On the Effect of Log-Barrier Regularization in Decentralized Softmax Gradient Play in Multiagent Systems

Runyu (Cathy) Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, Na Li

Abstract

Softmax policy gradient is a popular algorithm for policy optimization in single-agent reinforcement learning, particularly since projection is not needed for each gradient update. However, in multi-agent systems, the lack of central coordination introduces significant additional difficulties in the convergence analysis. Even for a stochastic game with identical interest, there can be multiple Nash Equilibria (NEs), which disables proof techniques that rely on the existence of a unique global optimum. Moreover, the softmax parameterization introduces non-NE policies with zero gradient, making NE-seeking difficult for gradient-based algorithms. In this paper, we study the finite time convergence of decentralized softmax gradient play in a special form of game, Markov Potential Games (MPGs), which includes the identical interest game as a special case. We investigate both gradient play and natural gradient play, with and without log-barrier regularization. Establishing convergence for the unregularized cases relies on an assumption that the stationary policies are isolated, and yields convergence bounds that contain a trajectory dependent constant that can be arbitrarily large. We introduce the log-barrier regularization to overcome these drawbacks, with the cost of slightly worse dependence on other factors such as the action set size. An empirical study on an identical interest matrix game confirms the theoretical findings.

1 Introduction

Multi-agent systems encounter vast application in real world scenarios, such as network routing (Tao et al., 2001; Claes et al., 2011), social and economic decision making (Ventre et al., 2013; Roscia et al., 2013), and robotic swarms (Liu & Wu, 2018; Iñigo-Blasco et al., 2012). In these problems, a system consists of a group of agents interacting in a shared environment. Given the recent success of reinforcement learning (RL), increasing attention has been drawn to the possibility of applying RL algorithms, such as policy gradient, to multi-agent systems. However, the theoretical foundations for multi-agent reinforcement learning (MARL) remain limited. Unlike single-agent RL, the actions of other agents affect the decision making outcome for each individual in the system, raising additional theoretical challenges when analyzing joint performance.

The stochastic game (SG) is a classical multi-agent model that has received extensive attention in recent MARL studies. In a stochastic game, the environment is represented by a state space that evolves based on the joint actions of agents. Each agent in a stochastic game tries to maximize its own total reward by making decisions independently, based on state information shared between agents. The stochastic game model was first introduced in Shapley (1953), with a series of followup works proposing NE-seeking algorithms, particularly in the RL setting (e.g. Littman, 1994; Bowling & Veloso, 2000; Shoham et al., 2003; Bucsoniu et al., 2010; Lanctot et al., 2017; Zhang et al., 2019 and citations therein). Given recent progress in the underlying theory of RL, many recent works have investigated finite time iteration and sample complexity for learning NE or other general equilibria notions, such as correlated and coarse correlated equilibria (e.g. Song et al., 2021).

There are different types of SGs, some with attributes that merit special attention; for example, two-player zero sum games (Bai & Jin, 2020; Daskalakis et al., 2021), which are widely used to model two player competitive games such as GO. In this paper, we will focus on another type of SG, the Markov potential game (MPG) (Macua et al., 2018; Mguni et al., 2021; Zhang et al., 2021; Leonards et al., 2021), which includes the identical interest game as a special case. The structure of a MPG enables efficient learning through the use of gradient-based algorithms such as gradient play. However, recent work (Zhang et al., 2021; Leonards et al., 2021) has focused on the iteration and sample complexity of finding a NE in an MPG under the direct policy parameterization, which is not practical in most real world scenarios, given the cost of projecting back to the probability simplex on every iteration. This drawback has motivated consideration of the softmax parameterization, which bypasses the projection step in the gradient update, and is perhaps the most popular approach to parameterizing policies in practice. Fox et al. (2021) have studied natural gradient play for MPG...
under softmax parameterization, but only address asymptotic behavior and leave finite time complexity open, while additionally requiring an assumption that the set of stationary policies is isolated (which is the same as Assumption 3.2 in this paper below).

From the perspective of analysis and practical performance, the extension from the direct to the softmax parameterization in policies is nontrivial. Even in the single agent case, as shown by Agarwal et al. (2020); Mei et al. (2020), there are policies in the softmax parameterization that have near-zero gradient and yet are far from being globally optimal, which creates difficulty for a gradient-based algorithm to escape suboptimal points. A similar issue exists for MPGs: due to the more complex interaction between agents, there is even a greater set of policies that obtain small gradient norm but are far from a NE. Based on our analysis and numerical results, even for natural gradient play—which is known to enjoy dimension free convergence in single agent learning (Agarwal et al., 2020)—we find in the multiagent setting that it can still become stuck in these undesirable regions. Such evidence suggests that preconditioning according to the Fisher information matrix is not sufficient to ensure fast convergence in multi-agent learning. A stronger form of regularization is required, which motivates the introduction of log-barrier regularization to avoid undesirable regions of policy space.

| Algorithm                                             | Single-agent MDP                                                                 | Multi-agent MPG                                                                 |
|-------------------------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Gradient play, direct parameterization                | \( O\left(\frac{|A|^{M^4}}{\epsilon^{2-M}}\right) \) (Agarwal et al., 2020)   | \( O\left(\frac{\text{max}_{i}(\epsilon\to\text{min})^2}{M}M^2\right) \) (Zhang et al., 2021; Leonardos et al., 2021) |
| Gradient play, softmax parameterization              | \( O\left(\frac{|A|^{M^4}}{\epsilon^{2-M}}\right) \) (Mei et al., 2020)       | \( O\left(\frac{\text{max}_{i}(\epsilon\to\text{min})^2}{M}M^2\right) \) * |
| Natural gradient play, softmax parameterization      | \( O\left(\frac{|A|^{M^4}}{\epsilon^{2-M}}\right) \) (Agarwal et al., 2020)   | \( O\left(\frac{\text{max}_{i}(\epsilon\to\text{min})^2}{M}M^2\right) \) * |
| Gradient play+log-barrier regularization,             | \( O\left(\frac{|A|^{M^4}}{\epsilon^{2-M}}\right) \) (Agarwal et al., 2020)   | \( O\left(\frac{\text{max}_{i}(\epsilon\to\text{min})^2}{M}M^2\right) \) * |
| softmax parameterization                              |                                                                                 |                                                                                 |
| Natural gradient play+log-barrier regularization      | Unknown                                                                        | \( O\left(\frac{\text{max}_{i}(\epsilon\to\text{min})^2}{M}M^2\right) \) |
| softmax parameterization                              |                                                                                 |                                                                                 |

Table 1: Summary of known convergence rate results for gradient based methods in Markov decision processes (MDPs) and MPGs respectively. The new results proved in this paper for MPGs are displayed in bold font. Complexity bounds with * depend on an additional assumption on the MPG (see Assumption 3.2). The definitions of variables \( M \) and \( c \) appearing in some bounds can be found in [1] and [10]. Note that the definition of \( M \) is slightly different from the “distribution mismatch coefficient” \( D_\infty \) defined in [Agarwal et al., 2020] (see more details in descriptions that follows Assumption 2.3). To make the complexity results more comparable, we slightly modify and re-derive the results in Agarwal et al. (2020); Mei et al. (2020); Zhang et al. (2021); Leonardos et al. (2021), which use \( D_\infty \) to measure the complexity, to using \( M \) instead while preserving the other major steps in the proofs of these results.

Our contribution: In this paper, we provide finite time iteration complexity results for gradient and natural gradient play under the softmax parameterization, considering both unregularized and log-barrier regularized dynamics. To the best of our knowledge, these are the first such results for MPGs under the direct parameterization. We summarize our new results and compare them to existing results for the direct parameterization and to the corresponding single agent cases in Table 1. These findings suggest that regularization is crucial for obtaining fast convergence to a NE under the softmax parameterization in a MPG. In Table 1 the results for the two unregularized algorithms in the multi-agent case rely on the assumption that the set of stationary policies is isolated (Assumption 3.2), and the corresponding complexity bounds contain an initialization dependent factor \( c \). By contrast, the log-barrier regularized algorithms overcome both drawbacks, but as a tradeoff, their bounds incur a slightly worse dependence on \(|A_i| \) and \( M \). We observe numerically that the log-barrier regularized algorithms are indeed more robust against becoming trapped near undesirable non-NE stationary points. Our results also convey the following two additional messages. First, finding the NE of a multi-agent MPG is harder than finding the global optimum of a Markov decision process (MDP), because multi-agent learning suffers greater risk of becoming trapped near undesirable stationary points. Our complexity bounds also indicate that, for MPGs, gradient-based algorithms exhibit complexities that scale with \( O\left(\frac{1}{\epsilon}\right) \), which is worse than \( O\left(\frac{1}{\epsilon^{1.5}}\right) \) (Agarwal et al., 2020) or even the exponential rate (Khodadadian et al., 2021; Mei et al., 2021; Cen et al., 2021) obtainable in the single agent case. Second, natural gradient play outperforms gradient play counterparts, suggesting that natural gradient play captures useful information
about the geometry of the parameter space that helps accelerate the learning process.

