\cdot\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\Big/\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2$. *Denote* $\theta_\zeta \coloneqq \theta + \zeta\cdot(\theta'-\theta)$ *with some* $\zeta\in[0,1]$. *We have,*

$$\left\|\frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial\theta_\zeta}\right\|_2 \le 2\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2. \tag{5.4}$$

Similar to Lemma 4, it can also be shown that the NŁ coefficient $\min_s \pi_\theta(a^*(s)|s)$ in Lemma 5 is lower bounded away from 0:

Lemma 8 (Non-vanishing NŁ coefficient)**.** *Let Assumption 1 hold. We have,* $c \coloneqq \inf_{s\in\mathcal{S},t\ge1}\pi_{\theta_t}(a^*(s)|s) > 0$*, where* $\{\theta_t\}_{t\ge1}$ *is generated by Algorithm 1.*

Note that the non-concave function $V^{\pi_\theta}(\rho)$ satisfies requirements (1b) in Section 4.2.1 with $\xi = 0$ *in each iteration of Algorithm 1.* Therefore, combining Lemmas 5 to 8, the global linear convergence rate $O(e^{-c\cdot t})$ of Algorithm 1 could be shown.

Theorem 2. *Let Assumption 1 hold and let* $\{\theta_t\}_{t\ge1}$ *be generated using Algorithm 1 with* $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma+4\cdot(C_\infty-(1-\gamma))}\cdot\frac{1}{\sqrt{S}}$*, where* $C_\infty \coloneqq \max_\pi\left\|\frac{d_\mu^\pi}{\mu}\right\|_\infty$*. Denote* $C'_\infty \coloneqq \max_\pi\left\|\frac{d_\rho^\pi}{\mu}\right\|_\infty$*. Let* c *be the positive constant from Lemma 8. We have, for all* $t\ge1$*,*

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \le \frac{(V^*(\mu)-V^{\pi_{\theta_1}}(\mu))\cdot C'_\infty}{1-\gamma}\cdot e^{-C\cdot(t-1)},$$

where $C = \frac{(1-\gamma)^2\cdot\gamma\cdot c}{12\cdot(1-\gamma)\cdot\gamma+8\cdot(C_\infty-(1-\gamma))}\cdot\frac{1}{S}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1}$.

Not only the $O(e^{-c\cdot t})$ rate in Theorem 2 is faster than $O(1/t)$ for standard PG without normalization, but also the constant is better than the standard PG as shown in [Mei *et al.*, Theorem 4]. The strictly better dependence c ($\gg c^2$ in PG) is related to faster escaping plateaus as shown later (Mei *et al.*).

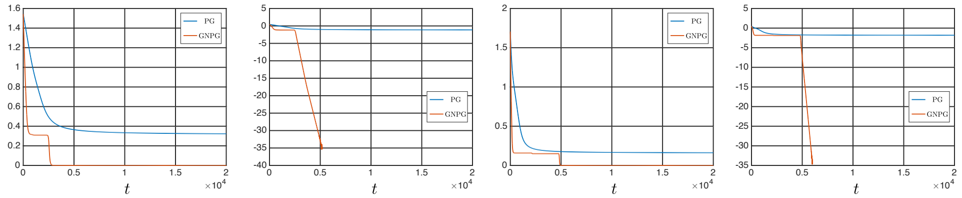
Standard softmax PG with bounded learning rate satisfies $\Omega(1/t)$ lower bound (Mei *et al.*), which is consistent with case (1) in Section 4.2.1. Our proposed Algorithm 1 achieves linear convergence rates, indicating that the adaptive update stepsize $\eta/\left\|\nabla V^{\pi_{\theta_t}}(\rho)\right\|_2$ is asymptotically unbounded, since $\left\|\nabla V^{\pi_{\theta_t}}(\rho)\right\|_2 \to 0$ as $t\to\infty$.

PG and GNPG are compared on the multi-state MDP. The environment is a synthetic tree with height h and branching factor ℓ. The total number of states is

$$S = \sum_{i=0}^{h-1} \ell^i. \tag{5.5}$$

The discount factor $\gamma = 0.99$, $\mu = \rho$, where $\rho(s_0) = 1$ for the root state s_0.

Fig. 5.3 (a) and (b) show the results for $h = \ell = 4$, and $S = 85$. The learning rate is $\eta = 0.02$ for PG and GNPG. Subfigures (c) and (d) show the results for $h = 5$ and $\ell = 4$, and $S = 341$. The learning rate is $\eta = 0.05$ for both PG and GNPG. Subfigures (a) and (c) show that GNPG escapes from the sub-optimal plateau significantly faster than PG; While subfigures (b) and (d) show that GNPG follows linear convergence $O(e^{-c\cdot t})$ of sub-optmality, verifying the theoretical results.



(a) $V^*(\rho) - V^{\pi_{\theta_t}}(\rho)$ (b) $\log\left(V^*(\rho) - V^{\pi_{\theta_t}}(\rho)\right)$ (c) $V^*(\rho) - V^{\pi_{\theta_t}}(\rho)$ (d) $\log\left(V^*(\rho) - V^{\pi_{\theta_t}}(\rho)\right)$

Figure 5.3: Results for PG and GNPG on tree MDPs. In (a) and (b), $S = 85$. In (c) and (d), $S = 341$.

Note that the conclusion of GNPG escapes plateaus faster than PG ($c \gg c^2$) arises from upper bounds (Theorem 2 and [Mei *et al.*, Theorem 4]), and is also supported by empirical evidence as shown in Figs. 5.2 and 5.3. In fact, there exists a lower bound that shows c cannot be removed for PG [Mei *et al.*, Theorem 1] under one-state MDP settings. For finite MDPs, Li *et al.* show that for softmax PG (without normalization), c can be very small in terms of the number of states. It remains open to consider whether the lower bound of c is reasonably large for GNPG.

To the best of my knowledge, existing PG variants can achieve linear convergence $O(e^{-c\cdot t})$ only if at least one of the following techniques is used : (a) **regularization**; Mei *et al.* prove that entropy regularized PG enjoys $O(e^{-c\cdot t})$ convergence toward the regularized optimal policy. (b) **natural gradient**; Cen *et al.* prove that entropy regularized natural PG achieves linear convergence. (c) **exact line-search**; Bhandari and Russo prove that without parameterization, PG variants with exact line-search achieve linear rates by approximating policy iteration.

Among the above techniques, regularization changes the problem to regularized MDPs, so this technique is not a direct one for solving the problem. Natural PG and line-search require solving expensive optimization problems to do updates, since each update is an arg max. On the contrary, Algorithm 1 enjoys global $O(e^{-c\cdot t})$ rate while preserving those two superiorities : *(i)* without using regularization, since Algorithm 1 directly works on the original MDPs; *(ii)* without solving optimization problems in each iteration, and the normalized PG update is cheap.

Chapter 6

GNGD Applied to Generalized Linear Model

In this section, generalized linear model (GLM) with quasi-maximum likelihood estimate (quasi-MLE), which applied widely in supervised learning is investigated. It can be shown that the mean squared error (MSE) of GLM (Hazan *et al.*) satisfies the case (2) in Section 4.2.1 with $\xi = 1/2$. Hence, both GD and GNGD achieve global linear convergence rates $O(e^{-c \cdot t})$, significantly improving the best existing results of $O(1/\sqrt{t})$ (Hazan *et al.*), experimental results verify the result. And GNGD escapes plateaus faster than GD. In this section, new understandings of using normalization in GLM based on non-uniform analysis is provided.

6.1 Settings and Convergence Rate of NGD

Given a training data set $\mathcal{D} = \{(x_i, y_i)\}_{i \in [N]}$ of size N, there is a feature map $x_i \mapsto \phi(x_i) \in \mathbb{R}^d$ for each pair $(x_i, y_i) \in \mathcal{D}$. Denote $\phi_i \coloneqq \phi(x_i)$ for conciseness. For each data point x_i, correspondingly, $y_i \in [0, 1]$ is the ground truth likelihood. Following Hazan *et al.*, the model is parameterized by a weight vector $\theta \in \mathbb{R}^d$ as ,

$$\pi_i = \sigma(\phi_i^\top \theta) = \frac{1}{1 + \exp\{-\phi_i^\top \theta\}}, \tag{6.1}$$

where $\sigma : \mathbb{R} \to (0,1)$ is the sigmoid activation. The problem is to minimize the mean squared error (MSE),

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta \in \mathbb{R}^d} \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - y_i)^2. \tag{6.2}$$

Assume $y_i = \pi_i^* := \sigma(\phi_i^\top \theta^*)$, where $\theta^* \in \mathbb{R}^d$, and $\|\theta^*\|_2 < \infty$, which means the target y_i is realizable and non-deterministic. According to Hazan *et al.*, the MSE in Eq. (6.2) is not quasi-convex (thus not convex). But Hazan *et al.* show that Eq. (6.2) satisfies a weaker Strictly-Locally-Quasi-Convex (SLQC) property.

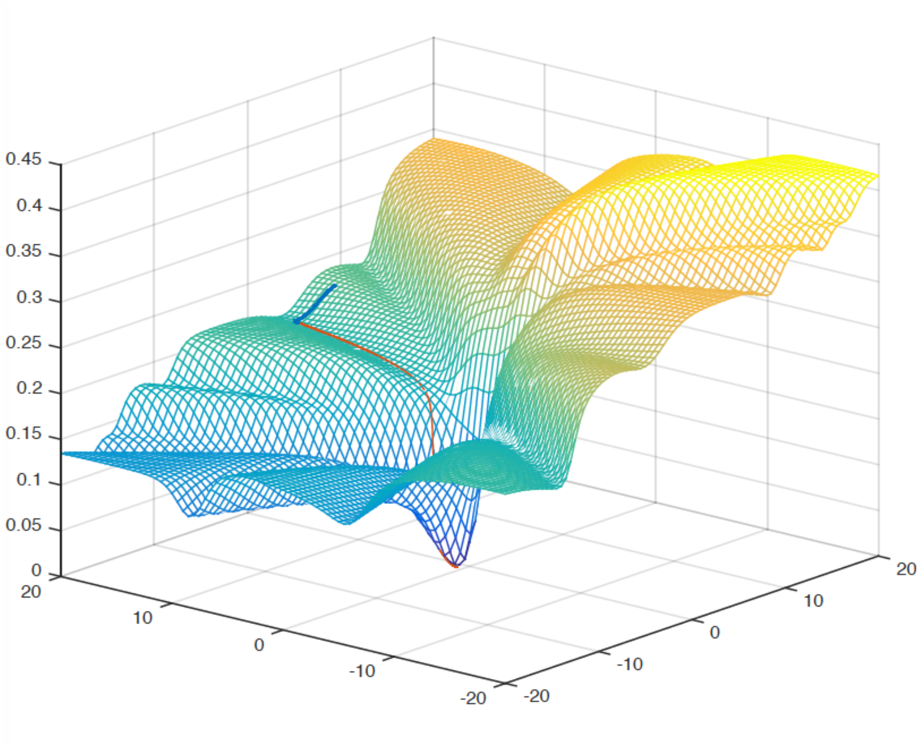


Figure 6.1: MSE Landscape in GLM

Fig. 6.1 visualizes the mean squared error (MSE) of a generalized linear model, which is non-convex and highly non-uniform.

Based on the SlQC property, Hazan *et al.* prove the following convergence result for NGD:

Theorem 3 (Hazan *et al.*). *With diminishing learning rate $\eta_t \in \Theta(1/\sqrt{t})$, the normalized gradient descent (NGD) update $\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \Big/ \left\| \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \right\|_2$ satisfies,*

$$\delta(\theta_t) \coloneqq \mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \in O(1/\sqrt{t}), \tag{6.3}$$

where $\theta^* \coloneqq \arg\min_\theta \mathcal{L}(\theta)$ is the global optimal solution.

Based on the $O(1/\sqrt{t})$ rate for NGD in Theorem 3, Hazan *et al.* propose to normalize gradient norm in MSE minimization. However, there is no lower bound for other methods including GD on GLM, and thus it is not clear if there exists a faster rate for GLM optimization.

6.2 Non-uniform Analysis

By non-uniform analysis of NS and NŁ, it can be shown that both GD and GNGD actually achieve much faster rates of $O(e^{-c \cdot t})$.

Firstly, it can be shown that the MSE in GLM satisfies a new NŁ inequality with $\xi = 1/2$:

Lemma 9 (NŁ). *Denote $u(\theta) \coloneqq \min_{i \in [N]} \left\{ \pi_i \cdot (1 - \pi_i) \right\}$, and $v \coloneqq \min_{i \in [N]} \left\{ \pi_i^* \cdot (1 - \pi_i^*) \right\}$. We have,*

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq C(\theta, \phi) \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}, \tag{6.4}$$

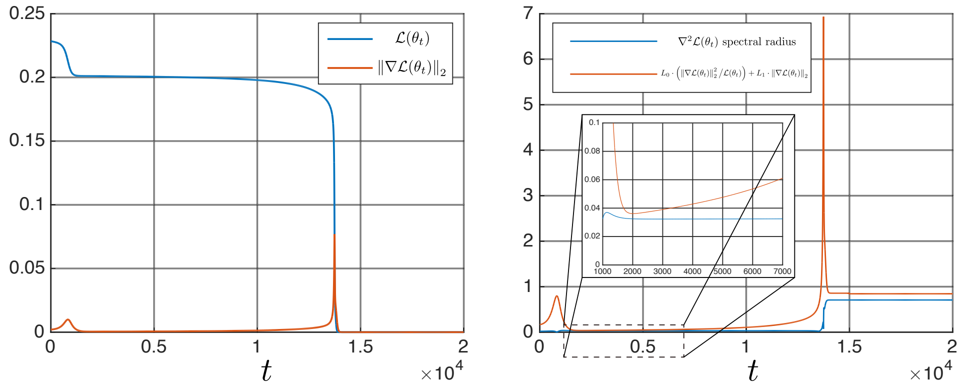
holds for all $\theta \in \mathbb{R}^d$, where

$$C(\theta, \phi) = 8 \cdot u(\theta) \cdot \min \left\{ u(\theta), v \right\} \cdot \sqrt{\lambda_\phi}, \tag{6.5}$$

and λ_ϕ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$.

Note that this NŁ lemma is for GLM with realizable and non-deterministic target. It is not clear if results similar to Lemma 9 hold without assuming: *(i)* realizable optimal prediction $y_i = \pi_i^* \coloneqq \sigma(\phi_i^\top \theta^*)$; *(ii)* non-deterministic optimal prediction $\left\| \theta^* \right\|_2 < \infty$. It's still an open question to study non-uniformity of GLM without the above assumptions.

In Lemma 9, λ_ϕ is determined by the feature ϕ. By Eq. (6.5) and definition of $u(\theta)$, when π_i is near deterministic, the gradient is vanishing, which is consistent with the fact that the sigmoid saturates and provides uninformative gradient as the parameter magnitude becomes large.



(a) $\mathcal{L}(\theta_t)$ and $\|\nabla \mathcal{L}(\theta_t)\|_2$

(b) Hessian spectral radius and NS

Figure 6.2: Experiments on GLM using GD.

Experiments are done by running GD on one GLM model with $N = 10$ and $d = 2$. As shown in Fig. 6.2, the gradient norm $\|\nabla \mathcal{L}(\theta_t)\|_2$ is close to zero at plateaus and near optimum. However, unlike the PG, the spectral radius of the Hessian $\nabla^2 \mathcal{L}(\theta_t)$ is only close to zero at plateaus, while it approaches positive constant near optimum. This indicates that unlike Lemmas 2 and 6, the spectral radius of Hessian of GLM is not simply bounded by the gradient norm. With some calculations, the following NS results can be shown:

Lemma 10 (Smoothness and NS)**.** *$\mathcal{L}(\theta)$ satisfies β smoothness with*

$$\beta = \frac{3}{8} \cdot \max_{i \in [N]} \|\phi_i\|_2^2, \tag{6.6}$$

and $\beta(\theta)$ NS with

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left(\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right).$$

The non-uniform smoothness of GLM satisfies the case (2) in Section 4.2.1 with $\xi = 1/2$. Combining Lemmas 9 and 10, the global linear convergence result can be shown:

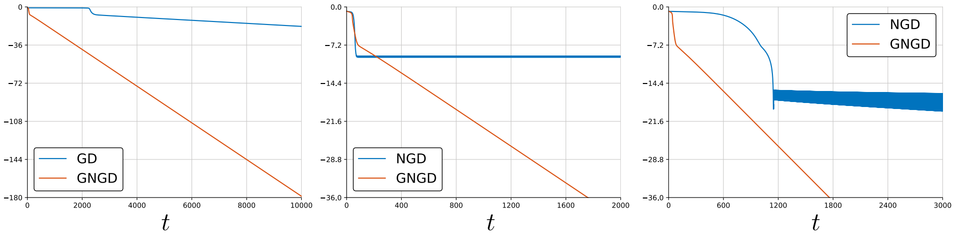
Theorem 4. *With $\eta = 1/\beta$, GD update satisfies for all $t \geq 1$, $\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C^2 \cdot (t-1)}$. With $\eta \in \Theta(1)$, GNGD update satisfies for all $t \geq 1$, $\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C \cdot (t-1)}$, where $C \in (0, 1)$, i.e., GNGD is strictly faster than GD.*

In Theorem 4, $C = \inf_{t \geq 1} C(\theta_t, \phi)$ is very close to zero if π_i is near deterministic, and GD suffers sub-optimality plateaus as shown in Fig. 6.1. GNGD has strictly (orders of magnitudes) better constant dependence $C \gg C^2$, and escapes plateaus significantly faster than GD. Intuitively, for the GLM in Fig. 6.1, C is lower bounded reasonably if θ_1 is initialized within some finite distance of the central valley containing θ^*.

By non-uniform analysis of the NŁ and NS properties (Lemmas 9 and 10), there's a new understanding of using normalization in GLM: *(i)* **First**, using standard NGD (Hazan *et al.*) for all $t \geq 1$ is not a good choice. By examining the asymptotic behaviour as $\theta \to \theta^*$, it's obvious that $\beta(\theta) \to \beta > 0$. However, the normalization of gradient norm in standard NGD gives incremental updates with adaptive stepsize $\to \infty$, which implies that NGD does not converge. To guarantee convergence, it is necessary to use $\eta_t \to 0$, which counteracts normalization and slows down the learning, since it could be not easy to find a suitable learning rate scheme. This is consistent with the $O(e^{-c \cdot t})$ result for GD with $\eta > 0$ and without normalization in Theorem 4. *(ii)* **Second**, using geometry-aware normalization $\beta(\theta_t)$ is a better choice than normalizing the gradient norm $\left\| \nabla \mathcal{L}(\theta_t) \right\|_2$. Later in this section, *both the asymptotic and the early-stage behaviours* are investigated using NS-NŁ. Since $\beta(\theta_t) \to \beta > 0$ asymptotically, GNGD is approaching GD as $\theta_t \to \theta^*$, which makes GNGD enjoy the same $O(e^{-c \cdot t})$ rate. On the other hand, when θ_t is far from θ^*, the NS parameter is of similar scale as gradient norm, i.e., $\beta(\theta_t) \leq c \cdot \left\| \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \right\|_2$, hence by normalizing $\beta(\theta_t)$, GNGD accelerates the optimization a lot. Then GNGD is close to NGD, but guarantees strictly better progresses than GD.

This is because of the progress of GNGD in each iteration at this time is of scale $\|\nabla\mathcal{L}(\theta_t)\|_2$, while the progress of GD is of scale $\|\nabla\mathcal{L}(\theta_t)\|_2^2$, and on plateaus, GNGD escapes plateaus faster. Using NŁ of Lemma 9, GNGD will have strictly better constant dependence C than C^2 in GD.

GD, NGD (Hazan *et al.*), and GNGD on GLM are compared by experiment. The performances are as shown in Fig. 6.3.



(a) $\log\delta(\theta_t)$ (b) $\log\delta(\theta_t)$ (c) $\log\delta(\theta_t)$

Figure 6.3: Convergence rates for GD, NGD, and GNGD on GLM.

Subfigure (a) presents the convergence rate of GD with $\eta = 0.09$ and GNGD with $\eta = 0.09$. As shown in the figure, both GD and GNGD achieve linear $O(e^{-c\cdot t})$ rates, verifying Theorem 4. GD stuck at the plateaus at the early-stage optimization, which verifies the explanations after Theorem 4. On the other hand, the slopes indicate that GNGD converges strictly faster than GD, which verifies the constant dependences ($C \geq C^2$) in Theorem 4. Subfigure (b) shows that standard NGD (Hazan *et al.*) with constant learning rate $\eta = 0.09$ does not converge. The NGD update keeps oscillating, indicating that using standard normalization for all $t \geq 1$ is not a good idea. Subfigure (c) presents the NGD using adaptive learning rate $\eta_t = \frac{0.09}{\sqrt{t}}$, which has faster convergence than NGD with constant η. However, GNGD with constant learning rate $\eta = 0.09$ still significantly outperforms NGD with $\eta_t = \frac{0.09}{\sqrt{t}}$, verifying the $O(e^{-c\cdot t})$ in Theorem 4 and $O(1/\sqrt{t})$ in Theorem 3.

Chapter 7

Robust Risk Averse Reinforcement Learning

This chapter is mainly situated in the broad area of safe RL [20]. In particular, a subgroup of works aims to improve robustness of learned policies by assuming two opposing learning processes: one that aims to disturb the most and another one that tries to control the perturbations [40]. This approach has been recently adapted to work with neural networks in the context of deep RL [46]. Moreover, Risk-Averse Robust Adversarial Reinforcement Learning (RARL) [45] extended this idea by combining with Averaged DQN [3], an algorithm that proposes averaging the previous k estimates to stabilize the training process. RARL trains two agents – protagonist and adversary in parallel, and the goal for those two agents are respectively to maximize/minimize the expected return as well as minimize/maximize the variance of expected return. RARL showed good experimental results in self-driving car simulations on the variance reduction and robustness. Multi-agent Q-learning [23] is useful for finding the optimal strategy when there exists a unique Nash equilibrium in general sum stochastic games, and this approach could also be used in adversarial RL.