## 2 Problem Settings

We consider an infinite time horizon $n$-agent stochastic game (SG, [Shapley, 1953]) $\mathcal{M} = (N, S, A = A_1 \times \cdots \times A_n, P, r = (r_1, \ldots, r_n), \gamma, \rho)$ which is specified by an agent set $N = \{1, 2, \ldots, n\}$, a finite state space $S$, a finite action space $A_i$ for each agent $i \in N$, a transition model $P$ (such that $P(s'|s, a) = P(s'|s, a_1, \ldots, a_n)$ is the probability of transitioning into state $s'$ upon taking action $a := (a_1, \ldots, a_n)$ in state $s$ where $a_i \in A_i$ is action of agent $i$), a reward function $r_i : S \times A \to [0, 1]$ for each agent $i$, a discount factor $\gamma \in [0, 1)$, and an initial state distribution $\rho$ over $S$. We use $s(t) \in S$ to denote the state at time step $t$, and $a(t) = (a_1(t), \ldots, a_n(t)) \in A$ to denote the total action at time step $t$.

A stochastic policy $\pi : S \to \Delta(A)$ (where $\Delta(A)$ is the probability simplex over $A$) specifies a strategy, where agents choose their actions jointly based on the current state in a stochastic fashion; i.e. $Pr(a(t)|s(t)) = \pi(a(t)|s(t))$. A decentralized stochastic policy is a special subclass of stochastic policies with $\pi = \pi_1 \times \cdots \times \pi_n$, such that $\pi_i : S \to \Delta(A_i)$, where $\pi_i$ is agent $i$’s own local policy. For decentralized stochastic policies, each agent takes its action based on the current state $s$ independently of other agents’ action choices; i.e.,

$$Pr(a(t)|s(t)) = \pi(a(t)|s(t)) = \prod_{i=1}^{n} \pi_i(a_i(t)|s(t)).$$

For notation simplicity, we define $\pi_i(a_i|s):=\prod_{j \in I} \pi_i(a_i|s)$, where $I \subseteq N$ is an index set. Further, we use the notation $\neg i$ to denote the index set $N \setminus \{i\}$.

In this paper we focus on tabular softmax parameterization for a policy, where policy $\pi_\theta = (\pi_{\theta_1}, \ldots, \pi_{\theta_n})$ is parameterized by a set of parameters $\theta = (\theta_1, \ldots, \theta_n)$, with $\theta_i = \{\theta_{s,a_i}\}_{s \in S, a_i \in A_i}$, and where

$$\pi_{\theta_i}(a_i|s) = \frac{\exp(\theta_{s,a_i})}{\sum_{a_i'} \exp(\theta_{s,a_i'})}.$$  \hspace{1cm} (1)

We primarily focus on a subclass of stochastic games called the potential game (MPG) defined as follows:

**Definition 2.1.** A stochastic game is called a Markov potential game (MPG, [Zazo et al., 2016; Macua et al. 2018; Zhang et al. 2021; Leonards et al. 2021; Mguni 2020]) if there exists a potential function $\phi : S \times A_1 \times \cdots \times A_n \to \mathbb{R}$ such that for any agent $i$ and any pair of policy parameters $(\theta'_i, \theta_{\neg i}), (\theta_i, \theta_{\neg i})$:

$$\mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_i(s(t), a(t)) \big| \pi = (\theta'_i, \theta_{\neg i}), s(0) = s \right] - \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_i(s(t), a(t)) \big| \pi = (\theta_i, \theta_{\neg i}), s(0) = s \right] = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s(t), a(t)) \big| \pi = (\theta'_i, \theta_{\neg i}), s(0) = s \right] - \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s(t), a(t)) \big| \pi = (\theta_i, \theta_{\neg i}), s(0) = s \right], \forall s.

The definition of Markov potential game is a generalization of the notion potential game in the one-shot setting [Monderer & Shapley 1996]. Note that identical reward game where agents share a same reward function naturally satisfies the above condition and serves as an important special case of MPG. For non-identical reward settings, [Macua et al. 2018; González-Sánchez & Hernández-Lerma 2013] found that continuous MPGs can model applications such as the great fish war [Levhari & Mirman 1980], the stochastic lake game [Dechert & O’Donnell 2006], medium access control [Macua et al. 2018] etc. For tabular MPGs, [Zhang et al. 2021; Leonards et al. 2021; Mguni 2020] also discuss necessary/sufficient conditions that implies a MPG, as well as its application and counterexamples.

We denote agent $i$’s total reward starting from initial states $s(0) \sim \rho$ as:

$$J_i(\theta) := \mathbb{E}_{s(0) \sim \rho} \left[ \sum_{t=0}^{\infty} \gamma^t r_i(s(t), a(t)) \big| \pi_\theta, s(0) = s \right].$$

Similarly, we define the total potential function $\Phi$ as:

$$\Phi(\theta) := \mathbb{E}_{s(0) \sim \rho} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s(t), a(t)) \big| \pi_\theta, s(0) = s \right].$$

In the game setting, agent $i$’s objective is to maximize its own total reward $J_i$. A Nash equilibrium (NE) is often used to characterize the equilibrium (a joint policy) where no agent has a unilateral incentive to deviate from it.
We will use $M$ to denote the following quantity
\[ M := \sup_{\theta} \max_s \frac{1}{d_\theta(s)}. \] (5)

Note that $M$ can be viewed as a measure of exploration sufficiency in the stochastic game, which is slightly different from the “distributional mismatch coefficient” introduced in \cite{agarwal2020distributional} defined by $\sup_{\theta,\theta'} \max_s \frac{d_{\theta'}(s)}{d_\theta(s)}$; however, both can be upper bounded by $\max_s \frac{1}{(1-\gamma)\rho(s)}$.

**Assumption 2.4.** The stage potential function $\phi$ is bounded, i.e., $\phi_{\min} \leq \phi(s,a) \leq \phi_{\max}$.
3 Unregularized Gradient Play

We first investigate the convergence to NE for gradient and natural gradient play respectively. Under the softmax parameterization, the two schemes are given by

\[
\text{Gradient Play: } \theta^{(t+1)} = \theta^{(t)} + \eta \nabla_{\theta} J_i(\theta^{(t)})
\]

\[
\text{Natural Gradient Play: } \theta_i^{(t+1)} = \theta_i^{(t)} + \eta F_i(\theta^{(t)}) \nabla_{\theta_i} J_i(\theta_i^{(t)})
\]

where $\dagger$ denotes the Moore-Penrose inverse and $F_i(\theta)$ is the Fisher information matrix for $\pi_{\theta_i}$:

\[
F_i(\theta) := \mathbb{E}_{\pi_{\theta_i}(\cdot)} \mathbb{E}_{a \sim \pi_{\theta_i}(\cdot | s)} \left[ \nabla \log \pi_{\theta_i}(a | s) \nabla \log \pi_{\theta}(a | s) \right].
\]

For notational simplicity, we abbreviate the variables $d_i^{(t)}$, $A_i^{(t)}$ and $\overline{A}_i^{(t)}$ as $d(t)$, $A^i(t)$ and $\overline{A}^i(t)$ respectively; and denote $\pi_{\theta_i}(a | s)$ and $\pi_{\theta_i}(a | s) = \pi^2(a | s)$ and $\pi_{\theta_i}(a | s)$ respectively.

For the softmax parameterization, we can establish the equivalence of natural gradient play and soft Q-learning (Haarnoja et al., 2017), formally stated in the following lemma.

**Lemma 3.1.** (Proof given in Appendix C) Natural gradient play is equivalent to

\[
\pi_i^{(t+1)}(a_i | s) \propto \pi_i^{(t)}(a_i | s) \exp \left( \frac{\eta \overline{A}_i^{(t)}(s, a_i)}{1 - \gamma} \right)
\]

**3.1 Asymptotic Convergence to Nash Equilibrium**

We first establish asymptotic convergence to NE for gradient and natural gradient play. Compared to the direct parameterization (see, e.g. Zhang et al. (2021) Leonardos et al. (2021)), the convergence analysis for the softmax parameterization is more challenging due to the existence of non-NE stationary policies; for example, any deterministic policy has zero gradient whether or not it is a NE. Meanwhile, compared to global convergence in the single agent setting Agarwal et al. (2020), the coupling of the agent dynamics in the multi-agent setting introduces extra difficulty, compelling us to introduce the same assumption as Fox et al. (2021) to establish asymptotic NE convergence.

**Assumption 3.2.** The set of stationary policies

\[
\mathcal{SP} = \{ \pi : \pi_i(a_i | s) \overline{A}^i(s, a_i) = 0, \forall s \in \mathcal{S}, a_i \in \mathcal{A}, i = 1, 2, \ldots, n \}
\]

is isolated; that is, for any $\pi \in \mathcal{SP}$, there exists a neighborhood $U_\delta(\pi) := \{ \pi' : ||\pi' - \pi||_2 \leq \delta, \pi' \neq \pi \}$ such that any $\pi' \in U_\delta(\pi)$ is not in $\mathcal{SP}$.

Note that Assumption 3.2 is imposed on the policy $\pi$ instead of the parameters $\theta$. Then, using similar techniques as Agarwal et al. (2020), we prove the following the asymptotic convergence result.

**Theorem 3.3.** (Proof given in Appendix D.1) Under Assumption 2.3, 2.4 and 3.2 for $\eta \leq \frac{(1 - \gamma)^2}{6n}$, running gradient play guarantees that $\lim_{t \to +\infty} \theta^{(t)} = \theta^{(\infty)}$, where $\theta^{(\infty)}$ is a NE. The same argument also holds for natural gradient play with $\eta \leq \frac{1}{2 \gamma (\theta_{\max} - \theta_{\min})^2}$.

We remark that Assumption 3.2 is introduced solely for the purpose of proving asymptotic convergence. The proof of Theorem 3.3 resembles the technique used in Agarwal et al. (2020) for the single agent case, which relies heavily on the fact that the sequence of Q-functions $Q^{(t)}(s, a)$ obtains a limit $Q^{(\infty)}(s, a)$. The existence of such a limit in the single agent case follows from the monotonicity of the Q-functions. However, generalizing this proof to the multi-agent case requires the assumption that the sequence of averaged Q-functions $\overline{Q}^{(t)}(s, a_i)$ (which can be non-monotonic, see, e.g. Figure 3 in Appendix) has a limit $\overline{Q}^{(\infty)}(s, a_i)$, which is not necessarily true in general. In particular, if the set of stationary points $\mathcal{SP}$ is not isolated, one cannot rule out the possibility that (natural) gradient play will not converge to a fixed point $\pi^{(\infty)}$ (see e.g. Absil et al. (2005) for counterexamples). Consequently, $\overline{Q}^{(t)}(s, a_i)$ might not converge to a single value. For the above reasons, we introduce Assumption 3.2 to ensure that $\pi^{(t)}$ converges to a fixed stationary policy $\pi^{(\infty)}$ and thus $\overline{Q}^{(t)}(s, a_i)$ obtains a limit. We believe that Assumption 3.2 is a conservative condition that is sufficient to imply asymptotic convergence. It remains an interesting open question to establish convergence without this assumption.
3.2 Convergence Rate

We next consider a convergence rate analysis for gradient play and natural gradient play. Corresponding results for the single-agent setting can be found in Mei et al. (2020) (for gradient play) and Agarwal et al. (2020); Khodadadian et al. (2021); Mei et al. (2021) (for natural gradient play). Some aspects of these analyses can be carried over to the multi-agent MPG setting; however, as will be discussed later, there are several fundamental differences that make the multi-agent case more challenging.

We begin by establishing a non-uniform Lojasiewicz condition (also known as gradient domination) for stochastic games.

**Lemma 3.4.** (Non-uniform Lojasiewicz inequality; proof given in Appendix D.2) Define

\[
M(\theta) := \max_s \frac{1}{d_\theta(s)},
\]

\[
c(\theta) := \min_s \sum_{a_i \in \text{argmax}_s Q_\theta(s,a_i)} \pi_{\theta}(a_i^*|s).
\]

Then we have that

\[
\text{NE-gap}_{i}(\theta) \leq \sqrt{|A_i| M(\theta) c(\theta) \|\nabla_\theta J_i(\theta)\|_2}.
\]

The Lojasiewicz condition (gradient domination) implies that the NE-gap of a policy can be bounded by the norm of its gradient, whereas the term ‘non-uniform’ refers to the factor \(\sqrt{|A_i| M(\theta) c(\theta)}\), which cannot be bounded uniformly for all \(\theta\). The counterpart of Lemma 3.4 for a single-agent MDP was first introduced in Mei et al. (2020), where they derive a similar relationship between the global optimality gap and the gradient norm. They also point out that the non-uniform term can be arbitrarily large for certain \(\theta\) (e.g. any \(\theta\) that corresponds to a deterministic policy), which implies that a \(\theta\) with gradient norm close to zero is not necessarily near a global optimum. Similarly, for multiagent MPG, a \(\theta\) with gradient norm close to zero is not necessarily near a NE (see Figure 1(a)-(c)).

To establish a valid convergence rate, Mei et al. (2020) leverage asymptotic convergence to show that the non-uniform term can be upper-bounded by a constant that depends on initialization. For MPG, we can apply a similar technique due to the following lemma.

**Lemma 3.5.** For any \(\theta^*\) that is a NE, \(c(\theta^*) = 1\).

Combining Theorem 3.3 and Lemma 3.5, we know that \(c(\theta(t))\) asymptotically converges to 1 for (natural) gradient play, and since \(c(\theta(t)) > 0\) for any softmax parameterized policy (because \(\pi_{\theta_i}(a_i|s) > 0\)), this gives

\[
c := \inf_t c(\theta(t)) > 0.
\]

We are now ready to give formal convergence rates for gradient and natural gradient play respectively.

**Theorem 3.6.** (Gradient play; proof given in D.3) Under Assumption 2.3, 2.4 and 3.2, for \(\eta = (1 - \gamma)^{\frac{3}{2n}}\) running gradient play (6) for \(T\) steps will guarantee:

\[
\sum_{t=0}^{T-1} \text{NE-gap}(\theta(t))^2 \leq O \left( \frac{1}{\gamma^3} \right),
\]

where \(O(\cdot)\) hides constant factors, \(c = \inf_{t \leq T} c(\theta(t)) > 0\), and \(M\) is defined as in (5).

**Theorem 3.7.** (Natural gradient play; proof given in D.4) Under Assumption 2.3, 2.4 and 3.2, for \(\eta = \frac{(1 - \gamma)^2}{2n(\phi_{\text{max}} - \phi_{\text{min}})}\) running natural gradient play (8) will guarantee:

\[
\sum_{t=0}^{T-1} \text{NE-gap}(\theta(t))^2 \leq O \left( \frac{1}{\gamma^3} \right).
\]
Proof in a glance: We provide a brief proof sketch for the convergence of gradient play. The analysis consists of two parts. The first part is to apply standard smooth nonconvex optimization techniques to show that the average of squared gradient norms scales as $O(\frac{1}{t})$. The second part is to link the gradient norm with the NE-gap via the non-uniform Lojasiewicz condition. To bound the NE-gap by the gradient norm uniformly along the trajectory, we need to take the maximum of the Lojasiewicz constant along the trajectory, i.e. $\sup_{t \leq T} \frac{L_{M}(\theta(t))}{c(\theta(t))}$, which explains why $\frac{1}{c}$ and $M$ show up in the bound. Note that the bound is only valid given the prior asymptotic convergence guarantee, without which $c = \inf_{t \leq T} c(\theta(t))$ could be zero.

Discussion on $\frac{1}{c}$: The complexity results in Theorem 3.6 and 3.7 both depend on $\frac{1}{c}$. However, this term can become arbitrarily large. In fact, [Li et al. 2021] show that $c$ can be exponentially small in terms of the number of states $|S|$ for a general finite MDP, even under uniform initialization, hence convergence can be very slow. This conclusion is also confirmed by numerical evidence. As pointed out by Mei et al. (2020), even for single agent settings, policy gradient can get stuck at regions with small gradient yet far from being global optimal. Similar or even worse phenomena can be observed for multi-agent MPG, as shown in Figure 1(a)-(c): even for a single state game ($|S| = 1$) with uniform initialization, unregularized gradient based algorithms can still enter regions with a relatively large NE-gap while the gradient norm and $c(\theta)$ are close to zero. For comparison, under similar conditions in the single agent case, i.e. $|S| = 1$ with uniform initialization, Mei et al. (2020) have proved that $c$ is lower-bounded by $\frac{1}{|S|}$, while for the multi-agent case $c$ can be much smaller than $\max_{i} |A_{i}|$ (see Figure 1(b)), indicating that multi-agent learning is harder than single-agent case.

More comparison with learning for single-agent MDP: For gradient play, we have established an iteration complexity of $O \left( \sum_{i=1}^{n} \max_{|A_{i}|} (\frac{|A_{i}| (\phi_{\max}-\phi_{\min})}{(1-\gamma)^{2}}) M^{2} \right)$ to find an $\epsilon$-NE, whereas Mei et al. (2020) show a complexity of $O \left( \frac{n \max_{i} |A_{i}| (\phi_{\max}-\phi_{\min}) M^{2}}{(1-\gamma)^{2} c^{2}} \right)$ to reach an $\epsilon$-global optimum for policy gradient in a single agent MDP. The dependence on $\frac{1}{c}$ is better in the single agent case because of the existence of a global optimal policy $\pi^{*}$ and optimal total reward $J^{*}$, which justify the definition of optimality gap $\delta_{i} = J(\theta(i)) - J^{*}$. This, combined with the non-uniform Lojasiewicz condition which bounds $\delta_{i}$ by the gradient norm, allows one to use techniques from convex smooth analysis to show that $\delta_{i}$ is on the scale of $\frac{1}{c}$. By contrast, for multi-agent learning, there can be multiple NEs with different values, hence $\delta_{i}$ is ill-defined. Further, note that the NE-gap is different from the optimality gap, hence gradient ascent no longer guarantees monotonic decreasing of NE-gap (Figure 1(c)). For comparison, in (Agarwal et al. 2020) for single-agent MDPs. (A better exponential convergence rate for natural PG has also been proved in (Khodadadian et al. 2021; Mei et al. 2021) with the exponential factor being problem dependent.) Nevertheless, the dependence on $\frac{1}{c}$ and $M$ is better than with gradient play, suggesting that the preconditioning of natural gradient play at least partially captures the geometry of the parameter space. We also note that the quadratic dependence on $|A_{i}| \phi_{\max} - \phi_{\min}$ might be a proof artifact. It remains an open question whether this can be reduced to a linear dependence.