Wainwright (2019) proposed a variance reduction Q-learning algorithm (V-QL) which can be seen as a variant of the SVRG algorithm in stochastic optimization [25]. Given an algorithm that converges to Q^*, one of its iterates $\bar{Q}$ could be used as a proxy for Q^*, and then recenter the ordinary Q-learning updates by a quantity $-\hat{\mathcal{T}}_k(\bar{Q}) + \mathcal{T}(\bar{Q})$, where $\hat{\mathcal{T}}_k$ is an empirical Bellman operator, $\mathcal{T}$ is the population Bellman operator, which is not computable, but an unbiased approximation of it could be used instead. This algorithm is theoretically shown to be convergent and enjoys minimax optimality up to a logarithmic factor.

Lastly, another group of works proposed the use of risk-averse objective functions [38] with the Q-learning algorithm. These ideas are highly related to the novel proposed algorithms in this chapter.

7.1 Risk Averse Q Learning

Shen *et al.* (2014) proposed a Q learning algorithm (Algorithm 2) that is theoretically shown to converge to the optimal of a risk-sensitive objective function, the training scheme is the same as Q learning, except that in each iteration, a utility function is applied to a TD-error. [54] showed how exponential utility function and normally distributed consumption give rise to a mean variance utility function where the agent's expected utility is a linear function of his mean return and the variance of his return.

Algorithm 2 Risk-Averse Q-Learning (RAQL) [49]

```
For $\forall(s,a)$, initialize $Q(s,a) = 0$; $N(s,a) = 0$.
for $t = 1$ to $T$ do
    At state $s_t$, choose action according to the $\epsilon$-greedy strategy.
    Observe $s_t, a_t, r_t, s_{t+1}$
    $N(s_t,a_t) = N(s_t,a_t) + 1$
    Set learning rate $\alpha_t = \frac{1}{N(s_t,a_t)}$
    Update Q :

        $Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha_t(s_t,a_t) \cdot \left[u\left(r_t + \gamma \ \cdot \max_a Q_t(s_{t+1},a) - Q_t(s_t,a_t)\right) - x_0\right] \quad (7.1)$

    where $u$ is a utility function, here we use $u(x) = -e^{\beta x}$ where $-1 < \beta < 0$; $x_0 = -1$
end for
Return Q.
```

Since the goal is to optimize the expected return as well as minimizing the variance of the expected return, an expected utility of the return could be used as the objective function instead [16]:

$$\tilde{J}_\pi = \frac{1}{\beta}\log \mathbb{E}_\pi \left[exp\left(\beta \sum_{t=0}^{\infty} \gamma^t r_t\right)\right]. \tag{7.2}$$

By a straightforward Taylor expansion, Eq. (7.2) yields

$$\begin{aligned}
&\frac{1}{\beta}\log \mathbb{E}_\pi \left[exp\left(\beta \sum_{t=0}^{\infty} \gamma^t r_t\right)\right] \\
&\approx \frac{1}{\beta}\log\left(1+\mathbb{E}_\pi\left[\beta\sum_{t=0}^{\infty}\gamma^t r_t + \frac{1}{2}\left(\beta\sum_{t=0}^{\infty}\gamma^t r_t\right)^2\right]\right) \\
&\approx \frac{1}{\beta}\left(\mathbb{E}_\pi\left[\beta\sum_{t=0}^{\infty}\gamma^t r_t + \frac{1}{2}\left(\beta\sum_{t=0}^{\infty}\gamma^t r_t\right)^2\right] - \frac{1}{2}\left(\mathbb{E}_\pi\left[\beta\sum_{t=0}^{\infty}\gamma^t r_t + \frac{1}{2}\left(\beta\sum_{t=0}^{\infty}\gamma^t r_t\right)^2\right]\right)^2\right) \\
&= \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t] + \frac{\beta}{2}\mathbb{E}_\pi\left[\left(\sum_{t=0}^{\infty}\gamma^t r_t\right)^2\right] - \frac{1}{2\beta}\left(\mathbb{E}_\pi\left[\beta\sum_{t=0}^{\infty}\gamma^t r_t + O(\beta^2)\right]\right)^2 \\
&= \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t] + \frac{\beta}{2}\mathbb{E}_\pi\left[\left(\sum_{t=0}^{\infty}\gamma^t r_t\right)^2\right] - \frac{\beta}{2}\left(\mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t]\right)^2 + O(\beta^2) + O(\beta^3) \qquad (-1<\beta<0) \\
&= \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t] + \frac{\beta}{2}\mathbb{V}ar[\sum_{t=0}^{\infty}\gamma^t r_t] + O(\beta^2)
\end{aligned}$$

where when $\beta < 0$ the objective function is risk-averse, when $\beta = 0$ the objective function is risk-neutral, and when $\beta > 0$ the objective function is risk-seeking.

Shen *et al.* (2014) proved that by applying a monotonically increasing concave utility function $u(x) = -exp(\beta x)$ where $\beta < 0$ to the TD error, Algorithm 2 converges to the optimal point of Eq. (7.2). Hence, it can be shown that:

Theorem 5. *(Theorem 3.2, Shen* et al. *2014) Running Algorithm 2 from an initial Q table, $Q \to Q^*$ w.p. 1, where Q^* is the unique solution to*

$$\mathbb{E}_{s'}\left[u\left(r(s,a) + \gamma\cdot \max_a Q^*(s',a) - Q^*(s,a)\right)\right] - x_0 = 0$$

$\forall(s,a)$. Where s' is sampled from $\mathcal{T}[\cdot|s,a]$. And the corresponding policy π^* of Q^* satisfies $\tilde{J}_{\pi^*} \geq \tilde{J}_\pi \; \forall\pi$.

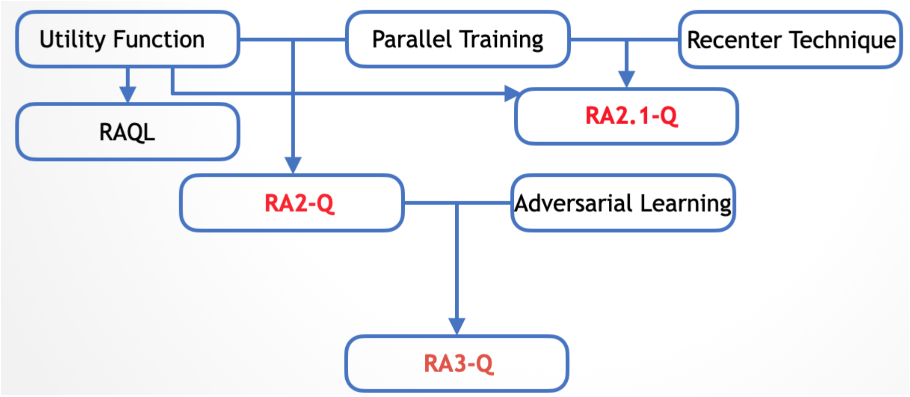


Figure 7.1: Building Blocks of Proposed Algorithms

7.2 Proposed Algorithms

In this section I'll describe my proposed algorithms continuing the results discussed in the previous sections.The two single agent algorithms RA2-Q and RA2.1-Q use a risk-averse utility function and reduce variance by training multiple Q tables in parallel. The last proposal, RA3-Q, keeps the adversarial component to improve robustness while relaxing the strong assumptions. The building blocks of those algorithms are as shown in Fig. 7.1. As a summary, Table 7.1 presents closely related works and the comparison with the novel proposed algorithms.

7.2.1 RA2-Q

Although it's already shown that RAQL converges to the optimal of risk-sensitive objective function with probability 1, the proof assumes visiting every state infinitely many times whereas the actual training time is finite. The main idea here is that we can reduce the training variance further by choosing more risk-averse actions during the finite training process.

Averaged DQN [3] reduces training variance by averaging multiple Q tables in the update. In a similar spirit, the proposed RA2-Q also trains multiple Q tables in parallel. RA2-Q trains k Q tables in parallel using Eq. (7.4) as update rule. To select more *stable* actions, RA2-Q uses the sample variance

Algorithm 3 Risk-Averse Averaged Q-Learning (RA2-Q)

```
Input : Training steps $T$; Exploration rate $\epsilon$; Number of models $k$; risk control parameter $\lambda_P$; Utility function parameter $\beta$.
Initialize $Q^i = \mathbf{0}$, $N^i = \mathbf{0}$, $\alpha^i = \mathbf{1}$ for $\forall i = 1, ..., k$.
Initialize Replay Buffer $RB = \emptyset$; Randomly sample action choosing head integers $H \in [1, k]$
for $t = 1$ to $T$ do
    $Q = Q^H$
    Compute $\hat{Q}$ by
        $$\hat{Q}(s,a) = Q(s,a) - \lambda_P \cdot \frac{\sum_{i=1}^{k} (Q^i(s,a) - \bar{Q}(s,a))^2}{k-1} \tag{7.3}$$
    where $\lambda_P > 0$ is a constant; $\bar{Q}(s,a) = \frac{1}{k}\sum_{i=1}^{k} Q^i(s,a)$
    Select action $a_t$ according to $\hat{Q}$ by applying $\epsilon$-greedy strategy.
    Execute actions and get $(s_t, a_t, r_t, s_{t+1})$, append to the replay buffer $RB = RB \cup \{(s_t, a_t, r_t, s_{t+1})\}$
    Generate mask $M \in \mathbb{R}^k \sim Poisson(1)$
    for $i = 1, ..., k$ do
        if $M_i = 1$ then
            Update $Q^i$ by
                $$Q^i(s_t,a_t) = Q^i(s_t,a_t) + \alpha^i(s_t,a_t) \cdot \left[u\left(r(s_t,a_t) + \gamma \cdot \max_a Q^i(s_{t+1},a) - Q^i(s_t,a_t)\right) - x_0\right] \tag{7.4}$$
            where $u$ is a utility function, here we use $u(x) = -e^{\beta x}$ where $-1 < \beta < 0$; $x_0 = -1$
            $N^i(s_t,a_t) = N^i(s_t,a_t) + 1$; Update learning rate $\alpha^i(s_t,a_t) = \frac{1}{N^i(s_t,a_t)}$.
        end if
    end for
    Update $H$ by randomly sampling integers from 1 to $k$.
end for
Return $\frac{1}{k}\sum_{i=1}^{k} Q^i$
```

of k Q tables as an approximation to the true variance and then compute a risk-averse $\hat{Q}$ table and select actions according to it. A detailed description is presented in Algorithm 3.

The objective function here is also Eq. (7.2), and it can be shown that Algorithm 3 also converges to the optimal.

Theorem 6. *Running Algorithm 3 for an initial Q table, then for all $i \in \{1, ..., k\}$, $Q^i \to Q^*$ w.p. 1, hence the returned table $\frac{1}{k}\sum_{i=1}^{k} Q^i \to Q^*$ w.p. 1, where Q^* is the unique solution to*

$$\mathop{\mathbb{E}}_{s'}\left[u\left(r(s,a) + \gamma \cdot \max_a Q^*(s',a) - Q^*(s,a)\right)\right] - x_0 = 0$$

for all (s,a). Where s' is sampled from $\mathcal{T}[\cdot|s,a]$. And the corresponding policy π^* of Q^* satisfies $\tilde{J}_{\pi^*} \geq \tilde{J}_{\pi}$ $\forall \pi$.

Theorem 6 follows directly from Theorem 5 (see Section A.4 for the detailed proof).

7.2.2 Variance Reduced Risk-Averse Q-Learning (RA2.1-Q)

Wainwright (2019) proposed Variance Reduced Q-learning which trains multiple Q tables in parallel and uses the averaged Q table in the update rule. It is shown that it guarantees a convergence rate which is minimax optimal. Inspired by that work, we propose our RA2.1-Q (Algorithm 4) which applies a utility function to the TD error during Q updates for the purpose of further reducing variance. To select more *stable* actions during training, we use the sample variance of k Q tables as an approximation to the true variance and then compute a risk-averse $\hat{Q}$ table and select actions according to it.

It's still not clear whether Algorithm 4 (RA2.1-Q) has a convergence guarantee, however, it obtained good empirical results (better than RAQL and RA2-Q as presented in Section 9.1). Furthermore, it could be interesting to study whether it also enjoys minimax optimality convergence rate up to a logarithmic factor as in [55].

7.2.3 Risk-Averse Adversarial Averaged Q-Learning (RA3-Q)

In complex scenarios such as financial markets, learned RL policies can be brittle. To improve robustness, algorithms could adapt ideas from adversarial learning to a multi-agent learning problem similar to [23].

In the adversarial setting, we can assume there are two learning processes happening simultaneously, a main protagonist *(P)* and an adversary *(A)*: the goal of protagonist is to maximize the total return as well as minimize the variance; the goal of adversary is to minimize the total return of protagonist as well as maximizing the variance. Here, we assume that each agent can observe its opposite's immediate reward. The process is as presented in Fig. 7.2.

Let r_t^P be the immediate reward received by protagonist at step t, and let

Algorithm 4 Variance Reduced Risk-Averse Q-Learning (RA2.1-Q)

```
Input : Training epochs $T$; Exploration rate $\epsilon$; Number of models $k$; Epoch length $K$; Recentering sample size $N$; Utility function parameter $\beta < 0$;
Initialize $\bar{Q}_0 = \mathbf{0}$; $m = 1$; $RB = \emptyset$.
for $m = 1$ to $T$ do
    Select action according to $\bar{Q}_{m-1}$ by applying $\epsilon-$greedy strategy
    Execute action and get $(s, a, r(s, a), s')$ and update the replay buffer $RB = RB \cup (s, a, r(s, a), s')$.
    for $i = 1, ..., N$ do
        Define the empirical Bellman operator $\ddot{\mathcal{T}}_i$ as
            $$\ddot{\mathcal{T}}_i(Q)(s, a) = u\left(r(s, a) + \gamma \cdot \max_{a'} Q(s_i, a')\right) - x_0$$
        where $s_i$ is randomly sampled from $\mathcal{T}[\cdot|s, a]$; $u$ is the utility function, and $u(x) = -e^{\beta x}$, $\beta < 0$ and $x_0 = -1$
    end for
    Define $\tilde{\mathcal{T}}_N(\bar{Q}_{m-1}) = \frac{1}{N}\sum_{i \in \mathcal{D}_N} \ddot{\mathcal{T}}_i(\bar{Q}_{m-1})$, where $\mathcal{D}_N$ is a collection of $N$ i.i.d. samples (i.e., matrices with samples for each state-action pair $(s, a)$ from $RB$).
    Define $Q_1 = \bar{Q}_{m-1}$.
    for $k = 1, ..., K$ do
        Compute stepsize $\lambda_k = \frac{1}{1+(1-\gamma)k}$
            $$Q_{k+1} = (1 - \lambda_k) \cdot Q_k + \lambda_k \cdot \left[\ddot{\mathcal{T}}_k(Q_k) - \ddot{\mathcal{T}}_k(\bar{Q}_{m-1}) + \tilde{\mathcal{T}}_N(\bar{Q}_{m-1})\right]. \quad (7.5)$$
        where $\ddot{\mathcal{T}}_k$ is empirical Bellman operator constructed using a sample not in $\mathcal{D}_N$, thus the random operators $\ddot{\mathcal{T}}_k$ and $\tilde{\mathcal{T}}_N$ are independent
    end for
    $\bar{Q}_m = Q_{K+1}$; $m = m + 1$
end for
Return $\bar{Q}_m$
```

r_t^A be the immediate reward received by adversary at step t. Then we choose the objective functions as follows:

The objective function for the protagonist is,

$$\tilde{J}_\pi^P = \frac{1}{\beta^P} \log \mathop{\mathbb{E}}_{\pi}\left[exp\left(\beta^P \sum_{t=0}^{\infty} \gamma^t \cdot r_t^P\right)\right] \qquad \beta^P < 0 \tag{7.6}$$

by a Taylor expansion, Eq. (7.6) yields,

$$\tilde{J}_\pi^P = \mathbb{E}\left[\sum_{t=0} \gamma^t \cdot r_t^P\right] + \frac{\beta^P}{2}\mathbb{V}ar\left[\sum_{t=0} \gamma^t \cdot r_t^P\right] + O((\beta^P)^2).$$

Similarly, the objective function for the adversary is,

$$\tilde{J}_\pi^A = \frac{1}{\beta^A} \log \mathop{\mathbb{E}}_{\pi}\left[exp\left(\beta^A \sum_{t=0}^{\infty} \gamma^t r_t^A\right)\right] \qquad \beta^A > 0 \tag{7.7}$$

and by Taylor expansion, Eq. (7.7) yields,

$$\tilde{J}_\pi^A = \mathbb{E}\left[\sum_{t=0} \gamma^t \cdot r_t^A\right] + \frac{\beta^A}{2}\mathbb{V}ar\left[\sum_{t=0} \gamma^t \cdot r_t^A\right] + O((\beta^A)^2).$$

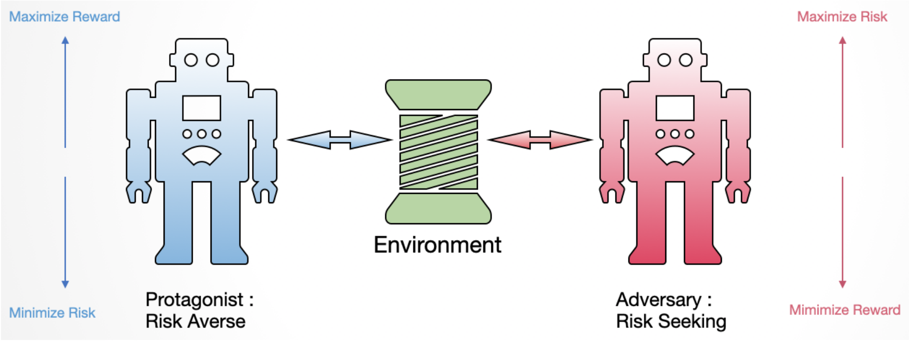


Figure 7.2: Protagonist V.S. Adversary Process in RA3-Q

Starting from the objective functions for the protagonist - Eq. (7.6), and adversary - Eq. (7.7), In order to optimize $\tilde{J}^P$ and $\tilde{J}^A$, we apply utility functions to TD errors when updating Q tables, and combining the idea of training multiple Q tables in parallel as Algorithm 3 to select actions with low variance, we get a novel Algorithm 5.

Algorithm 5 Risk-Averse Adversarial Averaged Q-Learning (RA3-Q)

Input : Training steps T; Exploration rate ϵ; Number of models k; Risk control parameters λ_P, λ_A; Utility function parameters $\beta^P < 0; \beta^A > 0$.

```
Initialize $Q_P^i(s, a_P, a_A) = 0$; $Q_A^i(s, a_P, a_A) = 0$ for $\forall i = 1, ..., k$ and $(s, a_A, a_P)$; $N = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{A}|}$;
Randomly sample action choosing head integers $H_P, H_A \in \{1, ..., k\}$.
for $t = 1$ to $T$ do
    $Q_P = Q_P^{H_P}$
    Compute $\hat{Q}_P$ by
        $\hat{Q}_P(s, a_P, a_A) = Q_P(s, a_P, a_A) - \lambda_P \cdot \frac{\sum_{i=1}^{k}(Q_P^i(s, a_P, a_A) - \bar{Q}_P(s, a_P, a_A))^2}{k-1} \qquad \lambda_P > 0 \qquad (7.8)$
    where $\bar{Q}_P(s, a_P, a_A) = \frac{1}{k}\sum_{i=1}^{k} Q_P^i(s, a_P, a_A)$
    $Q_A = Q_A^{H_A}$
    Compute $\hat{Q}_A$ by
        $\hat{Q}_A(s, a_P, a_A) = Q_A(s, a_P, a_A) + \lambda_A \cdot \frac{\sum_{i=1}^{k}(Q_A^i(s, a_P, a_A) - \bar{Q}_A(s, a_P, a_A))^2}{k-1} \qquad \lambda_A > 0 \qquad (7.9)$
    where $\bar{Q}_A(s, a_P, a_A) = \frac{1}{k}\sum_{i=1}^{k} Q_A^i(s, a_P, a_A)$
    The optimal actions $(a'_P, a'_A)$ are defined as
        $\hat{Q}_P(s_t, a'_P, a_A^0) = \max_{a_P, a_A} \hat{Q}_P(s_t, a_P, a_A) \qquad \text{for some } a_A^0 \qquad (7.10)$
        $\hat{Q}_A(s_t, a_P^0, a'_A) = \max_{a_P, a_A} \hat{Q}_A(s_t, a_P, a_A) \qquad \text{for some } a_P^0 \qquad (7.11)$
    Select actions $a_P, a_A$ according to $\hat{Q}_P, \hat{Q}_A$ by applying $\epsilon$-greedy strategy.
    Two agents respectively execute actions $a_P, a_A$ and observe $(s_t, a_P, a_A, r_t^A, r_t^P, s_{t+1})$
    Generate mask $M \in \mathbb{R}^k \sim Poisson(1)$
    $N(s_t, a_P, a_A) = N(s_t, a_P, a_A) + 1$
    $\alpha(s_t, a_P, a_A) = \frac{1}{N(s_t, a_P, a_A)}$
    for $i = 1, ..., k$ do
        if $M_i = 1$ then
            Update $Q_P^i$ by
                $Q_P^i(s_t, a_P, a_A) = Q_P^i(s_t, a_P, a_A) + \alpha(s_t, a_P, a_A) \cdot \left[u^P\left(r_t^P + \gamma \cdot \max_{a_P, a_A} Q_P^i(s_{t+1}, a_P, a_A) - Q_P^i(s_t, a_P, a_A)\right) - x_0\right] \qquad (7.12)$
            where $u^P$ is a utility function, here we use $u^P(x) = -e^{\beta^P x}$ where $-1 < \beta^P < 0$; $x_0 = -1$
        end if
    end for
    for $i = 1, ..., k$ do
        if $M_i = 1$ then
            Update $Q_A^i$ by
                $Q_A^i(s_t, a_P, a_A) = Q_A^i(s_t, a_P, a_A) + \alpha(s_t, a_P, a_A) \cdot \left[u^A\left(r_t^A + \gamma \cdot \max_{a_P, a_A} Q_A^i(s_{t+1}, a_P, a_A) - Q_A^i(s_t, a_P, a_A)\right) - x_1\right] \qquad (7.13)$
            where $u^A$ is a utility function, here we use $u(x) = e^{\beta^A \cdot x}$ where $0 < \beta^A < 1$; $x_1 = 1$
        end if
    end for
    Update $H_P$ and $H_A$ by randomly sampling integers from 1 to $k$
end for
Return $\frac{1}{k}\sum_{i=1}^{k} Q_P^i$; $\frac{1}{k}\sum_{i=1}^{k} Q_A^i$
```

Note that RA3-Q combines (i) risk-averse using utility functions (ii) variance reduction by training multiple Q tables and (iii) robustness by adversarial learning. Intuitively, as the adversary is getting stronger, the protagonist experiences harder challenges, thus enhancing robustness. Note that in the multi-agent learning scenario (when protagonist and adversary are learning simultaneously), RA3-Q does not have a convergence guarantee, however, it has several practical advantages including computational efficiency, simplicity (no strong assumptions) and more stable actions during training. For its empirical superiority, see Chapter 9. In Section A.4, I'll present a related result showing that Eq. (7.12) or Eq. (7.13) converge to optimal assuming the policy for the adversary (or protagonist) is fixed (thus, it is no longer a multi-agent learning setting).