4 Gradient Play with log-Barrier Regularization

The previous section has shown that, for unregularized objectives, the convergence rate for gradient based algorithms depends on a factor $\frac{1}{c}$ that can be arbitrarily large for bad initializations. This motivates us to investigate regularization, in hopes of removing the dependence on $\frac{1}{c}$, as well as removing Assumption 3.2. For this purpose, we consider log-barrier regularization:

$$J_{\epsilon}(\theta) = J(\theta) + \lambda \sum_{s,a} \log \pi_{\theta}(a_{i}|s).$$

Define:

$$\Phi(\theta) = \Phi(\theta) + \lambda \sum_{i=1}^{n} \sum_{s,a} \log \pi_{\theta}(a_{i}|s).$$

(13)

It is not hard to verify that the gradient with respect to $J_{\epsilon}$ is:

$$\frac{\partial J_{\epsilon}(\theta)}{\partial \theta_{s,a_{i}}} = \frac{\partial \Phi(\theta)}{\partial \theta_{s,a_{i}}}.$$
\[
= \frac{1}{1 - \gamma} d_\theta(s)\pi_\theta(s,a|s)\theta(s,a) + \lambda - \lambda|A_i|\pi_\theta(s,a|s).
\]

Before analyzing the resulting algorithm we first discuss the motivation for this regularizer. First, note that for each agent, the additional regularizer only depends on an agent’s own local policy, which is desirable for multiagent RL. As an alternative, one might impose regularization by choosing

\[
\tilde{\Phi}(\theta) = \Phi(\theta) + \lambda\mathbb{E}_{s \sim d_\theta(\cdot)} \sum_{i=1}^{n} \sum_{a_i} \log \pi_{\theta_i}(a_i|s);
\]
i.e., so that the regularization weight imposed on a state \(s\) depends on the state visitation probability \(d_\theta(s)\). However, in this case the gradient of the \(i\)-th agent \(\nabla_{\theta_i} \tilde{\Phi}(\theta)\) will not only depend on its own policy parameter \(\theta_i\), but also on other parameters of the other agents’ policies \(\theta_{-i}\). Thus, running gradient based algorithms with such a regularization scheme can no longer be executed in a fully decentralized manner using local policy information. Therefore, we prefer regularization \([13]\) which does not depend on \(d_\theta(s)\). Second, we adopt the log-barrier instead of entropy regularization to penalize small \(\pi_i(a_i|s)\) values more strongly, since this is where the geometry becomes close to singular. Intuitively, the log-barrier regularized gradient field repels the trajectory from singular regions more strongly, which in turn encourages better convergence properties. Although entropy regularization achieves fast exponential convergence in single agent learning (Cen et al., 2021; Mei et al., 2020), for multi-agent learning, we haven’t been able to obtain results as strong as the log-barrier regularization, so we focus on log-barrier regularization in this paper. It remains future work to determine whether entropy regularization can achieve the same, or even better convergence rates.

### 4.1 Gradient Play

We first consider gradient play algorithm, i.e.,

\[
\theta_{i}^{(t+1)} = \theta_{i}^{(t)} + \eta \nabla_{\theta_i} \tilde{J}_i(\theta^{(t)}).
\]

We start with the following lemma:

**Lemma 4.1.** Suppose \(\theta\) is such that

\[
\|\nabla_{\theta_i} \tilde{J}_i(\theta)\|_2 \leq \lambda
\]

then

\[
\text{NE-gap}_i(\theta) \leq \lambda M|A_i|,
\]

where \(M\) is defined as in Assumption 2.3.

Lemma 4.1 implies that any policy with gradient norm smaller than \(\lambda\) is also a \(\lambda M\) \(\max_i |A_i|\)-NE. Thus by properly choosing \(\lambda\), agents can find a \(\epsilon\)-NE by running gradient play.

**Theorem 4.2.** (Proof given in Appendix E.1) Under Assumption 2.3 and 2.4 for \(\eta = \frac{(1-\gamma)^2}{6n+2\max_i |A_i|(1-\gamma)^2}\), and \(\lambda = \frac{\epsilon}{\max_i |A_i|} M\), let \(\theta^{(0)}\) be the uniform random policy, i.e., \(\theta^{(0)} = \mathbf{0}\), then running gradient play \([14]\) for \(T\) steps, where

\[
T \gtrsim O\left(\frac{n \max_i |A_i|^2 (\phi_{\max} - \phi_{\min}) M^2}{(1 - \gamma)^4 \epsilon^2}\right)
\]

will guarantee that

\[
\min_{0 \leq t \leq T-1} \text{NE-gap}(\theta^{(t)}) \leq \epsilon.
\]

Note that compared to the unregularized case in Theorem 4.2 we no longer require Assumption 3.3 (isolated NEs), while the convergence rate is accelerated by eliminating the dependence on \(\frac{1}{\gamma}\). However, as a (worthy) tradeoff, the dependence on the action space size \(\max_i |A_i|\) now becomes quadratic. The key reason for these differences is that log-barrier regularization assures that any policy with sufficiently small gradient norm cannot be close to the boundary of the probability simplex where the non-uniform Lojasiewicz constant is large. As a tradeoff, a \(\lambda \max_i |A_i| M\) bias is introduced as suggested in Lemma 4.1 i.e., log-barrier regularization can find a \(\lambda \max_i |A_i| M\)-NE at best. Thus to recover a \(\epsilon\)-NE, the regularization constant \(\lambda\) must be smaller than \(\frac{\epsilon}{\max_i |A_i| M}\), which results in the quadratic dependence on \(\max_i |A_i|\).
| $a_1$ | $a_2$ | $a_3$ |
|-------|-------|-------|
| 1     | -1    | 0.14  |
| 2     | 0.16  | 0.15  |
| 3     | 0.2   | -1    |

Reward table

Figure 1: We consider a two-player identical reward matrix game as shown in the reward table. We run gradient play and natural gradient play (with and without log-barrier regularization) with initial policies being the uniform distribution (i.e., $\pi_1 = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$, $\pi_2 = \left[\frac{1}{2}, \frac{1}{2}\right]$). The subfigures (a)-(d) show how the NE-gap $\mathcal{N}(\theta(t))$, $\|\nabla_\theta \Phi(\theta(t))\|_2$, $c(\theta(t))$ (defined in (9)) and $\Phi(\theta(t))$ change with each iteration respectively. In Figure (c), we zoom in on the log$_{10}$ factor for natural gradient play. In Figure (d), we also zoom out the trajectory for running gradient play to iteration $2 \times 10^4$. Here the step sizes were chosen to be $\eta = 5$ while the regularization weight $\lambda$ was chosen to be $\lambda = 0.003$. In consideration of numerical stability issues, we truncate the update step of natural gradient play with log-barrier regularization by a maximum absolute value of 1 for each entry. For more numerical results and corresponding analysis see Appendix A.

4.2 Natural Gradient Play

In the unregularized setting, we have seen that natural gradient play enjoys a better convergence rate than gradient play, which motivates us to consider whether a similar advantage still holds for the regularized case.

In this section we consider natural gradient play

$$\theta_i(t+1) = \theta_i(t) + \eta F_i(\theta_i(t))^\dagger \nabla_\theta \tilde{J}_i(\theta_i(t)),$$

which is equivalent to (see the proof in Appendix C)

$$\pi_i(t+1)(a_i|s) \propto \pi_i(t)(a_i|s) \exp\left(\eta \frac{1}{1-\gamma} A_i(s,a_i) + \frac{\eta \lambda |A_i|}{d^{(i)}(s)\pi_i(t)(a_i|s) - d^{(i)}(s)}\right).$$

**Theorem 4.3.** (Proof given in Appendix E.2) Under Assumptions 2.3 and 2.4, for $\eta = \min\left\{\frac{1}{15(1-\gamma)^2 + \lambda |A_i| M}, \frac{1}{4(2\max_i |A_i| M^2) (1-\gamma)^2 + \frac{2\max_i |A_i| M^2}{(1-\gamma)^2}}\right\}$, running the NPG scheme (16) will guarantee that

$$\frac{1}{T} \sum_{t=0}^{T-1} \text{NE-gap}(\theta(t)) \leq \frac{1}{\eta \lambda T} \sum_{t=0}^{T-1} \left(\Phi(\theta(t)) - \Phi(\theta(0))\right) + \lambda \max_i |A_i| M.$$

Further, by setting $\lambda = \frac{\epsilon}{2 \max_i |A_i| M}$, $\theta(0) = \mathbf{0}$, for

$$T \gtrsim O\left(\frac{n \max_i |A_i| (\phi_{\max} - \phi_{\min}) M^2}{(1-\gamma)^2 \epsilon^2}\right),$$
we have \( \sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)}) \leq \epsilon. \)

Compared with gradient play, natural gradient play manages to reduce the time complexity by a \( \max |A_i| \) factor. Further, gradient play can only guarantee the minimal \( \text{NE-gap} \) is smaller than \( \epsilon \), while natural gradient play can guarantee the average \( \text{NE-gap} \) along the trajectory smaller than \( \epsilon \). The key step in the proof is that natural gradient play can implicitly guarantee that \( \pi_i^{(t)}(a_i|s) \geq \frac{\lambda}{4(|A_i||M| + \frac{1}{\epsilon^2})} \), \( \forall t \), hence the ascent value \( \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \) can be bounded by \( \text{NE-gap}(\theta^{(t)}) \) plus a \( \lambda \max |A_i|/M \) bias term. To the best of our knowledge, this is the best time complexity bound obtained for the softmax parameterization in a MPG.

5 An Illustrative Example

To give a better illustration of the dynamics of the four gradient play algorithms, (6), (8), (14), and (16), we consider a simple example of an identical reward matrix game with only one state. The reward table as well as the performance of the four algorithms are shown in Figure 1. Comparing the log-barrier regularized algorithms to the unregularized counterparts, one can see that the regularized dynamics converge faster but with a bias induced by the regularizer. This finding corroborates the analyses given in Theorem 4.2 and 4.3. By contrast, the unregularized dynamics are able to find a policy with zero \( \text{NE-gap} \) asymptotically, but tend to get stuck in regions where \( c(\theta^{(t)}) \) is very close to zero, as illustrated in Fig 1(a)(b). Specifically unregularized natural gradient play gets stuck around iteration 100-400 in a region where the gradient norm and \( c(\theta^{(t)}) \) are both close to zero while the \( \text{NE-gap} \) is not. This corroborates the finding in Lemma 3.4. Similar behavior can be observed for gradient play if we keep running the algorithm.

In comparing the natural gradient play to gradient play algorithms, natural gradient play generally converges faster, which matches with our complexity analysis. However, natural gradient play with log-barrier regularization can suffer from numerical instability due to the \( \frac{1}{\pi_i^{(0)}(a_i|s)} \) term in the exponential factor. In this case, the stepsize needs to be chosen carefully. To bypass the numerical instability, we truncate the update step of natural gradient play with log-barrier regularization by a maximum absolute value of 1 for each entry.

6 Discussion and Conclusion

We have established finite time iteration complexity bounds for gradient and natural gradient play under the softmax parameterization, considering both unregularized and log-barrier regularized dynamics, in the Markov potential game setting. The complexities of finding a \( \epsilon \)-NE are all on the scale of \( O\left(\frac{1}{\epsilon^2}\right) \). However, for the unregularized case, the complexity bounds have an unpleasant dependence on initialization, while log-barrier regularization manages to remove this dependence with a small compromise of introducing a slightly worse dependence on \( \max |A_i| \) and \( M \). In both the unregularized and regularized settings, natural gradient play achieves better performance than gradient play. This work leaves open a number of interesting questions: for example, how to prove asymptotic convergence for unregularized algorithms without assuming isolated stationary policies, and whether an \( O\left(\frac{1}{\epsilon^2}\right) \) complexity can also be obtained for entropy regularization. Other interesting questions include: first, how to design the reward and state so that the NEs of a MPG lead to good social welfare; second, whether the current results can be extended to a setting without full state observation, for example, where each agent can only observe its local states; and third, can these results be applied to practical multi-agent systems, such as network routing and submodular games.

Acknowledgements

Runyu (Cathy) Zhang would like to thank Shicong Cen for enlightening discussions.

References

Absil, P.-A., Mahony, R., and Andrews, B. Convergence of the iterates of descent methods for analytic cost functions. SIAM Journal on Optimization, 16(2):531–547, 2005.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.
Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bowling, M. and Veloso, M. An analysis of stochastic game theory for multiagent reinforcement learning. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2000.

Buşoniu, L., Babuška, R., and De Schutter, B. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-I*, pp. 183–221, 2010.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Claes, R., Holvoet, T., and Weyns, D. A decentralized approach for anticipatory vehicle routing using delegate multiagent systems. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):364–373, 2011.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *arXiv preprint arXiv:2101.04233*, 2021.

Dechert, W. D. and O’Donnell, S. The stochastic lake game: A numerical solution. *Journal of Economic Dynamics and Control*, 30(9-10):1569–1587, 2006.

Fox, R., McAleer, S., Overman, W., and Panageas, I. Independent natural policy gradient always converges in markov potential games. *CoRR*, abs/2110.10614, 2021. URL [https://arxiv.org/abs/2110.10614](https://arxiv.org/abs/2110.10614).

González-Sánchez, D. and Hernández-Lerma, O. *Discrete-time stochastic control and dynamic potential games: the Euler–Equation approach*. Springer Science & Business Media, 2013.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Inigo-Blasco, P., Diaz-del Rio, F., Romero-Ternero, M. C., Cagigas-Muñiz, D., and Vicente-Diaz, S. Robotics software frameworks for multi-agent robotic systems development. *Robotics and Autonomous Systems*, 60(6):803–821, 2012.

Kakade, S. M. and Langford, J. Approximately optimal approximate reinforcement learning. In Sammut, C. and Hoffmann, A. G. (eds.), *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pp. 267–274. Morgan Kaufmann, 2002.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*, 2021.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. *arXiv preprint arXiv:1711.00832*, 2017.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Levhari, D. and Mirman, L. The great fish war: An example using a dynamic cournot-nash solution. *Bell Journal of Economics*, 11(1):322–334, 1980.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. *CoRR*, abs/2102.11270, 2021. URL [https://arxiv.org/abs/2102.11270](https://arxiv.org/abs/2102.11270).

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Liu, J. and Wu, J. *Multiagent robotic systems*. CRC press, 2018.

Macua, S. V., Zazo, J., and Zazo, S. Learning parametric closed-loop policies for markov potential games. *CoRR*, abs/1802.00899, 2018. URL [http://arxiv.org/abs/1802.00899](http://arxiv.org/abs/1802.00899).
Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In III, H. D. and Singh, A. (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 6820–6829. PMLR, 13–18 Jul 2020.

Mei, J., Dai, B., Xiao, C., Szepesvari, C., and Schuurmans, D. Understanding the effect of stochasticity in policy optimization. Advances in Neural Information Processing Systems, 34, 2021.

Mguni, D. Stochastic potential games. arXiv preprint arXiv:2005.13527, 2020.

Mguni, D., Wu, Y., Du, Y., Yang, Y., Wang, Z., Li, M., Wen, Y., Jennings, J., and Wang, J. Learning in nonzero-sum stochastic games with potentials. arXiv preprint arXiv:2103.09284, 2021.

Monderer, D. and Shapley, L. S. Potential games. Games and economic behavior, 14(1):124–143, 1996.

Roscia, M., Longo, M., and Lazaroiu, G. C. Smart city by multi-agent systems. In 2013 International Conference on Renewable Energy Research and Applications (ICRERA), pp. 371–376. IEEE, 2013.

Shapley, L. S. Stochastic games. Proceedings of the national academy of sciences, 39(10):1095–1100, 1953.

Shoham, Y., Powers, R., and Grenager, T. Multi-agent reinforcement learning: a critical survey. Technical report, Technical report, Stanford University, 2003.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum markov games with a large number of players sample-efficiently?, 2021.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In NIPs, volume 99, pp. 1057–1063. Citeseer, 1999.

Tao, N., Baxter, J., and Weaver, L. A multi-agent, policy-gradient approach to network routing. In In: Proc. of the 18th Int. Conf. on Machine Learning. Citeseer, 2001.

Ventre, A. G., Maturo, A., Hošková-Mayerová, Š., and Kacprzyk, J. Multicriteria and Multiagent Decision Making with Applications to Economics and Social Sciences, volume 305. Springer, 2013.

Zazzo, S., Macua, S. V., Sánchez-Fernández, M., and Zazo, J. Dynamic potential games with constraints: Fundamentals and applications in communications. IEEE Transactions on Signal Processing, 64(14):3806–3821, 2016.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. arXiv preprint arXiv:1911.10633, 2019.

Zhang, R., Ren, Z., and Li, N. Gradient play in multi-agent markov stochastic games: stationary points, convergence, and sample complexity. CoRR, abs/2106.00198, 2021. URL https://arxiv.org/abs/2106.00198
Other notations: We use the abbreviation $\pi_{\theta,i}$ to denote the probability distribution $\pi_{\theta}(\cdot|s)$ (as well as the corresponding $|A_i|$ dimensional vector). We use $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ to denote the $\ell_1, \ell_2, \ell_\infty$ norm respectively. $\text{KL}(\cdot||\cdot)$ is used to denote the KL divergence of two probability distributions. We also define the value function, Q-function, advantage function, averaged Q-function and averaged advantage function with respect to potential function $\phi$ as
\[
V_{\theta}^\phi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s(t), a(t)) \mid \pi_{\theta}, s(0) = s, a(0) = a\right],
Q_{\theta}^\phi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s(t), a(t)) \mid \pi_{\theta}, s(0) = s, a(0) = a\right],
A_{\theta}^\phi(s, a) := Q_{\theta}^\phi(s, a) - V_{\theta}^\phi(s),
\]

Uniform Initialization  

Good Initialization  

Bad Initialization

Figure 2: This set of figures shows how each algorithm performs for different initializations. Uniform initialization is the same as described in Figure 1 for good initialization, we choose the initial parameter as $\pi_1 = [0.1, 0.1, 0.8], \pi_2 = [0.5, 0.5]$; for good initialization, we choose the initial parameter as $\pi_1 = [0.8, 0.1, 0.1], \pi_2 = [0.5, 0.5]$. Figures from top row to bottom plot out $\text{NE-gap}(\theta(t)), \|\nabla_{\theta} \Phi(\theta(t))\|_2, \Phi(\theta(t))$ and $c(\theta(t))$ respectively. Here we choose $\eta = 5, \lambda = 0.003$.