7.3 Summarize of Risk Averse Algorithms

Table 7.1: Comparison of related algorithms. Our proposed algorithms are marked with **bold** and are described in Sections 7.2.1 to 7.2.3

Algorithm	**Description**	**Guarantees**
Risk averse Q-Learning [49]	Q-Learning with a utility function applied to TD Error in Q update	Convergence to optimal of a risk-averse objective function
Variance reduced Q-learning [55]	Use average estimation of multiple Q tables in Q-table updates to reduce variance	Convergent to optimal of expected return. Convergence rate is minimax optimal up to a logarithmic factor.
Nash Q-learning [23]	Two-agent Q-Learning in multi-agent MDP setting	Convergence to Nash equilibrium of the two-agent game (if exists)
Risk-Averse Robust Adversarial Reinforcement Learning (RARL) [45]	Q-Learning with risk-averse/risk-seeking behaviors of protagonist/adversary with multiple Q tables	No convergence guarantee
Risk-Averse Averaged Q-Learning (RA2-Q)	Q-Learning with a utility function + a more stable choice of actions with multiple Q tables	Convergence to optimal of a risk-averse objective function and reduced training variance.
Variance Reduced Risk-Averse Q-Learning (RA2.1-Q)	Use average estimation of multiple Q tables in Q updates; Apply utility function in Q updates	No convergence guarantee
Risk-Averse Multiagent Q-learning (RAM-Q)	Multi-agent Nash Q-Learning with a utility function + a risk-averse/risk-seeking behaviors of protagonist/adversary + multiple Q tables	Convergence to Nash equilibrium (if exists) of the two-agent game (with Risk-Averse/Seeking payoffs respectively)
Risk-Averse Adversarial Averaged Q-Learning (RA3-Q)	Multi-agent Q-Learning with a utility function + a risk-averse/risk-seeking behaviors of protagonist/adversary + multiple Q tables	No convergence guarantee

Chapter 8

Empirical Game Theory Analysis on Risk Sensitive Measurements

When the environment is populated by many learning agents, a way to evaluate their performance is a necessity. Empirical Game Theory (EGT) is used to address this question. In EGT, each agent is considered as a player involved in rounds of strategic interaction (games). By meta-game analysis, the superiority of each strategy could be analyzed. In this section, my contribution is to theoretically prove that the Nash-Equilibrium of risk averse meta-game is an approximation of the Nash-Equilibrium of the population game, to my knowledge, this is the first work doing this type of risk-averse analysis.

8.1 Replicator dynamics

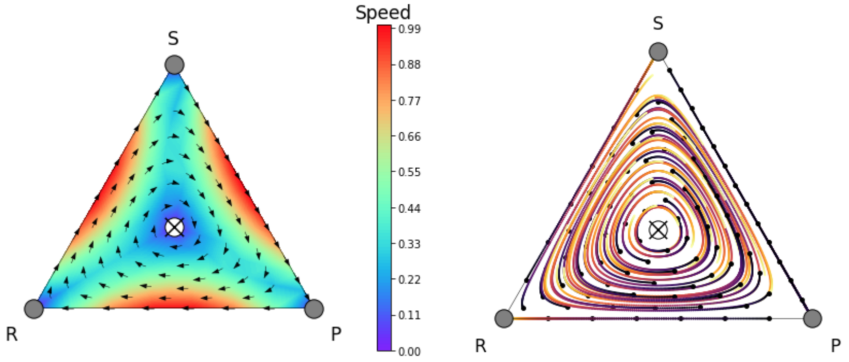
In EGT, the dominance of strategies could be visualized by plotting the meta-game payoff tables together with the replicator dynamics. A meta game payoff table could be seen as a combination of two matrices $(N|R)$, where each row N_i contains a discrete distribution of p players over k strategies, and each row yields a discrete profile $(n_{\pi_1}, ..., n_{\pi_k})$ indicating exactly how many players play each strategy with $\sum_j n_{\pi_j} = p$. A strategy profile $\mathbf{u} = \left(\frac{n_{\pi_1}}{p}, ..., \frac{n_{\pi_k}}{p}\right)$. And each row R_i captures the rewards corresponding to the rows in N.

For example, for a game A with 2 players, and 3 strategies $\{\pi_1, \pi_2, \pi_3\}$ to choose from, the meta game payoff table could be constructed as follows : In the left side of the table, we list all of the possible combinations of strategies. If there are p players and k strategies, then there are $\binom{p+k-1}{p}$ rows, hence in game A, there are 6 rows. See the tables & figures below for some concrete examples.

Once we have a meta-game payoff table and the replicator dynamics, a directional field plot is computed where arrows in the strategy space indicates the direction of flow, or change, of the population composition over the strategies. In Section 9.1, trading market experiments and results based on meta-game analysis with the performance of RAQL, RA2-Q and RA2.1-Q will be presented.

Table 8.1: Payoff Table of Rock-Paper-Scissors

N_{Rock}	N_{Paper}	$N_{Scissors}$	R_{Rock}	R_{Paper}	$R_{Scissors}$
2	0	0	0	0	0
1	1	0	-1	1	0
0	2	0	0	0	0
1	0	1	1	0	-1
0	0	2	0	0	0
0	1	1	0	-1	1



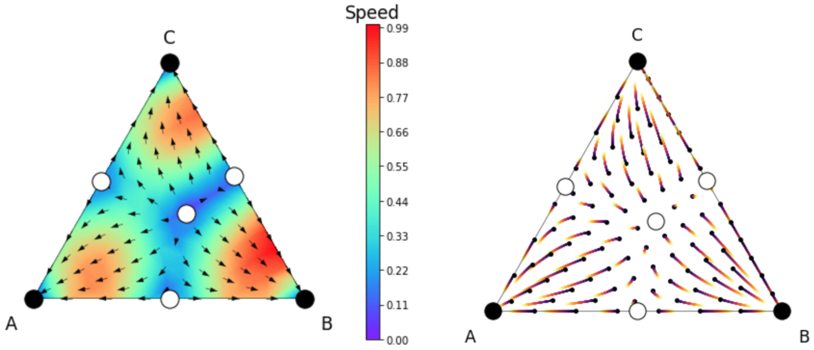
(a) (b)

Figure 8.1: Directional Field and Trajectory Plot of Rock-Paper-Scissors

The payoff table of a well-known game rock-scissors-papers is as shown in Table 8.1, its corresponding directional field and its trajectory plot are as shown in Fig. 8.1. It can be observed from Fig. 8.1 that the equilibrium of Rock-Paper-Scissors is the centroid of the strategies simplex.

Table 8.2: An example of a meta game payoff table of 2 players, 3 strategies.

N_{i1}	N_{i2}	N_{i3}	R_{i1}	R_{i2}	R_{i3}
2	0	0	0.5	0	0
1	1	0	0.3	0.7	0
0	2	0	0	0.9	0
1	0	1	0.35	0	0.45
0	0	2	0	0	0.6
0	1	1	0	0.66	0.38



(a)

(b)

Figure 8.2: Directional Field and Trajectory Plot of Table 8.2

8.2 Nash Equilibrium with risk neutral payoff

Previously, Tuyls *et al.* (2020) showed that for a game $r^i(\pi^i, ..., \pi^p) = \mathbb{E}[R^i(\pi^1, ..., \pi^p)]$, with a meta-payoff (empirical payoff) $\hat{r}^i(\pi^i, ..., \pi^p)$, the Nash Equilibrium of $\hat{r}$ is an approximation of Nash Equilibrium of r.

Lemma 11. *[53] If* $\mathbf{x}$ *is a Nash Equilibrium for the game* $\hat{r}^i(\pi^1, ..., \pi^p)$*, then it is a* 2ϵ*-Nash equilibrium for the game* $r^i(\pi^1, ..., \pi^p)$*, where* $\epsilon = \underset{\pi,i}{sup}\ |\hat{r}^i(\pi) - r^i(\pi)|$.

Lemma 11 implies that if for each player, we can bound the estimation error of empirical payoff, then we can use the Nash Equilibrium of meta game as an approximation of Nash Equilibrium of the game.

8.3 Risk averse payoff EGT

Recall that the objective in this chapter is to consider risk averse payoff to evaluate strategies. Hence, instead of letting

$$r^i(\pi^1, ..., \pi^p) = \mathbb{E}[R^i(\pi^1, ..., \pi^p)],$$

I choose

$$h^i(\pi^1, ..., \pi^p) = \mathbb{E}[R^i(\pi^1, ..., \pi^p)] - \beta \cdot \mathbb{V}ar[R^i(\pi^1, ..., \pi^p)]$$

(where $\beta > 0$) as the game payoff. Moreover, I use

$$\hat{h}^i(\pi^i, ..., \pi^p) = \bar{R}^i - \beta \cdot \left[\frac{1}{n-1} \sum_{j=1}^{n} \left(R^i_j - \bar{R}^i \right)^2 \right] \tag{8.1}$$

as meta-game payoff, where $\bar{R}^i = \frac{1}{n} \sum_{j=1}^{n} R^i_j$ and R^i_j is the stochastic payoff of player i in $j-$th experiment. To my knowledge, there is no previous work on empirical game theory analysis with risk sensitive payoff. Below in this section, I give the first theoretical analysis showing that for the risk-averse payoff game, the Nash Equilibrium could also be approximated by meta game.

Theorem 7. *Under Assumption 4, for a Normal Form Game with p players, and each player i chooses a strategy π^i from a set of strategies $S^i = \{\pi^i_1, ..., \pi^i_k\}$ and receives a meta payoff $h^i(\pi^1, ..., \pi^p)$ (Eq. (8.1)). If $\mathbf{x}$ is a Nash Equilibrium for the game $\hat{h}^i(\pi^1, ..., \pi^p)$, then it is a 2ϵ-Nash equilibrium for the game $h^i(\pi^1, ..., \pi^p)$ with probability $1 - \delta$ if we play the game for n times, where*

$$n \geq \max \left\{ \frac{8R^2}{\epsilon^2} \log \frac{|S^1| \times ... \times |S^p| \times p}{\delta} \; ; \; \frac{128R^4\beta^2}{\epsilon^2 n} \log \frac{|S^1| \times ... \times |S^p| \times p}{\delta} \right\} \tag{8.2}$$

For the proof of Theorem 7, please check Section A.5.

Chapter 9

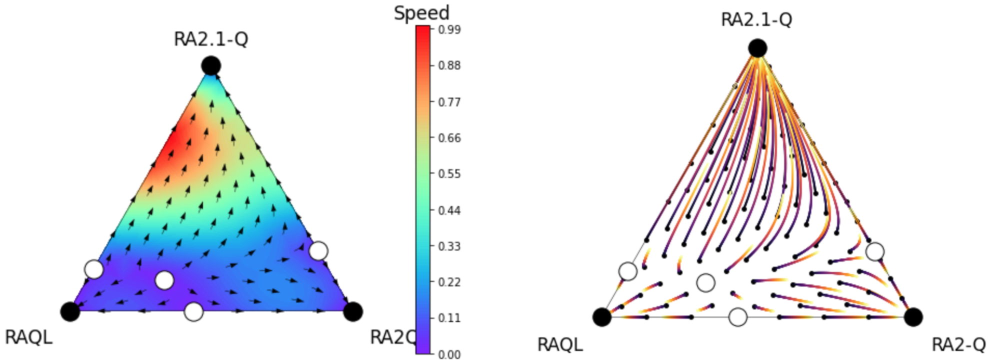
Empirical Results on Trading Market

In this section, experiments are done using the open-sourced ABIDES [15] market simulator in a simplified setting. The environment is generated by replaying publicly available real trading data for a single stock ticker.[1] The setting is composed of one non-learning agent that replays the market deterministically [6] and learning agents. The learning agents considered are: RAQL, RA2-Q, RA2.1-Q, and RA3-Q.

The experimental setting follows a similar setting to existing implementations in ABIDES[2] where the state space is defined by two features: current holdings and volume imbalance. Agents take one action at every time step (every second) selecting among: *buy/sell* with limit price $base + i \cdot K$, where $i \in \{1, 2, .., 6\}$ or *do nothing*. The immediate reward is defined by the change in the value of our portfolio (mark-to-market) and comparing against the previous time step. Our comparisons are in terms of Eq. (8.1), where β is carefully adjusted.

Table 9.1: Meta-payoff of 2 players, 3 strategies, respectively RAQL [49], **RA2-Q** and **RA2.1-Q** over 80 simulations.

N_{i1}	N_{i2}	N_{i3}	R_{i1}	R_{i2}	R_{i3}
2	0	0	0.9130	0	0
1	1	0	0.7311	0.7970	0
0	2	0	0	1.0298	0
1	0	1	0.6791	0	1.0786
0	0	2	0	0	2.2177
0	1	1	0	0.7766	1.4386



(a)

(b)

Figure 9.1: (a) Directional field plot and (b) Trajectory plot of the simplex of 3 strategies based on the meta-game payoff from Table 9.1. It can be seen that RA2.1-Q (top) is the the strongest attractor. White circles represent equilibria.

9.1 Single-agent Algorithms Comparison Using EGT

Table 9.1 shows the meta-payoff table of a two player-game among three strategies: RAQL, RA2-Q and RA2.1-Q. The results show that our two proposed algorithms RA2-Q and RA2.1-Q obtained better results than RAQL. With those payoffs I obtained the directional and trajectory plots shown in Fig. 9.1, where black solid circles denote globally-stable equilibria, and the white circles denote unstable equilibria (saddle-points), in (a) the plot is colored according to the speed at which the strategy mix is changing at each point; in (b) the lines show trajectories for some points over the simplex.

9.2 Robustness

Table 9.2: Comparison of two types of perturbations: The trained adversary from RA3-Q is used in testing time. Zero-intelligence agents are added to the simulation to perturb the market. RA3-Q obtains better results in both cases due to its enhanced robustness.

Algorithm/Setting	Adversarial Perturbation	ZI Agents Perturbation
RA2-Q	0.5269	0.9538
RA3-Q	0.9347	1.0692

The last experiment compares RA2-Q and RA3-Q in terms of robustness. In this setting I trained both agents under the same conditions as a first step. Then in testing phase I added two types of perturbations, one adversarial agent (trained within RA3-Q) or adding noise (aka. zero-intelligence) agents in the environment. In both cases, the agents will act in a perturbed environment. The results are presented in Table 9.2 using cross validation with 80 simulations.

9.3 Summarize of Algorithms

In summary, I proposed 3 different Q-learning style algorithms that augment

[1] https://lobsterdata.com/info/DataSamples.php

[2] https://github.com/abides-sim/abides/blob/master/agent/examples/QLearningAgent.py

reinforcement learning agents with risk-awareness, variance reduction, and robustness. RA2-Q and RA2.1-Q are risk-averse but use slightly different techniques to reduce variance. RA3-Q is a proposal that extend by adding an adversarial learning layer which is expected to improve its robustness. On the one side, theoretical results show convergence results for RA2-Q, on the other side, in empirical results RA2.1-Q and RA3-Q obtained better results in a simplified trading scenario.

Chapter 10

Conclusion and Future Work

One of the main contributions of this work concern a general characterization and analysis based on non-uniform properties, which can be applied in policy gradient methods, GLM and other machine learning cases that involve non-convex optimization problems. Also, they significantly improve convergence rates over previous work and even over classical lower bounds. A valuable open question regarding this is to incorporate stochastic gradient (Karimi *et al.*) and other adaptive gradient-based methods (Kingma and Ba) in the analysis, e.g., what convergence guarantees does stochastic geometry-aware gradient descent have for different functions? Another interesting question would be to push the analysis to other domains with more complex function approximators, including neural networks (Allen-Zhu *et al.*).

For the later half of this work, I have proposed 3 different Q-learning style algorithms that augment reinforcement learning agents with risk-awareness, variance reduction, and robustness. RA2-Q and RA2.1-Q are risk-averse but use slightly different techniques to reduce variance. RA3-Q is a proposal that extend by adding an adversarial learning layer which is expected to improve its robustness. On the one side, my theoretical results show convergence results for RA2-Q, on the other side, in the empirical results RA2.1-Q and RA3-Q obtained better results in a simplified trading scenario. Lastly, I contributed with risk-averse analysis of those algorithms using empirical game theory. As future work I want to perform a more extensive set of experiments to evaluate the algorithms under different conditions. Also, it's an interesting open question whether RA2.1-Q also enjoys minimax optimality convergence rate up to a logarithmic factor as in [55]. On the side of EGT analysis, previous works used average as payoff [53] and my work considers a risk-averse measure based on variance (second moment), studying higher moments and other measures is one interesting open question.

References

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

[2] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*, PMLR, 2019, pp. 242–252.

[3] O. Anschel, N. Baram, and N. Shimkin, "Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2017, pp. 176–185.

[4] M. Bagherzadeh, N. Kahani, and L. Briand, "Reinforcement learning for test case prioritization," *arXiv preprint arXiv:2011.01834*, 2020.

[5] S. Balakrishnan. (). "Intermediate statistics." C. M. University, Ed., [Online]. Available: `http://www.stat.cmu.edu/~siva/705/lec3.pdf`.

[6] T. H. Balch, M. Mahfouz, J. Lockhart, M. Hybinette, and D. Byrd, "How to evaluate trading strategies: Single agent market replay or multiple agent interactive simulation?" *arXiv preprint arXiv:1906.12010*, 2019.

[7] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, "Autonomous navigation of stratospheric balloons using reinforcement learning," *Nature*, vol. 588, no. 7836, pp. 77–82, 2020.

[8] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv preprint arXiv:1912.06680*, 2019.

[9] D. P. Bertsekas, "Neuro-dynamic programmingneuro-dynamic programming," in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Boston, MA: Springer US, 2009, pp. 2555–2560, ISBN: 978-0-387-74759-0. DOI: `10.1007/978-0-387-74759-0_440`. [Online]. Available: `https://doi.org/10.1007/978-0-387-74759-0_440`.

[10] J. Bhandari and D. Russo, "A note on the linear convergence of policy gradient methods," *arXiv preprint arXiv:2007.11120*, 2020.

[11] D. Bloembergen, D. Hennes, P. McBurney, and K. Tuyls, "Trading in markets with noisy information: An evolutionary analysis," *Connection Science*, vol. 27, no. 3, pp. 253–268, 2015.

[12] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, no. 2, pp. 215–250, 2002.

[13] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.

[14] V. V. Buldygin and Y. V. Kozachenko, "Sub-gaussian random variables," *Ukrainian Mathematical Journal*, vol. 32, no. 6, pp. 483–489, 1980. DOI: `10.1007/BF01087176`.

[15] D. Byrd, M. Hybinette, and T. H. Balch, "Abides: Towards high-fidelity market simulation for ai research," *arXiv preprint arXiv:1904.12066*, 2019.

[16] R. Cavazos-Cadena, "Optimality equations and inequalities in a class of risk-sensitive average cost markov decision chains," *Mathematical Methods of Operations Research*, vol. 71, no. 1, pp. 47–84, DOI: `10.1007/s00186-009-0285-6`. [Online]. Available: `https://doi.org/10.1007/s00186-009-0285-6`.

[17] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *arXiv preprint arXiv:2007.06558*, 2020.

[18] D. Di Castro, A. Tamar, and S. Mannor, "Policy gradients with variance related risk criteria," *arXiv preprint arXiv:1206.6404*, 2012.

[19] Y. Gao, K. Y. C. Lui, and P. Hernandez-Leal, *Robust risk-sensitive reinforcement learning agents for trading markets*, 2021. arXiv: `2107.08083 [cs.LG]`.

[20] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[21] E. Hazan, K. Levy, and S. Shalev-Shwartz, "Beyond convexity: Stochastic quasi-convex optimization," *Advances in neural information processing systems*, vol. 28, pp. 1594–1602, 2015.

[22] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[23] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 242–250, ISBN: 1558605568.

[24] C. Igel and M. Hüsken, *Improving the rprop learning algorithm*, 2000.

[25] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013.

[26] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *ICML*, vol. 2, 2002, pp. 267–274.

[27] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2016, pp. 795–811.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, 1998, pp. 769–783.

[30] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Softmax policy gradient methods can take exponential time to converge," *arXiv preprint arXiv:2102.11270*, 2021.

[31] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[32] Y. Li, C. Szepesvari, and D. Schuurmans, "Learning exercise policies for american options," in *Artificial Intelligence and Statistics*, PMLR, 2009, pp. 352–359.

[33] M. L. Littman, "Value-function reinforcement learning in markov games," *Cognitive systems research*, vol. 2, no. 1, pp. 55–66, 2001.