Figure 3: How $\overline{Q_{t}^{(i)}}(a_i), \pi_{t}^{(i)}(a_i)$ change with time when running unregularized natural gradient play. The left two figures plots out $\overline{Q_{1}^{(a)}}(a_1), \pi_{1}^{(a)}(a_1)$ and the right two figures plots out $\overline{Q_{2}^{(a)}}(a_2), \pi_{2}^{(a)}(a_2)$.
A Numerical Simulations

This section provides more material for the numerical example shown in Section 5. Figure 2 displays numerical performance for different initialization policies. All four algorithms perform well given a good initialization, i.e., initial policy close to a stable NE. However for bad initialization that is close to a non-NE stationary point, log-barrier regularized algorithms can escape bad regions and converge to NE much faster than unregularized dynamics.

To examine why multi-agent learning suffers more from getting stuck at undesirable stationary points, we plot out the trajectory for $Q_1(t)(a_1), \pi_1(t)(a_i)$ for both agents in Figure 3. We will mainly focus our attention on the two plots on the left. Note that for the first few steps, $Q_1(t)(a_1 = 2)$ is much larger than $Q_1(t)(a_1 = 3)$, thus the natural gradient play scheme 3 will drive $\pi_1(t)(a_1 = 2)$ close to 1 and $\pi_1(t)(a_1 = 3)$ close to 0 very quickly. However, at around iteration 70, $Q_1(t)(a_1 = 3)$ becomes slightly larger than $Q_1(t)(a_1 = 2)$. Unfortunately, at this stage, most of the probability is assigned to the suboptimal action $a_1 = 2$ and the optimal action receives $\pi_1(t)(a_1 = 3)$ close to zero. Thus it will take more steps to bring $\pi_1(t)(a_1 = 2)$ from 1 to 0 and $\pi_1(t)(a_1 = 3)$ from 0 to 1, which reflects as the trajectory being stuck at the non-NE stationary policy with $\pi_1(a_1 = 3) = 1$ in numerical behavior. From this simulation, we may conclude that one important reason for natural gradient play to get stuck at undesirable stationary points is due to the fact that the value of averaged $Q$-functions $\bar{Q}(s)$ for different actions might switch order during the learning process. In contrast, for single agent bandit learning, the averaged $Q$-function as well as the $Q$-function itself is the same as the reward value of a certain action $r(a)$, and thus will not change order, which explains why it can achieve dimension free convergence in single agent learning.

B Derivation of Gradient and Performance Difference Lemma

Proof. (of Equation 4) According to policy gradient theorem (Sutton et al. 1999):

$$\frac{\partial J_i(\theta)}{\partial \theta_{s,a_i}} = \frac{1}{1 - \gamma} \sum_{t' > t} \left[ \alpha_t(s') - \left( \sum_{a} \sum_{a'} \alpha_t(s) \pi\theta(a'|s') Q_i(s,a) \right) \right]$$

Since for softmax parameterization:

$$\frac{\partial \log \pi\theta(a'|s')}{\partial \theta_{s,a_i}} = 1 \{ a'_i = a_i, s' = s \} - 1 \{ s' = s \} \pi\theta_i(a_i|s)$$

Thus we have that:

$$\frac{\partial J_i(\theta)}{\partial \theta_{s,a_i}} = \frac{1}{1 - \gamma} \sum_{s'} \sum_{a'} d_\theta(s') \pi\theta_i(a'_i|s') \left( 1 \{ a'_i = a_i, s' = s \} - 1 \{ s' = s \} \pi\theta_i(a_i|s) \right) Q_i(s, a')$$

$$= \frac{1}{1 - \gamma} d_\theta(s) \pi\theta_i(a_i|s) \sum_{a'} \pi\theta(a'_i|s) Q_i(s, a_i, a'_i) - \frac{1}{1 - \gamma} d_\theta(s) \pi\theta_i(a_i|s) \sum_{a'} \pi\theta(a'_i|s) Q_i(s, a')$$

$$= \frac{1}{1 - \gamma} d_\theta(s) \pi\theta_i(a_i|s) Q_i(s, a_i, a'_i) - \frac{1}{1 - \gamma} d_\theta(s) \pi\theta_i(a_i|s) V_i(s)$$

$$= \frac{1}{1 - \gamma} d_\theta(s) \pi\theta_i(a_i|s) A_i'(s, a_i)$$

We also introduce a useful lemma used throughout the proof which is derived from the performance difference lemma in MDP (Kakade & Langford 2002).

Lemma B.1. Let $\theta'$ = $(\theta'_i, \theta_{-i})$.

$$J_i(\theta'_i, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s,a_i} d_\theta(s) \pi\theta'(a_i|s) A_i'(s, a_i)$$
Proof. From performance difference lemma \cite{Kakade & Langford 2002}

\[ J_i'(\theta_i, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s,a} d_{\theta}(s) \pi(s|a) A_i' (s, a) \]
\[ = \frac{1}{1 - \gamma} \sum_{s,a} d_{\theta}(s) \pi(s|a) \sum_{a_{-i}} \pi(a_{-i}|s) A_i' (s, a_i, a_{-i}) \]
\[ = \frac{1}{1 - \gamma} \sum_{s,a} d_{\theta}(s) \pi(s|a) A_i'(s, a_i). \]

\[ \square \]

C Derivation of Natural Gradient Play

Lemma C.1.

\[ E_{a \sim \pi_a(s)} \left[ \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s) \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s)^\top \right] = \text{diag}\{\pi_\theta(s)\} - \pi_\theta(s) \pi_\theta^\top(s) \coloneqq F_{i,s}(\theta_i,s), \]
where \( \text{diag}\{\cdot\} \) denotes the diagonal matrix generated by the corresponding vector, and \( \pi_\theta(s) \in \mathbb{R}^{|s|,1} \) is the vector that denotes \( \pi_\theta(s) \). Further, \( F_{i,s}(\theta_i,s) \) is a semi-positive definite matrix, where the eigenvalue 0 has the eigenspace of dimension 1 that is the span of the all-one-vector \( \mathbf{1} \).

Proof. Calculating the gradient using chain rule we have

\[ \frac{\partial \log \pi_{\theta_i}(a_i|s)}{\partial \theta_i(s)} = \mathbf{1}_{\{a_i' = a_i\}} - \pi_{\theta_i}(a_i|s). \]

Let \( \mathbf{1}_{a_i} \in \mathbb{R}^{1,|s|} \) denote the vector where the entry corresponds to \( a_i \) is 1 and other entries are zero. Then

\[ \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s) = \mathbf{1}_{a_i} - \pi_{\theta_i,s} \]
\[ \implies \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s) \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s)^\top = \text{diag}\{\pi_\theta(s)\} - \pi_\theta(s) \pi_\theta^\top(s) = \mathbf{1}_{\{a_i' = a_i\}} - \pi_\theta(s) \pi_\theta^\top(s) + \pi_\theta(s) \pi_\theta^\top(s) \]

Taking the expectation \( E_{a \sim \pi_a(s)} \) we have

\[ E_{a \sim \pi_a(s)} \left[ \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s) \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s)^\top \right] = \text{diag}\{\pi_\theta(s)\} - \pi_\theta(s) \pi_\theta^\top(s) = \text{diag}\{\pi_\theta(s)\} - \pi_\theta(s) \pi_\theta^\top(s) \]
\[ = \text{diag}\{\pi_\theta(s)\} - \pi_\theta(s) \pi_\theta^\top(s) \]

Further, for softmax parameterization, \( \pi_{\theta_i}(a_i|s) > 0, \forall a_i \). Thus \( F_{i,s}(\theta_i,s) \) is a (non-strict) diagonally dominant matrix with diagonal entries all being positive and off-diagonal entries all being negative, in which case the all-one vector \( \mathbf{1} \) is the only eigenvector for eigenvalue 0. \( \square \)

Corollary C.2.

\[ F_i(\theta) = \text{blkdiag}\{d_{\theta}(s) \bar{F}_{i,s}(\theta_i,s)\} \in \mathcal{S}, \]
where \( \text{blkdiag}\{\cdot\} \) denotes the block-diagonal matrix generated by corresponding sub-matrices.

Proof. This is a direct corollary of Lemma C.1, since

\[ \frac{\partial \log \pi_{\theta_i}(a_i|s)}{\partial \theta_i(s')} = 0, \text{ for } s' \neq s, \]

we have that

\[ F_i(\theta) = \mathbb{E}_{a \sim d_{\theta}(\cdot)} \mathbb{E}_{a \sim \pi_a(s)} \left[ \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s) \nabla_{\theta_i,s} \log \pi_{\theta_i}(a_i|s)^\top \right] = \text{blkdiag}\{d_{\theta}(s) \bar{F}_{i,s}(\theta_i,s)\} \]
\[ \square \]

Lemma C.3. For vector \( g: \mathcal{S} \times \mathcal{A} \to \mathbb{R} \), with \( \sum_{a_i} g(s, a_i) = 0, \forall s \in \mathcal{S} \), we have that

\[ [F_i(\theta)](s, a_i) = \frac{1}{d_{\theta}(s) \pi_{\theta_i}(a_i|s)} g(s, a_i) + c(s), \]

where \( c(s) \) is a function that depend on state \( s \) but not on \( a_i \).
Proof. Since $F_i(\theta)$ is a block diagonal matrix,

$$[F_1(\theta)^\top g]_{(s, \cdot)} = \frac{1}{d_\theta(s)} F_{i,s}(\theta_{i,s})^\top g(s, \cdot).$$

From Lemma C.1 since $F_{i,s}$ only has a one-dimensional eigenspace for eigenvalue 0, and the eigenspace is the span of the all-one vector $1$, we have that

$$F_{i,s}(\theta_{i,s})^\top F_{i,s}(\theta_{i,s}) = I - \frac{1}{|A_i|} 11^\top.$$

Let $f(s, a_i) := \frac{1}{d_\theta(s)} \pi_\theta(a_i|s) g(s, a_i)$

$$d_\theta(s) [F_{i,s}(\theta_{i,s}) f(s, \cdot)]_{a_i} = d_\theta(s) \left( \pi_\theta(a_i|s) f(s, a_i) - \pi_\theta(a_i'|s) \sum_{a_i'} \pi_\theta(a_i'|s) f(s, a_i') \right) = g(s, a_i) - \pi_\theta(a_i|s) \sum_{a_i'} g(s, a_i') = g(s, a_i),$$

i.e.,

$$d_\theta(s) F_{i,s}(\theta_{i,s}) f(s, \cdot) = g(s, \cdot)$$

$$\Rightarrow \frac{1}{d_\theta(s)} F_{i,s}(\theta_{i,s})^\top g(s, \cdot) = F_{i,s}(\theta_{i,s})^\top F_{i,s}(\theta_{i,s}) f(s, \cdot) = f(s, \cdot) - c(s) 1,$$

i.e.,

$$[F_1(\theta)^\top g]_{(s, a_i)} = f(s, a_i) - c(s),$$

which completes the proof. \(\square\)

Lemma C.4. Scheme (7) and (8) are equivalent. Similarly, (15) and (16) are equivalent.

Proof. It is not hard to check that $\nabla_\theta J_i(\theta), \nabla_\theta \tilde{J}_i(\theta)$ satisfies

$$\sum_{a_i} [\nabla_\theta J_i(\theta)]_{(s, a_i)} = 0, \quad \sum_{a_i} \left[ \nabla_\theta \tilde{J}_i(\theta) \right]_{(s, a_i)} = 0,$$

thus we can apply Lemma C.3 and conclude

$$F_i(\theta^{(t)})^\top \nabla_\theta J_i(\theta^{(t)}) = \frac{A_i^{(t)}(s, a_i)}{1 - \gamma} + c(s)$$

$$F_i(\theta^{(t)})^\top \nabla_\theta \tilde{J}_i(\theta^{(t)}) = \frac{A_i^{(t)}(s, a_i)}{1 - \gamma} + \frac{\lambda}{d^{(t)}(s) \pi_i^{(t)}(a_i|s)} - \frac{\lambda |A_i|}{d^{(t)}(s)} + c(s),$$

which completes the proof. \(\square\)

D Proofs for Section 3

D.1 Proof of Theorem 3.3

D.1.1 Asymptotic convergence for gradient play

Lemma D.1. For $\eta \leq \frac{(1-\gamma)^2}{\delta n}$, running scheme (6) will guarantee that $\lim_{t \to +\infty} \nabla \Phi(\theta^{(t)}) = 0$. 

16
Proof. Since $\Phi(\theta)$ is $\beta$-smooth w.r.t. $\theta$, where $\beta = \frac{6\alpha}{(1-\gamma)^2}$,

$$
\Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \left( \nabla \Phi(\theta^{(t)}), \theta^{(t+1)} - \theta^{(t)} \right) - \frac{\beta}{2} \| \theta^{(t+1)} - \theta^{(t)} \|_2^2
$$

$$
\geq \frac{\eta}{2} \| \nabla \Phi(\theta^{(t)}) \|_2^2 \geq 0
$$

which proves the monotonicity of $\Phi(\theta^{(t)})$. Since $\phi$ is a bounded function, this gives:

$$
\lim_{t \to +\infty} \| \nabla \Phi(\theta^{(t)}) \|_2 = 0.
$$

From Lemma [D.1] and Assumption [3.2] we know that the limit for $\theta^{(t)}$ exists, i.e., it is valid to define

$$
\theta(\infty) := \lim_{t \to +\infty} \theta^{(t)}.
$$

We abbreviate the related functions with respect to $\theta(\infty)$ as follows:

$$
Q_i^{(\infty)}(s,a) := Q_i^{(\infty)}(s,a), \quad V_i^{(\infty)}(s) := V_i^{(\infty)}(s), \quad A_i^{(\infty)}(s,a) := Q_i^{(\infty)}(s,a) - V_i^{(\infty)}(s)
$$

$$
Q_i^{(\infty)}(s,a) := \sum_{a_{i-1}} \pi_{i-1}(a_{i-1}|s)Q_i^{(\infty)}(s,a_{i-1}), \quad A_i^{(\infty)}(s,a) := \sum_{a_{i-1}} \pi_{i-1}(a_{i-1}|s)A_i^{(\infty)}(s,a_{i-1})
$$

Since $\theta(\infty)$ is the limit of $\theta^{(t)}$, we have that:

$$
\lim_{t \to +\infty} Q_i^{(t)}(s,a_i) = Q_i^{(\infty)}(s,a_i), \quad \lim_{t \to +\infty} A_i^{(t)}(s,a_i) = A_i^{(\infty)}(s,a_i)
$$

(17)

Define:

$$
I_{0}^{i,s} := \{ a_i | Q_i^{(\infty)}(s,a_i) = V^{(\infty)}(s) \} = \{ a_i | A_i^{(\infty)}(s,a_i) = 0 \}
$$

$$
I_{+}^{i,s} := \{ a_i | Q_i^{(\infty)}(s,a_i) > V^{(\infty)}(s) \} = \{ a_i | A_i^{(\infty)}(s,a_i) > 0 \}
$$

$$
I_{-}^{i,s} := \{ a_i | Q_i^{(\infty)}(s,a_i) < V^{(\infty)}(s) \} = \{ a_i | A_i^{(\infty)}(s,a_i) < 0 \}
$$

Let

$$
\Delta := \min_{i} \min_{(s,a_i) \neq 0} | A_i^{(\infty)}(s,a_i) |
$$

(18)

From Lemma [D.7] it is sufficient to show that $I_{0}^{i,s} = \emptyset, \forall i, s$.

From the Lemma [D.1] and the above definitions we have the following corollaries:

**Corollary D.2.** There exists $T_1$, such that $\forall t > T_1, \forall s \in S, \forall i \in \{1,2,\ldots,n\}$,

$$
A_i^{(t)}(s,a_i) < -\frac{\Delta}{4}, \quad \forall a_i \in I_{-}^{i,s}
$$

$$
A_i^{(t)}(s,a_i) > \frac{\Delta}{4}, \quad \forall a_i \in I_{+}^{i,s}
$$

$$
| A_i^{(t)}(s,a_i) | < \frac{\Delta}{4}, \quad \forall a_i \in I_{0}^{i,s}
$$

**Proof.** This is a direct corollary from [17] and [18].

**Corollary D.3.**

$$
\lim_{t \to +\infty} \sum_{a_i \in I_{0}^{i,s}} \pi_i^{(t)}(a_i|s) = 1
$$

$$
\lim_{t \to +\infty} \sum_{a_i \in I_{+}^{i,s} \cup I_{-}^{i,s}} \pi_i^{(t)}(a_i|s) = 0
$$
Proof. This is a direct corollary from Lemma \[D.1\]

\[
\lim_{t \to +\infty} \nabla \Phi(\theta^{(t)}) = 0
\]

\[
\Rightarrow \lim_{t \to +\infty} \frac{d\Phi(\theta^{(t)})}{d\theta_{s,a_i}} = \lim_{t \to +\infty} \frac{1}{1 - \gamma} d^{(t)}(s) \pi^{(t)}_i(a_i|s) A^{(t)}_i(s, a_i) = 0
\]

\[
\Rightarrow \lim_{t \to +\infty} \pi^{(t)}_i(a_i|s) = 0
\]

\[
\Rightarrow \lim_{t \to +\infty} \pi^{(t)}_i(a_i|s) = 0, \quad \forall a_i \notin I^t_0
\]

\[
\Rightarrow \lim_{t \to +\infty} \sum_{a_i \in I^t_0 \cup I^t_-} \pi^{(t)}_i(a_i|s) = 0
\]

Lemma \[D.4\]. \( \forall a_i \in I^t_+, \theta^{(t)}_{s,a_i} \) is bounded from below. \( \forall a_i \in I^t_-, \lim_{t \to +\infty} \theta^{(t)}_{s,a_i} = -\infty \).