[34] S. Łojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les équations aux dérivées partielles*, vol. 117, pp. 87–89, 1963.

[35] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans, *Leveraging non-uniformity in first-order non-convex optimization*, 2021. arXiv: `2105.06072 [cs.LG]`.

[36] J. Mei, C. Xiao, B. Dai, L. Li, C. Szepesvári, and D. Schuurmans, "Escaping the gravitational pull of softmax," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[37] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6820–6829.

[38] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, no. 2, pp. 267–290, 2002.

[39] A. Mirhoseini, A. Goldie, M. Yazgan, J. Jiang, E. Songhori, S. Wang, Y.-J. Lee, E. Johnson, O. Pathak, S. Bae, *et al.*, "Chip placement with deep reinforcement learning," *arXiv preprint arXiv:2004.10746*, 2020.

[40] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural computation*, vol. 17, no. 2, pp. 335–359, 2005.

[41] R. Murray, B. Swenson, and S. Kar, "Revisiting normalized gradient descent: Fast evasion of saddle points," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4818–4824, Nov. 2019, ISSN: 2334-3303. DOI: 10.1109/tac.2019.2914998. [Online]. Available: http://dx.doi.org/10.1109/TAC.2019.2914998.

[42] A. S. Nemirovski and D. B. Yudin, "Problem complexity and method efficiency in optimization," 1983.

[43] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003, vol. 87.

[44] B. Ning, F. H. T. Lin, and S. Jaimungal, "Double deep q-learning for optimal execution," *arXiv preprint arXiv:1812.06600*, 2018.

[45] X. Pan, D. Seita, Y. Gao, and J. Canny, "Risk averse robust adversarial reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 8522–8528.

[46] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2017, pp. 2817–2826.

[47] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[48] W. F. Sharpe, "The sharpe ratio," *Journal of portfolio management*, vol. 21, no. 1, pp. 49–58, 1994.

[49] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," *Neural computation*, vol. 26, no. 7, pp. 1298–1328, 2014.

[50] T. Spooner, J. Fearnley, R. Savani, and A. Koukorinis, "Market making via reinforcement learning," *arXiv preprint arXiv:1804.04216*, 2018.

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

[52] T. Théate and D. Ernst, "An application of deep reinforcement learning to algorithmic trading," *Expert Systems with Applications*, vol. 173, p. 114 632, 2021.

[53] K. Tuyls, J. Perolat, M. Lanctot, E. Hughes, R. Everett, J. Z. Leibo, C. Szepesvári, and T. Graepel, "Bounds and dynamics for empirical game theoretic analysis," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 1, pp. 1–30, 2020.

[54] T. A. University, Ed. (). "Mean variance utility," [Online]. Available: `https://www.tau.ac.il/~spiegel/teaching/corpfin/mean-variance.pdf`.

[55] M. J. Wainwright, "Variance-reduced q-learning is minimax optimal," *arXiv preprint arXiv:1906.04697*, 2019.

[56] W. E. Walsh, R. Das, G. Tesauro, and J. O. Kephart, "Analyzing complex strategic interactions in multi-agent systems," in *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents*, 2002, pp. 109–118.

[57] J. W. Weibull, *Evolutionary game theory*. MIT press, 1997.

[58] M. Weinberg and J. S. Rosenschein, "Best-response multiagent learning in non-stationary environments," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 2004, pp. 506–513.

[59] M. P. Wellman, "Methods for empirical game-theoretic analysis," in *AAAI*, 2006, pp. 1552–1556.

[60] A. Wilson, L. Mackey, and A. Wibisono, "Accelerating rescaled gradient descent: Fast optimization of smooth functions," *arXiv preprint arXiv:1902.08825*, 2019.

[61] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representations*, 2019.

Appendix A

Appendix

The appendix is organized as follows.

- Section A.1: proofs of policy gradient in Chapter 5.
  - Section A.1.1: proofs of one-state MDPs.
  - Section A.1.2: proofs of general MDPs.
- Section A.2: proofs of generalized linear model in Chapter 6.
- Section A.4: discuss of convergence guarantees of risk-averse RL algorithms.
- Section A.5: proofs of Theorem 7.
- Miscellaneous extra supporting results those are not mentioned in the main paper.

A.1 Proofs for Chapter 5

A.1.1 One-state MDP

Lemma 2 (NS) **.** Let $\pi_\theta = \text{softmax}(\theta)$ and $\pi_{\theta'} = \text{softmax}(\theta')$. Denote $\theta_\zeta := \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. For any $r \in [0, 1]^K$, $\theta \mapsto \pi_\theta^\top r$ is $\beta(\theta_\zeta)$ non-uniform smooth with $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$.

Proof. Let $S := S(r, \theta) \in \mathbb{R}^{K \times K}$ be the second derivative of the value map $\theta \mapsto \pi_\theta^\top r$. By Taylor's theorem, it suffices to show that the spectral radius of S is upper bounded. Denote $H(\pi_\theta) \coloneqq \operatorname{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top$ as the Jacobian of $\theta \mapsto \operatorname{softmax}(\theta)$. Now, by its definition we have

$$\begin{aligned}
S &= \frac{d}{d\theta} \left\{ \frac{d\pi_\theta^\top r}{d\theta} \right\} && \text{(A.1)} \\
&= \frac{d}{d\theta} \left\{ H(\pi_\theta) r \right\} && \text{(A.2)} \\
&= \frac{d}{d\theta} \left\{ (\operatorname{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top) r \right\}. && \text{(A.3)}
\end{aligned}$$

Continuing with our calculation fix $i, j \in [K]$. Then,

$$\begin{aligned}
S_{(i,j)} &= \frac{d\{\pi_\theta(i) \cdot (r(i) - \pi_\theta^\top r)\}}{d\theta(j)} && \text{(A.4)} \\
&= \frac{d\pi_\theta(i)}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + \pi_\theta(i) \cdot \frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)} && \text{(A.5)} \\
&= (\delta_{ij} \pi_\theta(j) - \pi_\theta(i) \pi_\theta(j)) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \cdot (\pi_\theta(j) r(j) - \pi_\theta(j) \pi_\theta^\top r) && \text{(A.6)} \\
&= \delta_{ij} \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(j) - \pi_\theta^\top r), && \text{(A.7)}
\end{aligned}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases} \tag{A.8}$$

is Kronecker's δ-function. To show the bound on the spectral radius of S, pick $y \in \mathbb{R}^K$. Then,

$$\begin{aligned}
\left| y^\top S y \right| &= \left| \sum_{i=1}^K \sum_{j=1}^K S_{(i,j)} \cdot y(i) \cdot y(j) \right| && \text{(A.9)} \\
&= \left| \sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i)^2 - 2 \sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i) \sum_j \pi_\theta(j) y(j) \right| && \text{(A.10)} \\
&= \left| (H(\pi_\theta) r)^\top (y \odot y) - 2 \cdot (H(\pi_\theta) r)^\top y \cdot \left(\pi_\theta^\top y \right) \right| && \text{(A.11)} \\
&\le \| H(\pi_\theta) r \|_\infty \cdot \| y \odot y \|_1 + 2 \cdot \| H(\pi_\theta) r \|_2 \cdot \| y \|_2 \cdot \| \pi_\theta \|_1 \cdot \| y \|_\infty && \text{(A.12)} \\
&\le 3 \cdot \| H(\pi_\theta) r \|_2 \cdot \| y \|_2^2 . && \text{(A.13)}
\end{aligned}$$

According to Taylor's theorem, $\forall \theta,\ \theta'$,

$$\left|(\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle\right| = \frac{1}{2} \cdot \left|(\theta' - \theta)^\top S(r, \theta_\zeta)\, (\theta' - \theta)\right| \tag{A.14}$$

$$\le \frac{3}{2} \cdot \left\|H(\pi_{\theta_\zeta}) r\right\|_2 \cdot \|\theta' - \theta\|_2^2 \qquad \text{(by Eq. (A.9))} \tag{A.15}$$

$$= \frac{3}{2} \cdot \left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Lemma 17)} \ \square$$

Lemma 3. Let

$$\theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2. \tag{A.16}$$

Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. We have, for all $\eta \in (0, 1/3)$,

$$\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \le \frac{1}{1 - 3\eta} \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2. \tag{A.17}$$

Proof. Denote $\zeta_1 \coloneqq \zeta$. Also denote $\theta_{\zeta_2} \coloneqq \theta + \zeta_2 \cdot (\theta_{\zeta_1} - \theta)$ with some $\zeta_2 \in [0, 1]$. We have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta}\right\|_2 = \left\|\int_0^1 \left\langle \frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2}, \theta_{\zeta_1} - \theta \right\rangle d\zeta_2\right\|_2 \tag{A.18}$$

$$\le \int_0^1 \left\|\frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2}\right\|_2 \cdot \left\|\theta_{\zeta_1} - \theta\right\|_2 d\zeta_2 \tag{A.19}$$

$$\le \int_0^1 3 \cdot \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 \cdot \zeta_1 \cdot \left\|\theta' - \theta\right\|_2 d\zeta_2 \qquad \text{(by Eq. (A.9))} \tag{A.20}$$

$$\le \int_0^1 3 \cdot \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 \cdot \eta\ d\zeta_2, \qquad \left(\zeta_1 \in [0, 1],\ \text{using } \theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right) \tag{A.21}$$

where the second last inequality is because of the Hessian is symmetric, and its operator norm is equal to its spectral radius. Therefore we have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}}\right\|_2 \le \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + \left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \qquad \text{(by triangle inequality)} \tag{A.22}$$

$$\le \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + 3\eta \cdot \int_0^1 \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 d\zeta_2. \qquad \text{(by Eq. (A.18))} \tag{A.23}$$

Denote $\theta_{\zeta_3} \coloneqq \theta + \zeta_3 \cdot (\theta_{\zeta_2} - \theta)$ with some $\zeta_3 \in [0,1]$. Using similar calculation as in Eq. (A.18), we have,

$$\begin{aligned} \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 &\leq \left\| \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2 + \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} - \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2 && \text{(A.24)} \\ &\leq \left\| \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2 + 3\eta \cdot \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}} \right\|_2 d\zeta_3. && \text{(A.25)} \end{aligned}$$

Combining Eqs. (A.22) and (A.24), we have,

$$\left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \leq (1+3\eta) \cdot \left\| \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2 + (3\eta)^2 \cdot \int_0^1 \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}} \right\|_2 d\zeta_3 d\zeta_2, \quad \text{(A.26)}$$

which implies,

$$\begin{aligned} \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 &\leq \left[\sum_{i=0}^{\infty} (3\eta)^i \right] \cdot \left\| \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2 && \text{(A.27)} \\ &= \frac{1}{1-3\eta} \cdot \left\| \frac{d\pi_{\theta}^\top r}{d\theta} \right\|_2. \qquad (\eta \in (0, 1/3)) && \square \end{aligned}$$

Lemma 4 (Non-vanishing NŁ coefficient) **.** Using normalized policy gradient method, we have $\inf_{t\geq 1} \pi_{\theta_t}(a^*) > 0$.

Proof. The proof is similar to [Mei *et al.*, Lemma 5]. Let

$$c = \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) \quad \text{(A.28)}$$

and

$$\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0 \quad \text{(A.29)}$$

denote the reward gap of r. We will prove that $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$, where $t_0 = \min\{t : \pi_{\theta_t}(a^*) \geq \frac{c}{c+1}\}$. Note that t_0 depends only on θ_1 and c, and c depends only on the problem. Define the following regions,

$$\begin{aligned} \mathcal{R}_1 &= \left\{ \theta : \frac{d\pi_{\theta}^\top r}{d\theta(a^*)} \geq \frac{d\pi_{\theta}^\top r}{d\theta(a)}, \ \forall a \neq a^* \right\}, && \text{(A.30)} \\ \mathcal{R}_2 &= \left\{ \theta : \pi_\theta(a^*) \geq \pi_\theta(a), \ \forall a \neq a^* \right\}, && \text{(A.31)} \\ \mathcal{N}_c &= \left\{ \theta : \pi_\theta(a^*) \geq \frac{c}{c+1} \right\}. && \text{(A.32)} \end{aligned}$$

We make the following three-part claim.

Claim 1. *The following hold :*

a) *Following a NPG update* $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$*, if* $\theta_t \in \mathcal{R}_1$*, then* (i) $\theta_{t+1} \in \mathcal{R}_1$ *and* (ii) $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$.

b) *We have* $\mathcal{R}_2 \subset \mathcal{R}_1$ *and* $\mathcal{N}_c \subset \mathcal{R}_1$.

c) *For* $\eta = 1/6$*, there exists a finite time* $t_0 \geq 1$*, such that* $\theta_{t_0} \in \mathcal{N}_c$*, and thus* $\theta_{t_0} \in \mathcal{R}_1$*, which implies that* $\inf_{t \geq 1} \pi_{\theta_t}(a^*) = \min_{1 \leq t \leq t_0} \pi_{\theta_t}(a^*)$..

Claim a) Part (i): We want to show that if $\theta_t \in \mathcal{R}_1$, then $\theta_{t+1} \in \mathcal{R}_1$. Let

$$\mathcal{R}_1(a) = \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)} \right\} . \tag{A.33}$$

Note that $\mathcal{R}_1 = \cap_{a \neq a^*} \mathcal{R}_1(a)$. Pick $a \neq a^*$. Clearly, it suffices to show that if $\theta_t \in \mathcal{R}_1(a)$ then $\theta_{t+1} \in \mathcal{R}_1(a)$. Hence, suppose that $\theta_t \in \mathcal{R}_1(a)$. We consider two cases.

Case (a): $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a)$. Since $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a)$, we also have $\theta_t(a^*) \geq \theta_t(a)$. After an update of the parameters,

$$\begin{aligned} \theta_{t+1}(a^*) &= \theta_t(a^*) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} && \text{(A.34)} \\ &\geq \theta_t(a) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} && \text{(A.35)} \\ &= \theta_{t+1}(a), && \text{(A.36)} \end{aligned}$$

which implies that $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_{t+1}}(a)$. Since $r(a^*) - \pi_{\theta_{t+1}}^\top r > 0$ and $r(a^*) > r(a)$,

$$\pi_{\theta_{t+1}}(a^*) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r \right) \geq \pi_{\theta_{t+1}}(a) \cdot \left(r(a) - \pi_{\theta_{t+1}}^\top r \right) , \tag{A.37}$$

which is equivalent to $\frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a^*)} \geq \frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a)}$, i.e., $\theta_{t+1} \in \mathcal{R}_1(a)$.

Case (b): Suppose now that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$. First note that for any θ and $a \neq a^*$, $\theta \in \mathcal{R}_1(a)$ holds if and only if

$$r(a^*) - r(a) \geq \left(1 - \frac{\pi_\theta(a^*)}{\pi_\theta(a)} \right) \cdot \left(r(a^*) - \pi_\theta^\top r \right) . \tag{A.38}$$

Indeed, from the condition $\frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)}$, we get

$$\begin{aligned} \pi_\theta(a^*) \cdot \left(r(a^*) - \pi_\theta^\top r\right) &\geq \pi_\theta(a) \cdot \left(r(a) - \pi_\theta^\top r\right) && \text{(A.39)} \\ &= \pi_\theta(a) \cdot \left(r(a^*) - \pi_\theta^\top r\right) - \pi_\theta(a) \cdot \left(r(a^*) - r(a)\right), && \text{(A.40)} \end{aligned}$$

which, after rearranging, is equivalent to Eq. (A.38). Hence, it suffices to show that Eq. (A.38) holds for θ_{t+1} provided it holds for θ_t. From the latter condition, we get

$$r(a^*) - r(a) \geq \left(1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right). \tag{A.41}$$

After an update of the parameters, according to Lemma 13 (or Eq. (A.56) below), $\pi_{\theta_{t+1}}^\top r \geq \pi_{\theta_t}^\top r$, i.e.,

$$0 < r(a^*) - \pi_{\theta_{t+1}}^\top r \leq r(a^*) - \pi_{\theta_t}^\top r. \tag{A.42}$$

On the other hand,

$$\begin{aligned} \theta_{t+1}(a^*) - \theta_{t+1}(a) &= \theta_t(a^*) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} - \theta_t(a) - \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} && \text{(A.43)} \\ &\geq \theta_t(a^*) - \theta_t(a) && \text{(A.44)} \end{aligned}$$

which implies that

$$1 - \exp\left\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\right\} \leq 1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}. \tag{A.45}$$

Furthermore, by our assumption that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$, we have $1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\} = 1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} > 0$. Putting things together, we get

$$\begin{aligned} \left(1 - \exp\left\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) &\leq \left(1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right) && \text{(A.46)} \\ &\leq r(a^*) - r(a), && \text{(A.47)} \end{aligned}$$

which is equivalent to

$$\left(1 - \frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)}\right) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) \leq r(a^*) - r(a), \tag{A.48}$$

and thus by our previous remark, $\theta_{t+1} \in \mathcal{R}_1(a)$, thus, finishing the proof of part (i).

Part (ii): Assume again that $\theta_t \in \mathcal{R}_1$. We want to show that $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$. Since $\theta_t \in \mathcal{R}_1$, we have $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \geq \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$, $\forall a \neq a^*$. Hence,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\left\{\theta_{t+1}(a^*)\right\}}{\sum_a \exp\left\{\theta_{t+1}(a)\right\}} \tag{A.49}$$

$$= \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}} \tag{A.50}$$

$$\geq \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}} \qquad \left(\text{using } \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \geq \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right) \tag{A.51}$$

$$= \frac{\exp\left\{\theta_t(a^*)\right\}}{\sum_a \exp\left\{\theta_t(a)\right\}} = \pi_{\theta_t}(a^*). \tag{A.52}$$

Claim b); Claim c) The proof of those claims are exactly the same as [Mei *et al.*, Lemma 5], since they do not involve the update rule. □

Theorem 1. Using NPG $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2$, with $\eta = 1/6$, for all $t \geq 1$, we have,

$$\left(\pi^* - \pi_{\theta_t}\right)^\top r \leq e^{-\frac{c\cdot(t-1)}{12}} \cdot \left(\pi^* - \pi_{\theta_1}\right)^\top r, \tag{A.53}$$

where $c = \inf_{t\geq 1} \pi_{\theta_t}(a^*) > 0$ is from Lemma 4, and c is a constant that depends on r and θ_1, but not on the time t.

Proof. Denote $\theta_{\zeta_t} := \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t)$ with some $\zeta_t \in [0,1]$. According to Lemma 2,

$$\left|\left(\pi_{\theta_{t+1}} - \pi_{\theta_t}\right)^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle\right| \leq \frac{3}{2} \cdot \left\|\frac{d\pi_{\theta_{\zeta_t}}^\top r}{d\theta_{\zeta_t}}\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{A.54}$$

$$\leq \frac{3}{2} \cdot \frac{1}{1-3\eta} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2, \qquad (\eta = 1/6, \text{ by Lemma 3}) \tag{A.55}$$

which implies,

$$
\begin{aligned}
\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r &\le -\Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle + \frac{3}{2 \cdot (1 - 3\eta)} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \| \theta_{t+1} - \theta_t \|_2^2 && \text{(A.56)} \\
&= -\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + \frac{3 \cdot \eta^2}{2 \cdot (1 - 3\eta)} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \qquad \left(\text{using } \theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \right) && \text{(A.57)} \\
&= -\frac{1}{12} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \qquad \text{(using } \eta = 1/6) && \text{(A.58)} \\
&\le -\frac{1}{12} \cdot \pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r \qquad \text{(by Lemma 1)} && \text{(A.59)} \\
&\le -\frac{1}{12} \cdot \inf_{t \ge 1} \pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r. && \text{(A.60)}
\end{aligned}
$$

According to Eq. (A.56), we have,

$$
\begin{aligned}
\left(\pi^* - \pi_{\theta_t}\right)^\top r &\le \left(1 - \frac{c}{12}\right) \cdot \left(\pi^* - \pi_{\theta_{t-1}}\right)^\top r \qquad \left(c \coloneqq \inf_{t \ge 1} \pi_{\theta_t}(a^*) > 0 \right) && \text{(A.61)} \\
&\le \exp\left\{-c/12\right\} \cdot \left(\pi^* - \pi_{\theta_{t-1}}\right)^\top r && \text{(A.62)} \\
&\le \exp\left\{-(t-1) \cdot c/12\right\} \cdot \left(\pi^* - \pi_{\theta_1}\right)^\top r. && \square
\end{aligned}
$$

A.1.2 Multi-state MDP

Lemma 5 (NŁ) **.** Denote $S \coloneqq |\mathcal{S}|$ as the total number of states. We have, for all $\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$,

$$
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left(V^*(\rho) - V^{\pi_\theta}(\rho) \right), \tag{A.63}
$$

where $a^*(s)$ is the action that π^* selects in state s.

Proof. See the proof in [Mei *et al.*, Lemma 8]. We include a proof for completeness. We have,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 = \left[\sum_{s,a}\left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)}\right)^2\right]^{\frac{1}{2}} \tag{A.64}$$

$$\geq \left[\sum_{s}\left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))}\right)^2\right]^{\frac{1}{2}} \tag{A.65}$$

$$\geq \frac{1}{\sqrt{S}}\sum_s \left|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))}\right| \qquad \text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1},\ |x|\rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2\text{)} \tag{A.66}$$

$$= \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s \left| d_\mu^{\pi_\theta}(s)\cdot \pi_\theta(a^*(s)|s)\cdot A^{\pi_\theta}(s,a^*(s))\right| \qquad \text{(by Lemma 16)} \tag{A.67}$$

$$= \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s d_\mu^{\pi_\theta}(s)\cdot \pi_\theta(a^*(s)|s)\cdot \left|A^{\pi_\theta}(s,a^*(s))\right|. \tag{A.68}$$

$$\big(\text{because } d_\mu^{\pi_\theta}(s)\geq 0 \text{ and } \pi_\theta(a^*(s)|s)\geq 0\big) \tag{A.69}$$

Define the distribution mismatch coefficient as $\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty = \max_s \frac{d_\rho^{\pi^*}(s)}{d_\mu^{\pi_\theta}(s)}$. We have,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)}\cdot d_\rho^{\pi^*}(s)\cdot \pi_\theta(a^*(s)|s)\cdot \left|A^{\pi_\theta}(s,a^*(s))\right| \tag{A.70}$$

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\sum_s d_\rho^{\pi^*}(s)\cdot\left|A^{\pi_\theta}(s,a^*(s))\right| \tag{A.71}$$

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\sum_s d_\rho^{\pi^*}(s)\cdot A^{\pi_\theta}(s,a^*(s)) \tag{A.72}$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^*}(s)\sum_a \pi^*(a|s)\cdot A^{\pi_\theta}(s,a) \tag{A.73}$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\left[V^*(\rho)-V^{\pi_\theta}(\rho)\right], \tag{A.74}$$

where the one but last equality used that π^* is deterministic and in state s chooses $a^*(s)$ with probability one, and the last equality uses the performance difference formula (Lemma 18). □

Lemma 6 (NS) **.** Let Assumption 1 hold and denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. $\theta \mapsto V^{\pi_\theta}(\mu)$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with

$$\beta(\theta_\zeta) = \left[3 + \frac{2 \cdot (C_\infty - (1 - \gamma))}{(1 - \gamma) \cdot \gamma} \right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta} \right\|_2, \tag{A.75}$$

where $C_\infty \coloneqq \max_\pi \left\| \frac{d_\mu^\pi}{\mu} \right\|_\infty \le \frac{1}{\min_s \mu(s)} < \infty$.

Proof. The main part is to prove that for all $y \in \mathbb{R}^{SA}$ and θ,

$$\left| y^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} y \right| \le \left[3 + \frac{2 \cdot (C_\infty - (1 - \gamma))}{(1 - \gamma) \cdot \gamma} \right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \cdot \|y\|_2^2. \tag{A.76}$$

We first calculate the second order derivative of $V^{\pi_\theta}(\mu)$ w.r.t. θ.

Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} &= \Big\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle & \text{(A.77)} \\ &= \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \Big\rangle & \text{(A.78)} \\ &= \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s, \cdot)}, u(s, \cdot) \Big\rangle \qquad \left(\frac{\partial \pi_\theta(a|s)}{\partial \theta(s', \cdot)} = \mathbf{0}, \ \forall s' \ne s \right) & \text{(A.79)} \end{aligned}$$

Similarly, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} &= \Big\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle & \text{(A.80)} \\ &= \Big\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle & \text{(A.81)} \\ &= \Big\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s, \cdot)} u(s, \cdot), u(s, \cdot) \Big\rangle. & \text{(A.82)} \end{aligned}$$

Define $\Pi(\alpha) \in \mathbb{R}^{S \times SA}$ as follows,

$$\Pi(\alpha) \coloneqq \begin{bmatrix} \pi_{\theta_\alpha}(\cdot|1)^\top & \mathbf{0}^\top & \cdots & \mathbf{0}^\top \\ \mathbf{0}^\top & \pi_{\theta_\alpha}(\cdot|2)^\top & \cdots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \pi_{\theta_\alpha}(\cdot|S)^\top \end{bmatrix}. \tag{A.83}$$

Denote $\mathcal{P} \in \mathbb{R}^{SA \times S}$ such that,

$$\mathcal{P}_{(sa, s')} \coloneqq \mathcal{P}(s'|s, a). \tag{A.84}$$

Define $P(\alpha) \coloneqq \Pi(\alpha)\mathcal{P} \in \mathbb{R}^{S\times S}$, where $\forall (s, s')$,

$$[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{A.85}$$

The derivative w.r.t. α is

$$\frac{\partial P(\alpha)}{\partial \alpha} = \frac{\partial \Pi(\alpha)\mathcal{P}}{\partial \alpha} = \frac{\partial \Pi(\alpha)}{\partial \alpha}\mathcal{P}. \tag{A.86}$$

And $\forall (s, s')$, we have,

$$\left[\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{A.87}$$

Next, consider the state value function of π_{θ_α},

$$V^{\pi_{\theta_\alpha}}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot V^{\pi_{\theta_\alpha}}(s'), \tag{A.88}$$

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha} \tag{A.89}$$

$$V^{\pi_{\theta_\alpha}}(\mu) = \mu^\top M(\alpha) r_{\theta_\alpha}, \tag{A.90}$$

where

$$M(\alpha) = \left(\mathbf{Id} - \gamma P(\alpha)\right)^{-1}, \tag{A.91}$$

and $r_{\theta_\alpha} \in \mathbb{R}^S$ is given by

$$r_{\theta_\alpha} = \Pi(\alpha) r, \tag{A.92}$$

where $r \in \mathbb{R}^{SA}$. Taking derivative w.r.t. α in Eq. (A.90),

$$\begin{aligned}
\frac{\partial V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha} &= \gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \mu^\top M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} && \text{(A.93)} \\
&= \mu^\top M(\alpha) \left[\gamma \cdot \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \right] && \text{(A.94)} \\
&= \mu^\top M(\alpha) \left[\gamma \cdot \frac{\partial \Pi(\alpha)}{\partial \alpha} \mathcal{P} M(\alpha) r_{\theta_\alpha} + \frac{\partial \Pi(\alpha)}{\partial \alpha} r \right] \qquad \text{(by Eqs. (A.86) and (A.92))} && \text{(A.95)} \\
&= \mu^\top M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}, && \text{(A.96)}
\end{aligned}$$

where $Q^{\pi_{\theta_\alpha}} \in \mathbb{R}^{SA}$ is the state-action value and it satisfies,

$$Q^{\pi_{\theta_\alpha}} = r + \gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} \tag{A.97}$$

$$= r + \gamma \cdot \mathcal{P} V^{\pi_{\theta_\alpha}} \qquad \text{(by Eq. (A.89))} \tag{A.98}$$

Similarly, taking second derivative w.r.t. α,

$$\frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha^2} = 2\gamma^2 \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \gamma \cdot \mu^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \tag{A.99}$$

$$+ 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + \mu^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \tag{A.100}$$

$$= 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} \left(\gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} + r \right) + \tag{A.101}$$

$$\mu^\top M(\alpha) \frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} \left(\gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} + r \right) \tag{A.102}$$

$$= 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} + \mu^\top M(\alpha) \frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}} \tag{A.103}$$

For the last term, we have,

$$\left[\frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}} \Big|_{\alpha=0} \right]_{(s)} = \sum_a \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \cdot Q^{\pi_\theta}(s,a) \tag{A.104}$$

$$= \sum_a \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \cdot Q^{\pi_\theta}(s,a) \qquad \text{(by Eq. (A.80))} \tag{A.105}$$

$$= u(s,\cdot)^\top \left[\sum_a \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] u(s,\cdot) \tag{A.106}$$

Let $S(a,\theta) = \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \in \mathbb{R}^{A \times A}$. $\forall i, j \in [A]$, the value of $S(a,\theta)$ is,

$$S_{(i,j)} = \frac{\partial \{ \delta_{ia} \pi_\theta(a|s) - \pi_\theta(a|s) \pi_\theta(i|s) \}}{\partial \theta(s,j)} \tag{A.107}$$

$$= \delta_{ia} \cdot [\delta_{ja} \pi_\theta(a|s) - \pi_\theta(a|s) \pi_\theta(j|s)] - \pi_\theta(a|s) \cdot [\delta_{ij} \pi_\theta(j|s) \tag{A.108}$$

$$- \pi_\theta(i|s) \pi_\theta(j|s)] - \pi_\theta(i|s) \cdot [\delta_{ja} \pi_\theta(a|s) - \pi_\theta(a|s) \pi_\theta(j|s)], \tag{A.109}$$

where the δ notation is as defined in Eq. (A.8). Then we have,

$$\left[\sum_a \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right]_{(i,j)} = \sum_a S_{(i,j)} \cdot Q^{\pi_\theta}(s,a) \tag{A.110}$$

$$= \delta_{ij} \cdot \pi_\theta(i|s) \cdot [Q^{\pi_\theta}(s,i) - V^{\pi_\theta}(s)] \tag{A.111}$$

$$- \pi_\theta(i|s) \cdot \pi_\theta(j|s) \cdot [Q^{\pi_\theta}(s,i) - V^{\pi_\theta}(s)] - \pi_\theta(i|s) \cdot \pi_\theta(j|s) \cdot [Q^{\pi_\theta}(s,j) - V^{\pi_\theta}(s)]. \tag{A.112}$$

Therefore we have,

$$\left[\frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)} = \sum_{i=1}^{A}\sum_{j=1}^{A} u(s,i)\cdot u(s,j)\cdot \left[\sum_a \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)}\cdot Q^{\pi_\theta}(s,a)\right]_{(i,j)} \tag{A.113}$$

$$= \left(H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right)^\top \left(u(s,\cdot)\odot u(s,\cdot)\right) \tag{A.114}$$

$$- 2\cdot \left[\left(H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right)^\top u(s,\cdot)\right]\cdot \left(\pi_\theta(\cdot|s)^\top u(s,\cdot)\right), \tag{A.115}$$

where $H(\pi) \coloneqq \operatorname{diag}(\pi) - \pi\pi^\top$. Combining the above results with Eq. (A.99), we have,

$$\left|\mu^\top M(\alpha)\frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right| \le \frac{1}{1-\gamma}\cdot \sum_s d_\mu^{\pi_\theta}(s)\cdot \left|\left[\frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)}\right| \tag{A.116}$$

$$\le \frac{1}{1-\gamma}\cdot \sum_s d_\mu^{\pi_\theta}(s)\cdot 3\cdot \left\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right\|_2\cdot \|u\|_2^2 \quad \text{(by Hölder's inequality)} \tag{A.117}$$

$$\le \frac{3\cdot\sqrt{S}}{1-\gamma}\cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot \left\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right\|_2^2\right]^{\frac{1}{2}}\cdot \|u\|_2^2 \quad \text{(by Cauchy-Schwarz)} \tag{A.118}$$

$$= 3\cdot\sqrt{S}\cdot \left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2\cdot \|u\|_2^2. \qquad \text{(by Lemma 17)} \tag{A.119}$$

For the first term in Eq. (A.99), we have,

$$\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0} = \sum_{s'}\left[\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0}\right]_{(s')}\cdot \tag{A.120}$$

$$\left[M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s')}, \tag{A.121}$$

since,

$$\left(\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}\right)^\top \in \mathbb{R}^S, \quad \text{and} \quad M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \in \mathbb{R}^S. \tag{A.122}$$

Next we have,

$$\left[M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \Big|_{\alpha=0} \right]_{(s')} = \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \left[\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \Big|_{\alpha=0} \right]_{(s)} \tag{A.123}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \sum_{a} \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \cdot Q^{\pi_\theta}(s,a) \tag{A.124}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \sum_{a} \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \Big\rangle \cdot Q^{\pi_\theta}(s,a) \qquad \text{(by Eq. (A.77))} \tag{A.125}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \Big\langle \sum_{a} \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a), u(s,\cdot) \Big\rangle \tag{A.126}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \left(H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right)^\top u(s,\cdot), \tag{A.127}$$

$$(H(\pi_\theta) \text{ is the Jacobian of } \theta \mapsto \text{softmax}(\theta)) \tag{A.128}$$

which implies,

$$\left| \left[M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \Big|_{\alpha=0} \right]_{(s')} \right| \le \frac{1}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \left\| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right\|_2 \cdot \left\| u(s,\cdot) \right\|_2 \tag{A.129}$$

$$\le \frac{\|u\|_2}{1-\gamma} \cdot \sum_{s} d_{s'}^{\pi_\theta}(s) \cdot \left\| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right\|_2. \tag{A.130}$$

On the other hand,

$$\left[\mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} \right]_{(s')} = \frac{1}{1-\gamma} \cdot \sum_{s} d_{\mu}^{\pi_\theta}(s) \cdot \left[\frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} \right]_{(s,s')} \qquad \left(\frac{\partial P(\alpha)}{\partial \alpha} \in \mathbb{R}^{S \times S} \right) \tag{A.131}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{\mu}^{\pi_\theta}(s) \cdot \sum_{a} \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \qquad \text{(by Eq. (A.87))} \tag{A.132}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{\mu}^{\pi_\theta}(s) \cdot \sum_{a} \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \Big\rangle \cdot \mathcal{P}(s'|s,a) \qquad \text{(by Eq. (A.77))} \tag{A.133}$$

$$= \frac{1}{1-\gamma} \cdot \sum_{s} d_{\mu}^{\pi_\theta}(s) \cdot \sum_{a} \pi_\theta(a|s) \cdot \mathcal{P}(s'|s,a) \cdot \left[u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right], \tag{A.134}$$

which implies,

$$\left| \left[\mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} \right]_{(s')} \right| \le \frac{1}{1-\gamma} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot \mathcal{P}(s'|s,a) \cdot 2 \cdot \left\| u(s,\cdot) \right\|_\infty \tag{A.135}$$

$$\le \frac{2 \cdot \left\| u \right\|_2}{1-\gamma} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot \mathcal{P}(s'|s,a). \tag{A.136}$$

According to

$$d_\mu^{\pi_\theta}(s') = (1-\gamma) \cdot \mu(s') + \gamma \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot \mathcal{P}(s'|s,a), \quad \forall s' \in \mathcal{S} \tag{A.137}$$

we have,

$$\left| \left[\mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} \right]_{(s')} \right| \le \frac{2 \cdot \left\| u \right\|_2}{(1-\gamma) \cdot \gamma} \cdot \left[d_\mu^{\pi_\theta}(s') - (1-\gamma) \cdot \mu(s') \right] \tag{A.138}$$

$$= \frac{2 \cdot \left\| u \right\|_2}{(1-\gamma) \cdot \gamma} \cdot \left[\frac{d_\mu^{\pi_\theta}(s')}{\mu(s')} \cdot \mu(s') - (1-\gamma) \cdot \mu(s') \right] \tag{A.139}$$

$$\le \frac{2 \cdot \left\| u \right\|_2}{(1-\gamma) \cdot \gamma} \cdot (C_\infty - (1-\gamma)) \cdot \mu(s'). \qquad \left(C_\infty \coloneqq \max_\pi \left\| \frac{d_\mu^\pi}{\mu} \right\|_\infty < \left\| \frac{1}{\mu} \right\|_\infty < \infty \right) \tag{A.140}$$

Combining Eqs. (A.120), (A.129) and (A.138), we have,

$$\left| \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \Big|_{\alpha=0} \right| \tag{A.141}$$

$$\le \sum_{s'} \frac{2 \cdot \left\| u \right\|_2}{(1-\gamma) \cdot \gamma} \cdot (C_\infty - (1-\gamma)) \cdot \mu(s') \cdot \frac{\left\| u \right\|_2}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left\| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right\|_2 \tag{A.142}$$

$$= \frac{2 \cdot (C_\infty - (1-\gamma))}{(1-\gamma)^2 \cdot \gamma} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \left\| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right\|_2 \cdot \left\| u \right\|_2^2 \tag{A.143}$$

$$\le \frac{2 \cdot (C_\infty - (1-\gamma)) \cdot \sqrt{S}}{(1-\gamma)^2 \cdot \gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \left\| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \right\|_2^2 \right]^{\frac{1}{2}} \cdot \left\| u \right\|_2^2 \tag{A.144}$$

$$\text{(by Cauchy-Schwarz)} \tag{A.145}$$

$$= \frac{2 \cdot (C_\infty - (1-\gamma)) \cdot \sqrt{S}}{(1-\gamma) \cdot \gamma} \cdot \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \cdot \left\| u \right\|_2^2. \qquad \text{(by Lemma 17)} \tag{A.146}$$

Combining Eqs. (A.99), (A.116) and (A.141),

$$\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \le \left[3 + \frac{2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right]\cdot\sqrt{S}\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \cdot \|u\|_2^2, \tag{A.147}$$

which implies for all $y \in \mathbb{R}^{SA}$ and θ,

$$\begin{aligned}
\left|y^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} y\right| &= \left|\left(\frac{y}{\|y\|_2}\right)^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2}\left(\frac{y}{\|y\|_2}\right)\right|\cdot\|y\|_2^2 && \text{(A.148)}\\
&\le \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} u, u\right\rangle\right|\cdot\|y\|_2^2 && \text{(A.149)}\\
&= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \theta_\alpha^2}\Big|_{\alpha=0}\frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right|\cdot\|y\|_2^2 && \text{(A.150)}\\
&= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha}\right\}\Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right|\cdot\|y\|_2^2 && \text{(A.151)}\\
&= \max_{\|u\|_2=1}\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha^2}\Big|_{\alpha=0}\right|\cdot\|y\|_2^2 && \text{(A.152)}\\
&\le \left[3 + \frac{2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right]\cdot\sqrt{S}\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2\cdot\|y\|_2^2. \quad \text{(by Eq. (A.147))} && \text{(A.153)}
\end{aligned}$$

Denote $\theta_\zeta = \theta + \zeta(\theta' - \theta)$, where $\zeta \in [0,1]$. According to Taylor's theorem, $\forall s$, $\forall \theta,\ \theta'$,

$$\begin{aligned}
\left|V^{\pi_{\theta'}}(\mu) - V^{\pi_\theta}(\mu) - \left\langle \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}, \theta' - \theta\right\rangle\right| &= \frac{1}{2}\cdot\left|(\theta'-\theta)^\top \frac{\partial^2 V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta^2}(\theta'-\theta)\right| && \text{(A.154)}\\
&\le \frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{2\cdot(1-\gamma)\cdot\gamma}\cdot\sqrt{S}\cdot\left\|\frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta}\right\|_2\cdot\|\theta'-\theta\|_2^2. && \text{(A.155)}
\end{aligned}$$

(by Eq. (A.148)) □

Lemma 8 (Non-vanishing NŁ coefficient) **.** Let Assumption 1 hold. We have, $c := \inf_{s\in\mathcal{S}, t\ge 1}\pi_{\theta_t}(a^*(s)|s) > 0$, where $\{\theta_t\}_{t\ge 1}$ is generated by Algorithm 1.

Proof. The proof is similar to [Mei *et al.*, Lemma 9] and is an extension of the proof for Lemma 4. Denote $\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a\ne a^*(s)} Q^*(s,a) > 0$ as the optimal value gap of state s, where $a^*(s)$ is the action that the optimal policy selects under state s, and $\Delta^* = \min_{s\in\mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. For each state $s \in \mathcal{S}$, define the following sets:

$$\mathcal{R}_1(s) = \left\{\theta : \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \geq \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a)},\ \forall a \neq a^*\right\}, \tag{A.156}$$

$$\mathcal{R}_2(s) = \left\{\theta : Q^{\pi_\theta}(s, a^*(s)) \geq Q^*(s, a^*(s)) - \Delta^*(s)/2\right\}, \tag{A.157}$$

$$\mathcal{R}_3(s) = \left\{\theta_t : V^{\pi_{\theta_t}}(s) \geq Q^{\pi_{\theta_t}}(s, a^*(s)) - \Delta^*(s)/2,\ \text{for all } t \geq 1 \text{ large enough}\right\}, \tag{A.158}$$

$$\mathcal{N}_c(s) = \left\{\theta : \pi_\theta(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}\right\},\ \text{where } c(s) = \frac{A}{(1-\gamma)\cdot\Delta^*(s)} - 1. \tag{A.159}$$

Similarly to the previous proof, we have the following claims:

Claim I. $\mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$ is a “nice” region, in the sense that, following a gradient update, (i) if $\theta_t \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$; while we also have (ii) $\pi_{\theta_{t+1}}(a^*(s)|s) \geq \pi_{\theta_t}(a^*(s)|s)$.

Claim II. $\mathcal{N}_c(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s) \subset \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$.

Claim III. There exists a finite time $t_0(s) \geq 1$, such that $\theta_{t_0(s)} \in \mathcal{N}_c(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, and thus $\theta_{t_0(s)} \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, which implies $\inf_{t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0(s)}\pi_{\theta_t}(a^*(s)|s)$.

Claim IV. Define $t_0 = \max_s t_0(s)$. Then, we have $\inf_{s\in\mathcal{S},t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0}\min_s \pi_{\theta_t}(a^*(s)|s)$.

Clearly, claim IV suffices to prove the lemma since for any θ, $\min_{s,a}\pi_\theta(a|s) > 0$. In what follows we provide the proofs of these four claims.

Claim I. First we prove part (i) of the claim. If $\theta_t \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$. Suppose $\theta_t \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$. We have $\theta_{t+1} \in \mathcal{R}_3(s)$ by the definition of $\mathcal{R}_3(s)$. We have,

$$Q^{\pi_{\theta_t}}(s, a^*(s)) \geq Q^*(s, a^*(s)) - \Delta^*(s)/2. \tag{A.160}$$

According to monotonic improvement of Eq. (A.213), we have $V^{\pi_{\theta_{t+1}}}(s') \geq V^{\pi_{\theta_t}}(s')$, and

$$\begin{aligned} Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) &= Q^{\pi_{\theta_t}}(s,a^*(s)) + Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_t}}(s,a^*(s)) && \text{(A.161)} \\ &= Q^{\pi_{\theta_t}}(s,a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s,a^*(s)) \cdot [V^{\pi_{\theta_{t+1}}}(s') - V^{\pi_{\theta_t}}(s')] && \text{(A.162)} \\ &\geq Q^{\pi_{\theta_t}}(s,a^*(s)) + 0 && \text{(A.163)} \\ &\geq Q^*(s,a^*(s)) - \Delta^*(s)/2, && \text{(A.164)} \end{aligned}$$

which means $\theta_{t+1} \in \mathcal{R}_2(s)$. Next we prove $\theta_{t+1} \in \mathcal{R}_1(s)$. Note that $\forall a \neq a^*(s)$,

$$\begin{aligned} Q^{\pi_{\theta_t}}(s,a^*(s)) - Q^{\pi_{\theta_t}}(s,a) &= Q^{\pi_{\theta_t}}(s,a^*(s)) - Q^*(s,a^*(s)) + Q^*(s,a^*(s)) - Q^{\pi_{\theta_t}}(s,a) && \text{(A.165)} \\ &\geq -\Delta^*(s)/2 + Q^*(s,a^*(s)) - Q^*(s,a) + Q^*(s,a) - Q^{\pi_{\theta_t}}(s,a) && \text{(A.166)} \\ &\geq -\Delta^*(s)/2 + Q^*(s,a^*(s)) - \max_{a \neq a^*(s)} Q^*(s,a) + Q^*(s,a) - Q^{\pi_{\theta_t}}(s,a) && \text{(A.167)} \\ &= -\Delta^*(s)/2 + \Delta^*(s) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \cdot [V^*(s') - V^{\pi_{\theta_t}}(s')] && \text{(A.168)} \\ &\geq -\Delta^*(s)/2 + \Delta^*(s) + 0 && \text{(A.169)} \\ &= \Delta^*(s)/2. && \text{(A.170)} \end{aligned}$$

Using similar arguments we also have $Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s,a) \geq \Delta^*(s)/2$. According to Lemma 16,

$$\begin{aligned} \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot A^{\pi_{\theta_t}}(s,a) && \text{(A.171)} \\ &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s) \right]. && \text{(A.172)} \end{aligned}$$

Furthermore, since $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}$, we have

$$\pi_{\theta_t}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \geq \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s) \right]. \tag{A.173}$$

Similarly to the first part in the proof for Lemma 4. There are two cases. Case (a): If $\pi_{\theta_t}(a^*(s)|s) \geq \pi_{\theta_t}(a|s)$, then $\theta_t(s,a^*(s)) \geq \theta_t(s,a)$. After an update of the parameters,

$$\theta_{t+1}(s,a^*(s)) = \theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{A.174}$$

$$\geq \theta_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 = \theta_{t+1}(s,a), \tag{A.175}$$

which implies $\pi_{\theta_{t+1}}(a^*(s)|s) \geq \pi_{\theta_{t+1}}(a|s)$. Since $Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s,a) \geq \Delta^*(s)/2 \geq 0$, $\forall a$, we have $Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) = Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - \sum_a \pi_{\theta_{t+1}}(a|s) \cdot Q^{\pi_{\theta_{t+1}}}(s,a) \geq 0$, and

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \right] \geq \pi_{\theta_{t+1}}(a|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s,a) - V^{\pi_{\theta_{t+1}}}(s) \right], \tag{A.176}$$

which is equivalent to $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a)}$, i.e., $\theta_{t+1} \in \mathcal{R}_1(s)$.

Case (b): If $\pi_{\theta_t}(a^*(s)|s) < \pi_{\theta_t}(a|s)$, then by $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}$,

$$\pi_{\theta_t}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \geq \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s) \right] \tag{A.177}$$

$$= \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) + Q^{\pi_{\theta_t}}(s,a) - Q^{\pi_{\theta_t}}(s,a^*(s)) \right], \tag{A.178}$$

which, after rearranging, is equivalent to

$$Q^{\pi_{\theta_t}}(s,a^*(s)) - Q^{\pi_{\theta_t}}(s,a) \geq \left(1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)} \right) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \tag{A.179}$$

$$= \left(1 - \exp\left\{ \theta_t(s,a^*(s)) - \theta_t(s,a) \right\} \right) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right]. \tag{A.180}$$

Since $\theta_{t+1} \in \mathcal{R}_3(s)$, we have,

$$Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \leq \Delta^*(s)/2 \leq Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s,a). \tag{A.181}$$

On the other hand,

$$\theta_{t+1}(s,a^*(s)) - \theta_{t+1}(s,a) \tag{A.182}$$

$$= \theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 - \theta_t(s,a) - \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{A.183}$$

$$\geq \theta_t(s,a^*(s)) - \theta_t(s,a), \tag{A.184}$$

which implies

$$1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\} \leq 1 - \exp\left\{\theta_t(s, a^*(s)) - \theta_t(s, a)\right\}. \tag{A.185}$$

Furthermore, since $1 - \exp\left\{\theta_t(s, a^*(s)) - \theta_t(s, a)\right\} = 1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)} > 0$ (in this case $\pi_{\theta_t}(a^*(s)|s) < \pi_{\theta_t}(a|s)$),

$$\begin{aligned}
&(1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\}) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \qquad (A.186)\\
&\leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a), \qquad (A.187)
\end{aligned}$$

which after rearranging is equivalent to

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \geq \pi_{\theta_{t+1}}(a|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)], \tag{A.188}$$

which means $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a)}$ i.e., $\theta_{t+1} \in \mathcal{R}_1(s)$. Now we have (i) if $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.

Let us now turn to proving part (ii). We have $\pi_{\theta_{t+1}}(a^*(s)|s) \geq \pi_{\theta_t}(a^*(s)|s)$. If $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}$, $\forall a \neq a^*$. After an update of the parameters,

$$\begin{aligned}
\pi_{\theta_{t+1}}(a^*(s)|s) &= \frac{\exp\left\{\theta_{t+1}(s, a^*(s))\right\}}{\sum_a \exp\left\{\theta_{t+1}(s, a)\right\}} && (A.189)\\
&= \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}} && (A.190)\\
&\geq \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}} && (A.191)\\
&\qquad \left(\text{because } \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}\right) && (A.192)\\
&= \frac{\exp\left\{\theta_t(s, a^*(s))\right\}}{\sum_a \exp\left\{\theta_t(s, a)\right\}} = \pi_{\theta_t}(a^*(s)|s). && (A.193)
\end{aligned}$$

Claim II, Claim III, Claim IV. The proof of those claims are exactly the same as [Mei *et al.*, Lemma 9], since they do not involve the update rule. □

Theorem 2. Let Assumption 1 hold and let $\{\theta_t\}_{t\geq 1}$ be generated using Algorithm 1 with

$$\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))}\cdot\frac{1}{\sqrt{S}}, \tag{A.194}$$

where $C_\infty \coloneqq \max_\pi \left\|\frac{d_\mu^\pi}{\mu}\right\|_\infty < \infty$. Denote $C'_\infty \coloneqq \max_\pi \left\|\frac{d_\rho^\pi}{\mu}\right\|_\infty$. Let c be the positive constant from Lemma 8. We have, for all $t \geq 1$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{(V^*(\mu) - V^{\pi_{\theta_1}}(\mu))\cdot C'_\infty}{1-\gamma}\cdot e^{-C\cdot(t-1)}, \tag{A.195}$$

where

$$C = \frac{(1-\gamma)^2\cdot\gamma\cdot c}{12\cdot(1-\gamma)\cdot\gamma + 8\cdot(C_\infty - (1-\gamma))}\cdot\frac{1}{S}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1}. \tag{A.196}$$

Proof. First note that for any θ and μ,

$$\begin{aligned}
d_\mu^{\pi_\theta}(s) &= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[d_\mu^{\pi_\theta}(s)\right] && \text{(A.197)}\\
&= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s|s_0, \pi_\theta, \mathcal{P})\right] && \text{(A.198)}\\
&\geq \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\Pr(s_0 = s|s_0)\right] && \text{(A.199)}\\
&= (1-\gamma)\cdot\mu(s). && \text{(A.200)}
\end{aligned}$$

Next, according to Lemma 19, we have,

$$
\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) && \text{(A.201)} \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) && \text{(A.202)} \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi_\theta}}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) && \text{(A.203)} \\
&\quad \left(\text{Note that } \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) \geq 0\right) && \text{(A.204)} \\
&\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{d_\rho^{\pi_\theta}}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) && \text{(A.205)} \\
&\quad \left(\text{by Eq. (A.197) and } \min_s \mu(s) > 0\right) && \text{(A.206)} \\
&\leq \frac{1}{(1-\gamma)^2} \cdot C_\infty^\prime \cdot \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) && \text{(A.207)} \\
&= \frac{1}{1-\gamma} \cdot C_\infty^\prime \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu)\right]. \qquad \text{(by Lemma 19)} && \text{(A.208)}
\end{aligned}
$$

Denote $\theta_{\zeta_t} \coloneqq \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t)$ with some $\zeta_t \in [0,1]$. And note $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}}$. According to Lemma 6, we have,

$$
\begin{aligned}
&\left| V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| && \text{(A.209)} \\
&\qquad \leq \frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{2\cdot(1-\gamma)\cdot\gamma} \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_{\zeta_t}}}(\mu)}{\partial \theta_{\zeta_t}} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 && \text{(A.210)} \\
&\qquad \leq \frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma} \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2. && \text{(A.211)} \\
&\qquad\quad \text{(by Lemma 7)} && \text{(A.212)}
\end{aligned}
$$

Denote $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. We have,

$$\begin{aligned}
\delta_{t+1} - \delta_t &= V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) && \text{(A.213)} \\
&\le -\Big\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \Big\rangle + && \text{(A.214)} \\
&\frac{3 \cdot (1-\gamma) \cdot \gamma + 2 \cdot (C_\infty - (1-\gamma))}{(1-\gamma) \cdot \gamma} \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 && \text{(A.215)} \\
&= -\frac{(1-\gamma) \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 && \text{(using the value of } \eta) \\
& && \text{(A.216)} \\
&\le -\frac{(1-\gamma) \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}} \cdot \frac{\min_s \pi_{\theta_t}(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\mu^{\pi^*} / d_\mu^{\pi_{\theta_t}} \right\|_\infty} \cdot \delta_t && \text{(by Lemma 5)} \\
& && \text{(A.217)} \\
&\le -\frac{(1-\gamma)^2 \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) \cdot \delta_t, \\
& && \text{(A.218)}
\end{aligned}$$

where the last inequality is by $d_\mu^{\pi_{\theta_t}}(s) \ge (1-\gamma) \cdot \mu(s)$ (cf. Eq. (A.197)). According to Lemma 8, $c = \inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) > 0$. Therefore we have,

$$\begin{aligned}
V^*(\mu) - V^{\pi_{\theta_t}}(\mu) &\le (V^*(\mu) - V^{\pi_{\theta_1}}(\mu)) \cdot && \text{(A.219)} \\
&\exp\left\{ -\frac{(1-\gamma)^2 \cdot \gamma \cdot c \cdot (t-1)}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right\}, && \text{(A.220)}
\end{aligned}$$

which leads to the final result,

$$\begin{aligned}
V^*(\rho) - V^{\pi_{\theta_t}}(\rho) &\le \frac{(V^*(\mu) - V^{\pi_{\theta_1}}(\mu)) \cdot C_\infty'}{1-\gamma} \cdot && \text{(A.221)} \\
&\exp\left\{ -\frac{(1-\gamma)^2 \cdot \gamma \cdot c \cdot (t-1)}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right\}, && \text{(A.222)}
\end{aligned}$$

thus, finishing the proof of convergence rate. □

A.2 Proofs for Chapter 6

Lemma 9 (NŁ) **.** Denote $u(\theta) \coloneqq \min_i \left\{ \pi_i \cdot (1 - \pi_i) \right\}$, and $v \coloneqq \min_i \left\{ \pi_i^* \cdot (1 - \pi_i^*) \right\}$. We have, for all $i \in [N]$,

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \ge 8 \cdot u(\theta) \cdot \min \left\{ u(\theta), v \right\} \cdot \sqrt{\lambda_\phi} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\pi_i - \pi_i^* \right)^2 \right]^{\frac{1}{2}}, \tag{A.223}$$

where λ_ϕ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$.

Proof. Denote $\pi_i' \coloneqq \sigma(z_i')$, where $z_i' \coloneqq \phi_i^\top \theta + \zeta \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right)$ for some $\zeta \in [0, 1]$. We have,

$$\left(\pi_i - \pi_i^* \right)^2 = \left(\pi_i - \pi_i^* \right) \cdot \frac{d\sigma(z_i')}{dz_i'} \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right) \qquad \text{(by the mean value theorem)} \tag{A.224}$$

$$= \pi_i' \cdot (1 - \pi_i') \cdot \left(\pi_i - \pi_i^* \right) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right) \tag{A.225}$$

$$\le \frac{1}{4} \cdot \left(\pi_i - \pi_i^* \right) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right). \tag{A.226}$$

$$\left(\text{Since } x \cdot (1 - x) \le \frac{1}{4},\ \forall x \in [0, 1];\ \left(\pi_i - \pi_i^* \right) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^* \right) \ge 0 \right) \tag{A.227}$$

Therefore we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \leq \frac{1}{4N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \qquad \text{(by Eq. (A.224))} \tag{A.228}$$

$$= \frac{1}{4N} \cdot \sum_{i=1}^{N} \frac{1}{\pi_i \cdot (1 - \pi_i)} \cdot \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \tag{A.229}$$

$$\leq \frac{1}{4N} \cdot \frac{1}{\min_i \pi_i \cdot (1 - \pi_i)} \cdot \sum_{i=1}^{N} \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \tag{A.230}$$

$$\left(\text{Note that}\, (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \geq 0\right) \tag{A.231}$$

$$= \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1 - \pi_i)} \cdot \left(\frac{2}{N} \cdot \sum_{i=1}^{N} \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i\right)^\top (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{A.232}$$

$$= \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1 - \pi_i)} \cdot \left(\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right)^\top (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{A.233}$$

$$\left(\text{where } \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{2}{N} \cdot \sum_{i=1}^{N} \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i\right) \tag{A.234}$$

$$\leq \frac{1}{8} \cdot \frac{1}{\min_i \pi_i \cdot (1 - \pi_i)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2 \qquad \text{(by Cauchy-Schwarz)} \tag{A.235}$$

$$= \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2 \,, \qquad \left(u(\theta) \coloneqq \min_i \left\{\pi_i \cdot (1 - \pi_i)\right\}\right) \tag{A.236}$$

where $v_{\phi,\perp}$ is orthogonal to the space Span $\{\phi_1, \phi_2, \ldots, \phi_N\}$, and $\theta - \theta^* - c \cdot v_{\phi,\perp}$ refers to the vector after cutting off all the components $v_{\phi,\perp}$ from $\theta - \theta^*$, such that $\theta - \theta^* - c \cdot v_{\phi,\perp} \in \text{Span}\left\{\phi_1, \phi_2, \ldots, \phi_N\right\}$. Next, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\pi_i - \pi_i^*\right)^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{d\sigma(z_i')}{dz_i'}\right)^2 \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right)^2 \qquad \text{(by the mean value theorem)} \tag{A.237}$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i')^2 \cdot (1 - \pi_i')^2 \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right)^2 \qquad \text{(by Eq. (A.224))} \tag{A.238}$$

$$\geq \min_i \left\{(\pi_i')^2 \cdot (1 - \pi_i')^2\right\} \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right)^2 \tag{A.239}$$

$$= \min_i \left\{(\pi_i')^2 \cdot (1 - \pi_i')^2\right\} \cdot (\theta - \theta^*)^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top\right) (\theta - \theta^*) \tag{A.240}$$

$$= \min_i \left\{(\pi_i')^2 \cdot (1 - \pi_i')^2\right\} \cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top\right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{A.241}$$

$$\geq \min\left\{u(\theta)^2, v^2\right\} \cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top\right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{A.242}$$

$$\left(\text{Note } v \coloneqq \min_i \left\{\pi_i^* \cdot (1 - \pi_i^*)\right\}\right) \tag{A.243}$$

$$\geq \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi \cdot \left\|\theta - \theta^* - c \cdot v_{\phi,\perp}\right\|_2^2, \tag{A.244}$$

where λ_ϕ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$. Therefore, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\pi_i - \pi_i^*\right)^2 \leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 \cdot \left\|\theta - \theta^* - c \cdot v_{\phi,\perp}\right\|_2 \qquad \text{(by Eq. (A.228))} \tag{A.245}$$

$$\leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 \cdot \frac{1}{\min\left\{u(\theta), v\right\}} \cdot \frac{1}{\sqrt{\lambda_\phi}} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\pi_i - \pi_i^*\right)^2\right]^{\frac{1}{2}}, \tag{A.246}$$

$$\text{(by Eq. (A.237))} \tag{A.247}$$

which implies,

$$\left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 \geq 8 \cdot u(\theta) \cdot \min\left\{u(\theta), v\right\} \cdot \sqrt{\lambda_\phi} \cdot \left[\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\pi_i - \pi_i^*\right)^2\right]^{\frac{1}{2}}. \qquad \square$$

Lemma 10. Denote $u(\theta) \coloneqq \min_i \{\pi_i \cdot (1 - \pi_i)\}$, $v \coloneqq \min_i \{\pi_i^* \cdot (1 - \pi_i^*)\}$, and λ_ϕ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^N \phi_i \phi_i^\top$. We have, $\mathcal{L}(\theta)$ satisfies β smoothness with

$$\beta = \frac{3}{8} \cdot \max_{i \in [N]} \|\phi_i\|_2^2, \tag{A.248}$$

and $\beta(\theta)$ NS with

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left(\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right). \tag{A.249}$$

where

$$L_1 = \frac{\max_i \|\phi_i\|_2^2}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}}, \text{ and } L_0 = \frac{17 \cdot \max_i \|\phi_i\|_2^2}{512 \cdot u(\theta)^2 \cdot \min\{u(\theta)^2, v^2\} \cdot \lambda_\phi}. \tag{A.250}$$

Proof. Note that the gradient of $\mathcal{L}(\theta)$ is,

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{2}{N} \cdot \sum_{i=1}^N \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i \in \mathbb{R}^d. \tag{A.251}$$

Denote the second order derivative (Hessian) of $\mathcal{L}(\theta)$ as,

$$S(\theta) \coloneqq \frac{\partial}{\partial \theta} \left\{ \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\} \in \mathbb{R}^{d \times d}. \tag{A.252}$$

For all $j, k \in [d]$, we calculate the corresponding component value of $S(\theta)$ matrix as follows,

$$S_{(j,k)} = \frac{d}{d\theta(k)} \left\{ \frac{2}{N} \cdot \sum_{i=1}^N \pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i(j) \right\} \tag{A.253}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \frac{d\{\pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*)\}}{d\theta(k)} \cdot \phi_i(j) \tag{A.254}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \frac{d\{\pi_i \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*)\}}{d\{\phi_i^\top \theta\}} \cdot \frac{d\{\phi_i^\top \theta\}}{d\theta(k)} \cdot \phi_i(j) \tag{A.255}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \left[\pi_i \cdot (1 - \pi_i)^2 \cdot (\pi_i - \pi_i^*) - \pi_i^2 \cdot (1 - \pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1 - \pi_i)^2 \right] \cdot \phi_i(k) \cdot \phi_i(j) \tag{A.256}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \left[\pi_i \cdot (1 - \pi_i) \cdot (1 - 2\pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1 - \pi_i)^2 \right] \cdot \phi_i(k) \cdot \phi_i(j). \tag{A.257}$$

To calculate the smoothness coefficient, take a vector $z \in \mathbb{R}^d$. We have,

$$\left| z^\top S(\theta) z \right| = \left| \sum_{j=1}^{d} \sum_{k=1}^{d} S_{(j,k)} \cdot z(j) \cdot z(k) \right| \tag{A.258}$$

$$= \left| \frac{2}{N} \cdot \sum_{i=1}^{N} \left[\pi_i \cdot (1-\pi_i) \cdot (1-2\pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1-\pi_i)^2 \right] \cdot \left(\phi_i^\top z \right)^2 \right| \tag{A.259}$$

$$\text{(by Eq. (A.253))} \tag{A.260}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z \right)^2 \cdot \sum_{i=1}^{N} \left| \pi_i \cdot (1-\pi_i) \cdot (1-2\pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1-\pi_i)^2 \right| \tag{A.261}$$

$$\text{(by Hölder's inequality)} \tag{A.262}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z \right)^2 \cdot \sum_{i=1}^{N} \left[\pi_i \cdot (1-\pi_i) \cdot |1-2\pi_i| \cdot |\pi_i - \pi_i^*| + \pi_i^2 \cdot (1-\pi_i)^2 \right] \tag{A.263}$$

$$\text{(by triangle inequality)} \tag{A.264}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z \right)^2 \cdot \sum_{i=1}^{N} \left[\frac{1}{8} + \frac{1}{16} \right] \tag{A.265}$$

$$\text{(Note that} x \cdot (1-x) \leq 1/4, \text{ and } x \cdot (1-x) \cdot |1-2x| \leq 1/8, \ \forall x \in [0,1]) \tag{A.266}$$

$$= \frac{3}{8} \cdot \max_i \left[\phi_i^\top \left(\frac{z}{\|z\|_2} \right) \right]^2 \cdot \|z\|_2^2 \tag{A.267}$$

$$\leq \frac{3}{8} \cdot \max_i \|\phi_i\|_2^2 \cdot \|z\|_2^2 . \tag{A.268}$$

Therefore, $\mathcal{L}(\theta)$ satisfies β (uniform) smoothness with $\beta = \frac{3}{8} \cdot \max_i \|\phi_i\|_2^2$. Next, we calculate the NS. We have,

$$\sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \cdot \mathcal{L}(\theta) = \sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \cdot \frac{1}{N} \cdot \sum_{j=1}^{N} (\pi_j - \pi_j^*)^2 \tag{A.269}$$

$$\leq \frac{N}{16} \cdot \frac{1}{N} \cdot \sum_{j=1}^{N} (\pi_j - \pi_j^*)^2 \tag{A.270}$$

$$\leq \frac{N}{16} \cdot \frac{1}{64 \cdot u(\theta)^2 \cdot \min\{u(\theta)^2, v^2\} \cdot \lambda_\phi} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 , \qquad \text{(by Lemma 9)} \tag{A.271}$$

which implies,

$$\sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \leq \frac{N}{2} \cdot \frac{1}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta). \tag{A.272}$$

According to Eq. (A.237), we have

$$\begin{aligned}
\sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^{3/2}} &\geq \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)}} \cdot (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \frac{1}{\mathcal{L}(\theta)^{3/4}} && \text{(A.273)} \\
&= (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\mathcal{L}(\theta)^{5/4}} && \text{(A.274)} \\
&= N \cdot (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \frac{\mathcal{L}(\theta)}{\mathcal{L}(\theta)^{5/4}} && \text{(A.275)} \\
&\geq N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}. \qquad (\mathcal{L}(\theta) \in (0,1]) && \text{(A.276)}
\end{aligned}$$

Therefore we have,

$$\sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |1-2\pi_i| \cdot |\pi_i - \pi_i^*| \le \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*| \tag{A.277}$$

$$\le \left(\sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*|\right) \cdot \left(\sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^{3/2}}\right) \cdot \Bigg(\tag{A.278}$$

$$\frac{1}{N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}}\Bigg) \tag{A.279}$$

$$= \frac{1}{N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left(\sum_{i=1}^{N} \frac{\pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*|}{\sqrt{\|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2}}\right) \cdot \tag{A.280}$$

$$\left(\sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2}\right) \tag{A.281}$$

$$\le \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left(\sum_{i=1}^{N} \frac{\pi_i^2 \cdot (1-\pi_i)^2 \cdot (\pi_i - \pi_i^*)^2}{2 \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2} + \right. \tag{A.282}$$

$$\left. \frac{(\pi_i - \pi_i^*)^4}{2 \cdot \mathcal{L}(\theta) \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^2}\right) \tag{A.283}$$

$$\le \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left(\frac{1}{32} \cdot \sum_{i=1}^{N} \frac{\pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot (\phi_i^\top \theta - \phi_i^\top \theta^*)}{\|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2}\right) \tag{A.284}$$

$$+ \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left(\frac{1}{32 \cdot u(\theta)^2} \cdot \right. \tag{A.285}$$

$$\left. \sum_{i=1}^{N} \frac{\pi_i^2 \cdot (1-\pi_i)^2 \cdot (\pi_i - \pi_i^*)^2 \cdot (\phi_i^\top \theta - \phi_i^\top \theta^*)^2}{\mathcal{L}(\theta) \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|^2}\right) \tag{A.286}$$

$$\le \frac{N}{64 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 + \tag{A.287}$$

$$\frac{N}{64 \cdot u(\theta)^2 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta), \tag{A.288}$$

where the second inequality is according to,

$$\left(\sum_{i=1}^{N} a_i\right) \cdot \left(\sum_{i=1}^{N} b_i\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i \cdot b_j \le \frac{1}{2} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \left(a_i^2 + b_j^2\right) = \frac{N}{2} \cdot \sum_{i=1}^{N} \left(a_i^2 + b_i^2\right), \tag{A.289}$$

and the last inequality is from the intermediate results in Eq. (A.228). Combining Eqs. (A.258), (A.272) and (A.277), we have

$$\left|z^\top S(\theta) z\right| \le \frac{2}{N} \cdot \max_i \left(\phi_i^\top z\right)^2 \cdot \left[\sum_{i=1}^N \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*| + \sum_{i=1}^N \pi_i^2 \cdot (1-\pi_i)^2\right] \tag{A.290}$$

$$\le \max_i \left(\phi_i^\top z\right)^2 \cdot \left(\frac{1}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 + \right. \tag{A.291}$$

$$\left. \frac{17}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right) \tag{A.292}$$

$$\le \max_i \|\phi_i\|_2^2 \cdot \|z\|_2^2 \cdot \left(\frac{1}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 + \right. \tag{A.293}$$

$$\left. \frac{17}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right). \tag{A.294}$$

Therefore, $\mathcal{L}(\theta)$ satisfies $\beta(\theta)$ NS with

$$\beta(\theta) = L_1 \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 + L_0 \cdot \left(\left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right), \tag{A.295}$$

where

$$L_1 = \frac{\max_i \|\phi_i\|_2^2}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}}, \text{ and } L_0 = \frac{17 \cdot \max_i \|\phi_i\|_2^2}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi}.$$

□

Theorem 4. With $\eta = 1/\beta$, GD update satisfies for all $t \ge 1$, $\mathcal{L}(\theta_t) \le \mathcal{L}(\theta_1) \cdot e^{-C^2 \cdot (t-1)}$. With $\eta \in \Theta(1)$, GNGD update satisfies for all $t \ge 1$, $\mathcal{L}(\theta_t) \le \mathcal{L}(\theta_1) \cdot e^{-C \cdot (t-1)}$, where $C \in (0, 1)$, i.e., GNGD is strictly faster than GD.

Proof. Combining Lemmas 9 and 10, and the second part of (2b) in Section 4.2.1, we have the results for GD. Using the fourth part of (2b) in Section 4.2.1, we have the results for GNGD. □

A.3 Miscellaneous Extra Supporting Results

Lemma 12 (Descent lemma for smooth function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function, $\theta \in \mathbb{R}^d$ and $\theta' = \theta - \eta \cdot \frac{\partial f(\theta)}{\partial \theta}$. We have, for any $0 < \eta < 2/\beta$,*

$$f(\theta') \le f(\theta). \tag{A.296}$$

In particular, for $\eta = \frac{1}{\beta}$, we have,

$$f(\theta') \le f(\theta) - \frac{1}{2\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \tag{A.297}$$

Proof. According to Definition 4, we have,

$$\left| f(\theta') - f(\theta) - \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle \right| \le \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2, \tag{A.298}$$

which implies,

$$\begin{aligned} f(\theta') - f(\theta) &\le \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle + \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2 && \text{(A.299)} \\ &= \eta \cdot \left(-1 + \frac{\beta}{2} \cdot \eta \right) \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 \qquad \left(\theta' = \theta - \eta \cdot \frac{\partial f(\theta)}{\partial \theta} \right) && \text{(A.300)} \\ &\le 0 \qquad \left(0 < \eta < \frac{2}{\beta} \right). && \text{(A.301)} \end{aligned}$$

Let $\eta = \frac{1}{\beta}$ in Eq. (A.300), we have Eq. (A.297). □

Lemma 13 (Descent lemma for NS function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies NS with $\beta(\theta) > 0$, for all $\theta \in \mathbb{R}^d$ and $\theta' = \theta - \frac{1}{\beta(\theta)} \cdot \frac{\partial f(\theta)}{\partial \theta}$. We have,*

$$f(\theta') \le f(\theta) - \frac{1}{2 \cdot \beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \tag{A.302}$$

Proof. According to Definition 5, we have,

$$\begin{aligned} f(\theta') - f(\theta) &\le \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle + \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2 && \text{(A.303)} \\ &= -\frac{1}{\beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 + \frac{1}{2 \cdot \beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 \qquad \left(\theta' = \theta - \frac{1}{\beta(\theta)} \cdot \frac{\partial f(\theta)}{\partial \theta} \right) && \text{(A.304)} \\ &= -\frac{1}{2 \cdot \beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \end{aligned}$$

□

Lemma 14. *Given any $\alpha > 0$, we have, for all $x \in [0, 1]$,*

$$\frac{1}{\alpha} \cdot (1 - x^\alpha) \ge x^\alpha \cdot (1 - x). \tag{A.305}$$

Proof. Define $f : x \mapsto \frac{1}{\alpha} \cdot (1 - x^{\alpha}) - x^{\alpha} \cdot (1 - x)$. We show that $f(x) \geq 0$ for all $x \in [0, 1]$. Note that,

$$f(0) = \frac{1}{\alpha} > 0, \text{ and } f(1) = 0. \tag{A.306}$$

On the other hand,

$$\begin{aligned} f'(x) &= -x^{\alpha-1} - \alpha \cdot x^{\alpha-1} \cdot (1 - x) + x^{\alpha} && \text{(A.307)} \\ &= -x^{\alpha-1} \cdot [1 + \alpha \cdot (1 - x) - x] && \text{(A.308)} \\ &= -x^{\alpha-1} \cdot (1 + \alpha) \cdot (1 - x) && \text{(A.309)} \\ &\leq 0, \qquad (\alpha > 0, \text{ and } x \in [0, 1]) && \text{(A.310)} \end{aligned}$$

which means f is monotonically decreasing over $[0, 1]$. Therefore $f(x) \geq 0$ for all $x \in [0, 1]$, finishing the proof. □

Lemma 15. *Given any $\alpha > 0$, we have, for all $x \in \left[\frac{2\alpha+1}{2\alpha+2}, 1\right]$,*

$$\frac{1}{2\alpha} \cdot (1 - x^{\alpha}) \leq x^{\alpha} \cdot (1 - x) . \tag{A.311}$$

Proof. Define $g : x \mapsto x^{\alpha} \cdot (1 - x) - \frac{1}{2\alpha} \cdot (1 - x^{\alpha})$. The derivative of g is,

$$\begin{aligned} g'(x) &= \alpha \cdot x^{\alpha-1} \cdot (1 - x) - x^{\alpha} + (1/2) \cdot x^{\alpha-1} && \text{(A.312)} \\ &= x^{\alpha-1} \cdot [\alpha \cdot (1 - x) - x + 1/2] && \text{(A.313)} \\ &= x^{\alpha-1} \cdot [(1 + \alpha) \cdot (1 - x) - 1/2] . && \text{(A.314)} \end{aligned}$$

Then we have,

$$\begin{aligned} &g'(x) > 0 \text{ for all } x \in [0, (2\alpha + 1)/(2\alpha + 2)) , \text{ and} && \text{(A.315)} \\ &g'(x) \leq 0 \text{ for all } x \in [(2\alpha + 1)/(2\alpha + 2), 1] , && \text{(A.316)} \end{aligned}$$

which means g is monotonically increasing over $[0, (2\alpha + 1)/(2\alpha + 2))$ and decreasing over $[(2\alpha+1)/(2\alpha+2), 1]$. On the other hand,

$$\begin{aligned}
g((2\alpha+1)/(2\alpha+2)) &= \left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha} \cdot \left(1 - \frac{2\alpha+1}{2\alpha+2}\right) - \frac{1}{2\alpha} \cdot \left[1 - \left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha}\right] && \text{(A.317)} \\
&= \frac{1}{2\alpha} \cdot \left[\left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha} \cdot \frac{2\alpha+1}{\alpha+1} - 1\right] && \text{(A.318)} \\
&= \frac{1}{2\alpha} \cdot \left[\exp\left\{\log\left(\frac{2\alpha+1}{\alpha+1}\right) - \alpha \cdot \log\left(1 + \frac{1}{2\alpha+1}\right)\right\} - 1\right] && \text{(A.319)} \\
&\geq \frac{1}{2\alpha} \cdot \left[\exp\left\{\log\left(\frac{2\alpha+1}{\alpha+1}\right) - \frac{\alpha}{2\alpha+1}\right\} - 1\right] \qquad (1 + x \leq e^{x}) && \text{(A.320)} \\
&\geq \frac{1}{2\alpha} \cdot \left[\exp\left\{\frac{\alpha}{2\alpha+1} - \frac{\alpha}{2\alpha+1}\right\} - 1\right] \qquad (\log(x) \geq 1 - 1/x \text{ for } x > 0) && \text{(A.321)} \\
&= 0. && \text{(A.322)}
\end{aligned}$$

Also note that $g(1) = 0$. Therefore we have $g(x) \geq 0$ for all $x \in [(2\alpha+1)/(2\alpha+2), 1]$, finishing the proof. □

Lemma 16. *Denote $H(\pi) := diag(\pi) - \pi\pi^{\top}$. Softmax policy gradient w.r.t. θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot), \quad \forall s \in \mathcal{S}. \tag{A.323}$$

Proof. See the proof in [Mei *et al.*, Lemma 1]. Here's the proof for completeness.

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_{\mu}^{\pi_\theta}} \left[\sum_{a} \frac{\partial \pi_\theta(a|s')}{\partial \theta} \cdot Q^{\pi_\theta}(s', a)\right]. \tag{A.324}$$

For $s' \neq s$, $\frac{\partial \pi_\theta(a|s')}{\partial \theta(s,\cdot)} = \mathbf{0}$ since $\pi_\theta(a|s')$ does not depend on $\theta(s,\cdot)$. Therefore, we have,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \sum_{s'} d_\mu^{\pi_\theta}(s') \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s')}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s',a) \right] \tag{A.325}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] \qquad \left(\frac{\partial \pi_\theta(a|s')}{\partial \theta(s,\cdot)} = \mathbf{0}, \ \forall s' \neq s \right) \tag{A.326}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} \right)^\top Q^{\pi_\theta}(s,\cdot) \tag{A.327}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot). \qquad (H(\pi_\theta) \text{ is the Jacobian of } \theta \mapsto \text{softmax}(\theta)) \tag{A.328}$$

Note that in one-state MDPs, we have,

$$\frac{d\pi_\theta^\top r}{d\theta} = \left(\frac{d\pi_\theta}{d\theta} \right)^\top r = H(\pi_\theta) r. \qquad \square$$

Lemma 17. *Softmax policy gradient norm is*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \|_2^2 \right]^{\frac{1}{2}}. \tag{A.329}$$

Proof. We have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \left[\sum_{s,a} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \tag{A.330}$$

$$= \left[\sum_s \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2^2 \right]^{\frac{1}{2}} \tag{A.331}$$

$$= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \| H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \|_2^2 \right]^{\frac{1}{2}}. \qquad \text{(by Lemma } \square \text{16)}$$

Lemma 18 (Performance difference lemma [Kakade and Langford)**.** *] For any policies π and π',*

$$V^{\pi'}(\rho) - V^\pi(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^\pi(s,a) \tag{A.332}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a \pi'(a|s) \cdot A^\pi(s,a). \tag{A.333}$$

Proof. According to the definition of value function,

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) &= \sum_a \pi'(a|s) \cdot Q^{\pi'}(s,a) - \sum_a \pi(a|s) \cdot Q^{\pi}(s,a) && \text{(A.334)} \\
&= \sum_a \pi'(a|s) \cdot \Big(Q^{\pi'}(s,a) - Q^{\pi}(s,a) \Big) + \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi}(s,a) && \text{(A.335)} \\
&= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi}(s,a) + && \text{(A.336)} \\
&\quad \gamma \sum_a \pi'(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \Big[V^{\pi'}(s') - V^{\pi}(s') \Big] && \text{(A.337)} \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} (\pi'(a'|s') - \pi(a'|s')) \cdot Q^{\pi}(s',a') && \text{(A.338)} \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot (Q^{\pi}(s',a') - V^{\pi}(s')) && \text{(A.339)} \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot A^{\pi}(s',a'). && \square
\end{aligned}
$$

Lemma 19 (Value sub-optimality lemma). *For any policy π,*

$$
V^*(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi}(s) \sum_a (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s,a). \tag{A.340}
$$

Proof. See the proof in [Mei *et al.*, Lemma 21]. We include a proof for completeness.

We denote $V^*(s) \coloneqq V^{\pi^*}(s)$ and $Q^*(s,a) \coloneqq Q^{\pi^*}(s,a)$ for conciseness. We have, for any policy π,

$$
\begin{aligned}
V^*(s) - V^{\pi}(s) &= \sum_a \pi^*(a|s) \cdot Q^*(s,a) - \sum_a \pi(a|s) \cdot Q^{\pi}(s,a) && \text{(A.341)} \\
&= \sum_a (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s,a) + \sum_a \pi(a|s) \cdot (Q^*(s,a) - Q^{\pi}(s,a)) && \text{(A.342)} \\
&= \sum_a (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s,a) + && \text{(A.343)} \\
&\quad \gamma \sum_a \pi(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left[V^{\pi^*}(s') - V^{\pi}(s') \right] && \text{(A.344)} \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \sum_{a'} (\pi^*(a'|s') - \pi(a'|s')) \cdot Q^*(s',a'). && \square
\end{aligned}
$$

A.4 Proof of Convergence of Algorithms

Proof of Theorem 5

This proof is originally proved by [49], but I describe it here in detail because it will be useful for later proofs for the proposed algorithms. First, show the following Lemma :

Lemma 20. *For the iterative procedure*

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \left[u \left(r_t + \gamma \cdot \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t) \right) - x_0 \right] \tag{A.345}$$

where $\alpha_t \geq 0$ satisfy that for any (s, a), $\sum_{t=0}^{\infty} \alpha_t(s, a) = \infty$; and $\sum_{t=0}^{\infty} \alpha_t^2(s, a) < \infty$, then $Q_t \to Q^*$, where Q^* is the solution of the Bellman equation

$$(H^A Q^*)(s, a) = \alpha \cdot \mathop{\mathbb{E}}_{s,a} \left[\tilde{u} \left(r_t + \gamma \cdot \max_a Q^*(s_{t+1}, a) - Q^*(s, a) \right) \right] + Q^*(s, a) = Q^*(s, a) \tag{A.346}$$

$$\forall (s, a) \tag{A.347}$$

If Lemma 20 holds, then it's shown in [49] that the corresponding policy optimizes the objective function Eq. (7.2).

Before proving the convergence, consider a more general update rule

$$q_{t+1}(i) = (1 - \alpha_t(i)) q_t(i) + \alpha_t(i) \left[(H q_t)(i) + w_t(i) \right] \tag{A.348}$$

where i is the independent variable (e.g., in single agent Q learning, it's the state-action pair (s, a)), $q_t \in \mathbb{R}^d$, $H : \mathbb{R}^d \to \mathbb{R}^d$ is an operator, w_t denotes some random noise term and α_t is learning rate with the understanding that $\alpha_t(i) = 0$ if $q(i)$ is not updated at time t. Denote by $\mathcal{F}_t$ the history of the algorithm up to time t,

$$\mathcal{F}_t = \{ q_0(i), ..., q_t(i), w_0(i), ..., w_t(i), \alpha_0(i), ..., \alpha_t(i) \} \tag{A.349}$$

Recall the following essential proposition :

Proposition 1. *[9] Let q_t be the sequence generated by the iteration Eq.* (A.348), *if assuming the following hold :*

(a) The Learning rates $\alpha_t(i)$ satisfy :

$$\alpha_t(i) \geq 0; \qquad \sum_{t=0}^{\infty} \alpha_t(i) = \infty; \qquad \sum_{t=0}^{\infty} \alpha_t^2(i) < \infty; \quad \forall i \tag{A.350}$$

(b) The noise terms $w_t(i)$ satisfy

(i) $\mathbb{E}[w_t(i)|\mathcal{F}_t] = 0$ for all i and t;

(ii) There exist constants A and B such that $\mathbb{E}[w_t^2(i)|\mathcal{F}_t] \leq A + B \|q_t\|^2$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$.

(c) The mapping H is a contraction under sup-norm.

Then q_t converges to the unique solution q^* of the equation $Hq^* = q^*$ with probability 1.

In order to apply Proposition 1, reformulate the update rule Eq. (7.4) by letting

$$q_{t+1}(s,a) = \left(1 - \frac{\alpha_t(s,a)}{\alpha}\right) q_t(s,a) + \frac{\alpha_t(s,a)}{\alpha}[\alpha \cdot u(d_t) - \alpha \cdot x_0 + q_t(s,a)] \tag{A.351}$$

where $\tilde{u}(x) := u(x) - x_0$; $d_t := r_t + \gamma \cdot \underset{a}{\max} q_t(s_{t+1}, a) - q_t(s,a)$. And we set

$$(Hq_t)(s,a) = \alpha \cdot \mathop{\mathbb{E}}_{s,a}\left[\tilde{u}\left(r_t + \gamma \cdot \underset{a}{\max} q_t(s_{t+1}, a) - q_t(s,a)\right)\right] + q_t(s,a) \tag{A.352}$$

$$w_t(s,a) = \alpha \cdot \tilde{u}(d_t) - \alpha \cdot \mathop{\mathbb{E}}_{s,a}\left[\tilde{u}(r_t + \gamma \cdot \underset{a}{\max} q_t(s', a) - q_t(s,a))\right] \tag{A.353}$$

where s' is sampled from $\mathcal{T}[\cdot|s,a]$.

More explicitly, Hq is defined as

$$(Hq)(s,a) = \alpha \cdot \sum_{s'} \mathcal{T}[s'|s,a] \cdot \tilde{u}\left(r(s,a) + \gamma \cdot \max_{a'} q(s',a') - q(s,a)\right) + q(s,a) \tag{A.354}$$

Next, show that H is a contraction under sup-norm.

Note that here it's assumed that the utility function satisfy :

Assumption 2. *(i) The utility function u is strictly increasing and there exists some $y_0 \in \mathbb{R}$ such that $u(y_0) = x_0$.*

(ii) There exist positive constants ϵ, L such that $0 < \epsilon \leq \frac{u(x)-u(y)}{x-y} \leq L$ for all $x \neq y \in \mathbb{R}$.

Note that Assumption 2 seems to exclude several important types of utility functions like the exponential function $u(x) = exp(c \cdot x)$ since it does not satisfy the global Lipschitz. But this can be solved by a truncation when x is very large and by an approximation when x is very close to 0. For more details see Shen *et al.* (2014).

And it's also assumed that the immediate reward r_t always satisfy a sub-Gaussian tail assumption. This allows the reward to be unbounded, which is closer to practical settings with tail events, for example, in financial markets. :

Assumption 3. *r_t is uniformly sub-Gaussian over t with variance proxy σ^2, i.e.,*

$$\mathbb{E}[r_t] = 0 \tag{A.355}$$

$$\mathbb{E}[exp(c \cdot r_t)] \leq exp\left(\frac{\sigma^2 c^2}{2}\right) \qquad \forall c \in \mathbb{R} \tag{A.356}$$

The above uniform sub-Gaussian assumption is equivalent to the following form, commonly seen in statistics and machine learning: there exists $C > 0, \alpha$ such that for every $K > 0$ and every r_t, we have:

$$\mathbb{P}(|r_t| > K) \leq Ce^{-\alpha K^2} \tag{A.357}$$

Proposition 2. *Suppose that Assumption 2 and Assumption 3 hold and $0 < \alpha < \min(L^{-1}, 1)$. Then there exists a real number $\bar{\alpha} \in [0, 1)$ such that for all $q, q' \in \mathbb{R}^d$, $\|Hq - Hq'\|_\infty \leq \bar{\alpha} \|q - q'\|_\infty$.*

Proof. Define $v(s) := \max_a q(s, a)$ and $v'(s) := \max_a q'(s, a)$. Thus,

$$|v(s) - v'(s)| \leq \max_{s,a} |q(s, a) - q'(s, a)| = \|q - q'\|_\infty \tag{A.358}$$

By Assumption 2, and the monotonicity of $\tilde{u}$, there exists a $\xi_{(x,y)} \in [\epsilon, L]$ such that $\tilde{u}(x) - \tilde{u}(y) = \xi_{(x,y)} \cdot (x - y)$. Then we can obtain

$$(Hq)(s,a) - (Hq')(s,a) \tag{A.359}$$

$$= \sum_{s'} \mathcal{T}[s'|s,a] \cdot \Big\{ \alpha \xi_{(s,a,s',q,q')} \cdot [\gamma v(s') - \gamma v'(s') - q(s,a) + q'(s,a)] + (q(s,a) - q'(s,a)) \Big\} \tag{A.360}$$

$$\leq \left(1 - \alpha(1-\gamma) \sum_{s'} \mathcal{T}[s'|s,a] \cdot \xi_{(s,a,s',q,q')} \right) \|q - q'\|_\infty \tag{A.361}$$

$$\leq (1 - \alpha(1-\gamma)\epsilon) \|q - q'\|_\infty \tag{A.362}$$

Hence, $\bar{\alpha} = 1 - \alpha(1-\gamma)\epsilon$ is the required constant. □

Now that it's already shown that the requirements (a) and (c) of Proposition 1 hold, it remains to check (b). By Eq. (A.352), $\mathbb{E}[w_t(s,a)|\mathcal{F}_t] = 0$. Next, prove (b)(ii).

$$\mathbb{E}[w_t^2(s,a)|\mathcal{F}_t] = \alpha^2 \, \mathbb{E}[(\tilde{u}(d_t))^2|\mathcal{F}_t] - \alpha^2 (\mathbb{E}[\tilde{u}(d_t)|\mathcal{F}_t])^2 \tag{A.363}$$

$$\leq \alpha^2 \, \mathbb{E}[(\tilde{u}(d_t))^2|\mathcal{F}_t] \tag{A.364}$$

By Assumption 3, $\mathbb{E}\,|r_t| < (2\sigma)^{\frac{1}{2}} \Gamma(\frac{1}{2})$, where $\Gamma(\cdot)$ is the Gamma function (see [14] for details). Denote the upper bound for $\mathbb{E}[|r_t|]$ as R_1. Then $\mathbb{E}[|d_t|] \leq R_1 + 2\|q_t\|_\infty$, due to Assumption 2, it implies that

$$\mathbb{E}\left[|\tilde{u}(d_t) - \tilde{u}(0)|\right] \leq \mathbb{E}\left[L \cdot d_t\right] \leq L(R_1 + 2\|q_t\|_\infty) \tag{A.365}$$

Hence by triangle inequality,

$$\mathbb{E}[|\tilde{u}(d_t)|] \leq \tilde{u}(0) + LR_1 + 2L\|q_t\|_\infty \tag{A.366}$$

And since

$$(a+b)^2 \leq 2a^2 + 2b^2 \qquad \forall a, b \in \mathbb{R} \tag{A.367}$$

, we have

$$(|\tilde{u}(0)| + LR_1 + 2L\|q_t\|_\infty)^2 \leq 2(|\tilde{u}(0)| + LR_1)^2 + 8L^2\|q_t\|_\infty^2 \tag{A.368}$$

And since

$$\mathbb{E}\left[(\tilde{u}(d_t)-\tilde{u}(0))^2\,|\mathcal{F}_t\right] \le \mathbb{E}\left[L\cdot d_t^2\right] \tag{A.369}$$

$$= \mathbb{E}\left[L\cdot\left(r_t+\gamma\cdot\max_a q_t(s',a)-q_t(s,a)\right)^2\right] \tag{A.370}$$

$$= \mathbb{E}\left[L\cdot\left(r_t^2+2r_t\cdot(\gamma\cdot\max_a q_t(s',a)-q_t(s,a))+\right.\right. \tag{A.371}$$

$$\left.\left.(\gamma\cdot\max_a q_t(s',a)-q_t(s,a))^2\right)\right] \tag{A.372}$$

$$= LR_2+2LR_1(1-\gamma)\cdot\|q_t\|_\infty+L(1-\gamma)^2\cdot\|q_t\|_\infty^2 \tag{A.373}$$

where R_2 is the upper bound for $\mathbb{E}[r_t^2]$ due to Assumption 3 ($\mathbb{E}[r_t^2]\le 4\sigma^2\cdot\Gamma(1)$ [14]). Note that here $\tilde{u}(0)=0$, hence we have

$$\alpha^2\,\mathbb{E}[(\tilde{u}(d_t))^2|\mathcal{F}_t] \le \alpha^2\cdot\left(LR_2+2LR_1(1-\gamma)\cdot\|q_t\|_\infty+L(1-\gamma)^2\cdot\|q_t\|_\infty^2\right) \tag{A.374}$$

Hence,

$$\mathbb{E}[w_t^2(s,a)|\mathcal{F}_t] \le 2\alpha^2\cdot\left(LR_2+2LR_1(1-\gamma)\cdot\|q_t\|_\infty+L(1-\gamma)^2\cdot\|q_t\|_\infty^2\right) \tag{A.375}$$

if $\|q_t\|_\infty\le 1$, then

$$\mathbb{E}[w_t^2(s,a)|\mathcal{F}_t] \le 2\alpha^2\cdot\left(LR_2+2LR_1(1-\gamma)+L(1-\gamma)^2\cdot\|q_t\|_\infty^2\right) \tag{A.376}$$

if $\|q_t\|_\infty > 1$, then

$$\mathbb{E}[w_t^2(s,a)|\mathcal{F}_t] \le 2\alpha^2\cdot\left(LR_2+(2LR_1(1-\gamma)+L(1-\gamma)^2)\cdot\|q_t\|_\infty^2\right) \tag{A.377}$$

Then it's shown that q_t satisfy all of the requirements in Proposition 1, then $q_t\to q^*$ with probability 1.

A.4.1 Proof of Theorem 6

Poisson masks $M\sim Poisson(1)$ provides parallel learning since $Binomial(T,\frac{1}{T})\to Poisson(1)$ as $T\to\infty$, so each Q table Q^i is trained in parallel. The proof of convergence of Q^i for all $i\in\{1,...,k\}$ is exactly same as Section A.4. Hence $\frac{1}{k}\sum_{i=1}^k Q^i\to Q^*$ w.p. 1.

A.4.2 Discussion of RA3-Q

In this section, I'll discuss convergence issues on RA3-Q. First I'll discuss a simplified setting where I show that if the adversary's policy is a *fixed* policy π_0^A, the update rule for protagonist Eq. (7.12) converges to the optimal of $J^P(s, :, \pi_0^A)$. Similarly, if the protagonist's policy is a *fixed* policy π_0^P, the update rule for adversary Eq. (7.13) converges to the optimal of $J^A(s, \pi_0^P, :)$.

Poisson masks $M \sim Poisson(1)$ provides parallel learning since $Binomial(T, \frac{1}{T}) \to Poisson(1)$ as $T \to \infty$, so each Q table of protagonist/adversary, Q_P^i, Q_A^i, are trained in parallel respectively.

Similar to Section A.4, I need to prove the convergence of the iterative procedure. Take agent protagonist as an example, and the proof for adversary is similar.

Fix the policy for adversary, then according to [[49] **Proposition 3.1**], for any random variable X, the following statements are equivalent

$$\text{(i) } \frac{1}{\beta^P} \log \mathop{\mathbb{E}}_{\mu} \left[exp \left(\beta^P \cdot X \right) \right] = m^*$$

$$\text{(ii) } \mathop{\mathbb{E}}_{\mu} \left[u^P(X - m^*) \right] = x_0$$

I'll use this proposition in the following context to show that the convergent point is the optimal of the objective function $\tilde{J}^P(s, :, \pi_0^A)$.

Compared to Algorithm 2 (RAQL), RA3-Q uses multi-agent extension of MDP (where the transition function is $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \mathbb{R}^{\mathcal{S}}$. We reformulate the update rule Eq. (7.12) by letting

$$q_{t+1}^P(s, a_P, a_A) = \left(1 - \frac{\alpha_t(s, a_P, a_A)}{\alpha} \right) q_t^P(s, a_P, a_A) + \tag{A.378}$$

$$\frac{\alpha_t(s, a_P, a_A)}{\alpha} \cdot \left[\alpha \cdot u(d_t) - x_0 + q_t^P(s, a_P, a_A) \right] \tag{A.379}$$

$$\text{where } d_t := r_t^P + \gamma \cdot \max_{a_P, a_A} q_t^P(s', a_P, a_A) - q_t^P(s, a_P, a_A) \quad x_0 = -1 \quad \alpha \in (0, \min(L^{-1}, 1)] \tag{A.380}$$

And we set

$$(H^P q_t^P)(s, a_P, a_A) = \alpha \cdot \mathop{\mathbb{E}}_{s,a_P,a_A} \left[\tilde{u} \left(r_t^P + \gamma \cdot \max_{a_P,a_A} q_t^P(s', a_P, a_A) - q_t^P(s, a_P, a_A) \right) \right] + \tag{A.381}$$

$$q_t^P(s, a_P, a_A) \tag{A.382}$$

$$w_t(s, a_P, a_A) = \alpha \cdot \tilde{u}(d_t) - \alpha \cdot \mathop{\mathbb{E}}_{s,a_P,a_A} \left[\tilde{u} \left(r_t^P + \gamma \cdot \max_{a_P,a_A} q_t^P(s', a_P, a_A) - q_t^P(s, a_P, a_A) \right) \right] \tag{A.383}$$

$$\tilde{u}(x) = u(x) - x_0 \tag{A.384}$$

Next, show that H^P is a $(1 - \alpha(1 - \gamma)\epsilon)$-contractor under Assumption 2:

For any two q tables q, q', define $v^P(s) := \max_{a_P,a_A} q(s, a_P, a_A)$ and $v^{P'}(s) := \max_{a_P,a_A} q'(s, a_P, a_A)$. Thus,

$$|v^P(s) - v^{P'}(s)| \leq \max_{s,a_P,a_A} |q(s, a_P, a_A) - q'(s, a_P, a_A)| = \|q - q'\|_\infty \tag{A.385}$$

By Assumption 2 and monotonicity of $\tilde{u}$, for given $x, y \in \mathbb{R}$, there exists $\xi_{(x,y)} \in [\epsilon, L]$ such that

$$\tilde{u}(x) - \tilde{u}(y) = \xi_{(x,y)} \cdot (x - y).$$

Then it can be obtained that

$$(H^P q)(s, a_P, a_A) - (H^P q')(s, a_P, a_A) \tag{A.386}$$

$$= \sum_{s'} \mathcal{P}[s'|s, a_P, a_A] \cdot \Big\{ \alpha \xi_{(s,a_P,a_A,s',q,q')} \cdot [\gamma \cdot v^P(s') - \gamma \cdot v^{P'}(s') - \tag{A.387}$$

$$q(s, a_P, a_A) + q'(s, a_P, a_A)] + (q(s, a_P, a_A) - q'(s, a_P, a_A)) \Big\} \tag{A.388}$$

$$\leq \left(1 - \alpha(1 - \gamma) \sum_{s'} \mathcal{P}[s'|s, a_P, a_A] \cdot \xi_{(s,a_P,a_A,s',q,q')} \right) \|q - q'\|_\infty \tag{A.389}$$

$$\leq (1 - \alpha(1 - \gamma)\epsilon) \|q - q'\|_\infty \tag{A.390}$$

Hence H^P is a contractor.

By Eq. (A.383), $\mathbb{E}[w_t(s, a_P, a_A)|\mathcal{F}_t] = 0$. Hence it remains to prove b(ii) in Proposition 1.

$$\mathbb{E}\left[w_t^2(s, a_P, a_A)|\mathcal{F}_t\right] = \alpha^2 \cdot \mathbb{E}\left[(\tilde{u}(d_t))^2|\mathcal{F}_t\right] - \alpha^2 (\mathbb{E}[\tilde{u}(d_t)|\mathcal{F}_t])^2 \leq \alpha^2 \cdot \mathbb{E}\left[(\tilde{u}(d_t))^2|\mathcal{F}_t\right] \tag{A.391}$$

Following from the same procedures as Section A.4, condition b(ii) of Proposition 1 also holds in this case. And recall that the learning rate satisfies condition a, hence by Proposition 1, $q \to q^*$, where q^* is the solution to the Bellman equation

$$\mathop{\mathbb{E}}_{s,a_P,a_A}\left[u^P\left(r_t^P + \gamma \cdot \max_{a_P,a_A} q(s', a_P, a_A) - q(s, a_P, a_A)\right)\right] = x_0 \qquad \pi_0^A \text{ is fixed} \tag{A.392}$$

for $\forall(s, a_P, a_A)$. Where s' is sampled from $\mathcal{P}[\cdot|s, a_P, a_A]$. Similarly, it can be shown that for a fixed policy for protagonist, the update rule Eq. (7.13) will guarantee that $q_A \to q_A^*$, where q_A^* is the solution to the Bellman equation

$$\mathop{\mathbb{E}}_{s,a_P,a_A}\left[u^A\left(r_t^A + \gamma \cdot \max_{a_P,a_A} q(s', a_P, a_A) - q(s, a_P, a_A)\right)\right] = x_1 \qquad \pi_0^P \text{ is fixed} \tag{A.393}$$

for $\forall(s, a_P, a_A)$. Where s' is sampled from $\mathcal{P}[\cdot|s, a_P, a_A]$.

Note that this does not imply a convergence guarantee of RA3-Q because of the *protagonist/adversary's policy is fixed* assumption. Only if one of the agents (say protagonist) stops learning (and its policy becomes fixed) at some point, then the other agent (adversary) will also converge. Note that in the general multi-agent learning case this is always a challenge and it is often hard to a balance between theoretical algorithms (with convergence guarantees) and practical algorithms (loosing guarantees but with good empirical results), see the experimental results in Section 9.1 and related literature [12], [33], [58].

A.5 Proof of Theorem 7

Theorem 7 For a Normal Form Game with p players, and each player i chooses a strategy π^i from a set of strategies $S^i = \{\pi_1^i, ..., \pi_k^i\}$ and receives a risk averse payoff $h^i(\pi^1, ..., \pi^p) : S^1 \times ... \times S^p \to \mathbb{R}$ satisfying Assumption 4. If $\mathbf{x}$ is a Nash Equilibrium for the game $\hat{h}^i(\pi^1, ..., \pi^p)$, then it is a 2ϵ-Nash equilibrium for the game $h^i(\pi^1, ..., \pi^p)$ with probability $1 - \delta$ if we play the game for n times, where

$$n \geq \max\left\{\frac{8R^2}{\epsilon^2}\log\frac{|S^1| \times ... \times |S^p| \times p}{\delta}\ ;\ \frac{128R^4\beta^2}{\epsilon^2 n}\log\frac{|S^1| \times ... \times |S^p| \times p}{\delta}\right\} \tag{A.394}$$

Assumption 4. *The stochastic return h (for each player and each strategy) for each simulation has a sub-Gaussian tail. i,e, there exists $\omega > 0$ s.t.*

$$\mathbb{E}\left[exp\left(c \cdot (h - \mathbb{E}[h])\right)\right] \leq exp\left(\frac{\omega^2 c^2}{2}\right) \qquad \forall c \in \mathbb{R} \tag{A.395}$$

And we also select $R > 0$ s.t. $h \in [-R, R]$ almost surely.

Proof. Note that we have the following relation:

$$\mathbb{E}_{\pi\sim\mathbf{x}}\left[h^i(\pi)\right] = \mathbb{E}_{\pi\sim\mathbf{x}}\left[\hat{h}^i(\pi)\right] + \mathbb{E}_{\pi\sim\mathbf{x}}\left[h^i(\pi) - \hat{h}^i(\pi)\right] \tag{A.396}$$

Then

$$\mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i})\right] = \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[\hat{h}^i(\pi^i, \pi^{-i})\right] + \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i}) - \hat{h}^i(\pi^i, \pi^{-i})\right] \tag{A.397}$$

$$\max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i})\right] \leq \max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[\hat{h}^i(\pi^i, \pi^{-i})\right] + \tag{A.398}$$

$$\max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i}) - \hat{h}^i(\pi^i, \pi^{-i})\right] \tag{A.399}$$

Hence,

$$\max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i})\right] - \mathbb{E}_{\pi\sim\mathbf{x}}\left[h^i(\pi)\right] \tag{A.400}$$

$$\leq \underbrace{\max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[\hat{h}^i(\pi^i, \pi^{-i})\right] - \mathbb{E}_{\pi\sim\mathbf{x}}\left[\hat{h}^i(\pi)\right]}_{=0 \text{ since } \mathbf{x} \text{ is a Nash Equilibrium for } \hat{h}^i} + \underbrace{\max_{\pi^i} \mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i, \pi^{-i}) - \hat{h}^i(\pi^i, \pi^{-i})\right]}_{\leq\epsilon} + \tag{A.401}$$

$$\underbrace{\mathbb{E}_{\pi\sim\mathbf{x}}\left[\hat{h}^i(\pi) - h^i(\pi)\right]}_{\leq\epsilon} \tag{A.402}$$

Hence, if we can control the difference between $|h^i(\pi) - \hat{h}^i(\pi)|$ uniformly over players and actions, then an equilibrium for the empirical game is almost an equilibrium for the game defined by the reward function. Hence the question is how many samples n do we need to assess that a Nash equilibrium for $\hat{h}$ is a 2ϵ-Nash equilibrium for h for a fixed confidence δ and a fixed ϵ.

In the following, in short, we fix player i and the joint strategy $\pi = (\pi^1, ..., \pi^p)$ for p players and and in short, denote $h^i = h^i(\pi)$, $\hat{h}^i = \hat{h}^i(\pi)$. By Hoeffding inequality,

$$\mathbb{P}\left[\left|\bar{R}^i - \mathbb{E}[R^i]\right| \geq \frac{\epsilon}{2}\right] \leq 2 \cdot exp\left(-\frac{\epsilon^2 n}{8R^2}\right) \tag{A.403}$$

Now, it remains to give a batch scenario for the unbiased estimator of variance penalty term. Denote $V_n^2 = \frac{1}{n-1}\sum_{j=1}^n \left(R_j^i - \bar{R}^i\right)^2$, then $\mathbb{E}[V_n^2] = \mathbb{V}ar[R^i] = \sigma^2$, i.e., it's an unbiased estimator of the game variance.

By McDiarmid's inequality [5],

$$\mathbb{P}\left[\left|V_n^2 - \mathbb{V}ar[R^i]\right| \geq \frac{\epsilon}{2\beta}\right] \leq 2 \cdot exp\left(-n\frac{(\epsilon/2\beta)^2}{32R^4}\right) = 2 \cdot exp\left(-\frac{\epsilon^2 n}{128R^4\beta^2}\right) \tag{A.404}$$

By triangle inequality,

$$\begin{aligned}
\mathbb{P}\left[\left|h^i - \hat{h}^i\right| \geq \epsilon\right] &\leq \mathbb{P}\left[\left|\mathbb{E}[R^i] - \bar{R}^i\right| + \beta \cdot \left|V_n^2 - \mathbb{V}ar[R^i]\right| \geq \epsilon\right] && \text{(A.405)}\\
&\leq \mathbb{P}\left[\left|\mathbb{E}[R^i] - \bar{R}^i\right| \geq \frac{\epsilon}{2} \; or \; \beta \cdot \left|V_n^2 - \mathbb{V}ar[R^i]\right| \geq \frac{\epsilon}{2}\right] && \text{(A.406)}\\
&\leq \mathbb{P}\left[\left|\mathbb{E}[R^i] - \bar{R}^i\right| \geq \frac{\epsilon}{2}\right] + \mathbb{P}\left[\left|V_n^2 - \mathbb{V}ar[R^i]\right| \geq \frac{\epsilon}{2\beta}\right] && \text{(A.407)}\\
&\leq 2 \cdot exp\left(-\frac{\epsilon^2 n}{8R^2}\right) + 2 \cdot exp\left(-\frac{\epsilon^2 n}{128R^4\beta^2}\right) && \text{(A.408)}\\
&= f(n, \epsilon). && \text{(A.409)}
\end{aligned}$$

Hence, for per joint strategies π and per player i, we have the following bound :

$$\begin{aligned}
\mathbb{P}\left[\sup_{\pi,i}\left|h^i(\pi) - \hat{h}^i(\pi)\right| \geq \epsilon\right] &\leq \sum_{\pi,i}\mathbb{P}\left[\left|h^i - \hat{h}^i\right| \geq \epsilon\right] \quad \text{By union bound} && \text{(A.410)}\\
&\leq |S^1| \times ... \times |S^p| \times p \times f(n, \epsilon) && \text{(A.411)}
\end{aligned}$$

Hence for

$$f(n, \epsilon) \leq \frac{\delta}{|S^1| \times ... \times |S^p| \times p} \tag{A.412}$$

we have $\mathbb{P}\left[\sup_{\pi,i}\left|h^i(\pi)-\hat{h}^i(\pi)\right|\geq\epsilon\right]\leq\delta$ Hence, for

$$n\geq\max\left\{\frac{8R^2}{\epsilon^2}\log\frac{|S^1|\times...\times|S^p|\times p}{\delta}\ ;\ \frac{128R^4\beta^2}{\epsilon^2 n}\log\frac{|S^1|\times...\times|S^p|\times p}{\delta}\right\} \tag{A.413}$$

we have $\mathbb{P}\left[\sup_{\pi,i}\left|h^i(\pi)-\hat{h}^i(\pi)\right|<\epsilon\right]\geq 1-\delta$.

Plugging the result into Eq. (A.400), we have

$$\max_{\pi^i}\mathbb{E}_{\pi^{-i}\sim\mathbf{x}^{-i}}\left[h^i(\pi^i,\pi^{-i})\right]-\mathbb{E}_{\pi\sim\mathbf{x}}\left[h^i(\pi)\right]\leq 2\epsilon \tag{A.414}$$

□