Proof. The first statement, \( \forall a_i \in I^t_+, \theta^{(t)}_{s,a_i} \) is bounded from below, is trivial from Corollary \[D.2\]. We only need to prove the second statement. The key observation is that:

\[
\sum_{a_i} \frac{d\Phi(\theta^{(t)})}{d\theta_{s,a_i}} = \frac{1}{1 - \gamma} d^{(t)}(s) \sum_{a_i} \pi^{(t)}_i(a_i|s) A^{(t)}_i(s, a_i) = 0
\]

Thus

\[
\sum_{a_i} \theta^{(t)}_{s,a_i} = \sum_{a_i} \theta^{(0)}_{s,a_i}
\]

From Corollary \[D.3\] we have that

\[
\lim_{t \to +\infty} \sum_{a_i \in I^t_0 \cup I^t_-} \pi^{(t)}_i(a_i|s) = 0
\]

\[
\Rightarrow \exists a_i \in I^t_0, \ s.t. \ \lim_{t \to +\infty} \theta^{(t)}_{s,a_i} = +\infty
\]

And since all \( \theta^{(t)}_{s,a_i} \), sum up to a constant and that \( \forall a_i \in I^t_+, \theta^{(t)}_{s,a_i} \) is bounded from below, we have that:

\[
\exists \ \pi_t \in I^t_0 \cup I^t_-, \ s.t. \ \liminf_{t \to +\infty} \theta^{(t)}_{s,\pi_t} = -\infty.
\]

From Corollary \[D.2\] for \( a_i \in I^t_-, \theta^{(t)}_{s,a_i} \) is monotonically decreasing for \( t > T_1 \), thus

\[
\lim_{t \to +\infty} \theta^{(t)}_{s,a_i} := \theta^{(\infty)}_{s,a_i},
\]

where \( \theta^{(\infty)}_{s,a_i} \) is either a constant or \(-\infty\). We’ll prove by contradiction. Suppose \( \theta^{(\infty)}_{s,a_i} \) is a constant, then for any \( \delta > 0 \) there exists \( T_1' \geq T_1 \) such that \( \forall t \geq T_1', |\theta^{(t)}_{s,a_i} - \theta^{(\infty)}_{s,a_i}| \leq \delta \).

Let \( \pi_t \in A_i \) be defined as in \[\tau(1)\], define:

\[
\tau(t) := \begin{cases} 
  t + 1, & \text{if } \theta^{(t)}_{s,\pi_t} > \theta^{(\infty)}_{s,a_i} - \delta \\
  \min_t (T_1' \leq t' \leq t) \theta^{(t')}_{s,\pi_t} \leq \theta^{(\infty)}_{s,a_i} - \delta, \ \forall t' \leq t & \text{otherwise} 
\end{cases}
\]

We will focus on the set where \( \{t | \tau(t) \leq t\} \). Since \( \lim_{t \to +\infty} \theta^{(t)}_{s,\pi_t} = -\infty \), there are infinitely many elements in this set.

For all \( \tau(t) \leq t \), we have that:

\[
\frac{d\Phi(\theta^{(t)})}{d\theta_{s,a_i}} = \frac{\pi^{(t)}_i(a_i|s) A^{(t)}_i(s, a_i)}{\pi^{(t)}_i(\pi_t|s) A^{(t)}_i(s, \pi_t)} = \exp \left( \theta^{(t)}_{s,a_i} - \theta^{(t)}_{s,\pi_t} \right) \frac{A^{(t)}_i(s, a_i)}{A^{(t)}_i(s, \pi_t)}
\]
Thus

\[ \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} \leq \frac{\Delta(1-\gamma)}{4} \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,\pi}} \], \quad \tau(t) \leq \tau \leq t

\[ \Rightarrow \frac{1}{\eta} (\theta_{s,a_i}^{(t+1)} - \theta_{s,a_i}^{(t)}) = \sum_{\tau(t)} \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} \leq \frac{\Delta(1-\gamma)}{4} \sum_{\tau(t)} \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,\pi}} = \frac{\Delta(1-\gamma)}{4\eta}(\theta_{s,\pi}^{(t+1)} - \theta_{s,\pi}^{(t)}) \quad (20) \]

Since:

\[ \theta_{s,\pi}^{(t)} \geq \theta_{s,\pi}^{(t-1)} - \eta \frac{1}{(1-\gamma)^2} \geq \theta_{s,a_i}^{(\infty)} - \delta - \eta \frac{1}{(1-\gamma)^2} \]

is bounded from below, and that \( \theta_{s,a_i}^{(t)} \) is also bounded from above by \( \theta_{s,a_i}^{(T_i)} \), thus taking \( \lim_{t \to +\infty} \) on both sides of eq. (20) will give

\[ \lim_{t \to +\infty} \theta_{s,a_i}^{(t+1)} - \theta_{s,a_i}^{(t)} \leq \frac{\Delta(1-\gamma)}{4} \left( \lim_{t \to +\infty} \theta_{s,\pi}^{(t+1)} - \theta_{s,\pi}^{(\infty)} + \delta + \eta \frac{1}{(1-\gamma)^2} \right) = -\infty \]

\[ \Rightarrow \lim_{t \to +\infty} \theta_{s,a_i}^{(t)} = -\infty \]

which contradicts the assumption that \( \theta_{s,a_i}^{(\infty)} \) is a constant, and thus we can conclude that

\[ \lim_{t \to +\infty} \theta_{s,a_i}^{(t)} = -\infty, \quad \forall a_i \in I_{++} \]  \( \square \)

**Lemma D.5.** \( \forall a_i^+ \in I_{++} \) for any \( a \in I_{++} \), if there exists \( t \geq T_1 \) such that \( \pi_i^{(t)}(a_i|s) \leq \pi_i^{(t)}(a_i^+|s) \), then for all \( \tau \geq t \), \( \pi_i^{(\tau)}(a_i|s) \leq \pi_i^{(\tau)}(a_i^+|s) \)

**Proof.** We will prove by induction. Suppose for a certain \( \tau \geq t \), it holds that \( \pi_i^{(\tau)}(a_i|s) \leq \pi_i^{(\tau)}(a_i^+|s) \), then:

\[ \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} = \frac{1}{1-\gamma} d^{(\tau)}(s)\pi_i^{(\tau)}(a_i^+|s)\overline{A}_i^{(\tau)}(s,a_i) \]

\[ \geq \frac{1}{1-\gamma} d^{(\tau)}(s)\pi_i^{(\tau)}(a_i|s)\overline{A}_i^{(\tau)}(s,a_i) \]

\[ \geq \frac{1}{1-\gamma} d^{(\tau)}(s)\pi_i^{(\tau)}(a_i|s)\overline{A}_i^{(\tau)}(s,a_i) = \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} \]

Since \( \pi_i^{(\tau)}(a_i|s) \leq \pi_i^{(\tau)}(a_i^+|s) \) \( \Rightarrow \theta_{s,a_i}^{(\tau)} \leq \theta_{s,a_i}^{(\tau)} \), we have:

\[ \theta_{s,a_i}^{(\tau+1)} = \theta_{s,a_i}^{(\tau)} + \eta \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} \geq \theta_{s,a_i}^{(\tau)} + \eta \frac{\partial \Phi(\theta(t))}{\partial \theta_{s,a_i}} = \theta_{s,a_i}^{(\tau+1)} \]

Thus \( \pi_i^{(\tau+1)}(a_i|s) \leq \pi_i^{(\tau+1)}(a_i^+|s) \) also holds, which completes the proof.  \( \square \)

**Lemma D.6.** \( I_{++} = \emptyset \)

**Proof.** We will prove by contradiction. If \( I_{++} \neq \emptyset \), select an arbitrary \( a_i^+ \in I_{++} \) and define

\[ B_0^{(t)}(a_i^+) := \{ a_i \in I_{++} \mid \pi_i^{(t)}(a_i|s) \leq \pi_i^{(t)}(a_i^+|s), \forall t \geq T_1 \} \]

From Lemma D.4 we have that for any \( a_i \in I_{++} \) \( \lim_{t \to +\infty} \pi_i^{(t)}(a_i|t) = 0 \), thus there exists \( T_2 > T_1 \) such that for any \( t \geq T_2 \),

\[ \pi_i^{(t)}(a_i|t) \leq \frac{(1-\gamma)\Delta}{16|A_i|}, \quad \forall a_i \in I_{++}. \]
Additionally, since for any \( a_i \in l_0^{-s} \), \( \lim_{t \to +\infty} A^{(t)}_i(s, a_i) = 0 \), there exists \( T_3 > T_1 \) such that for any \( t \geq T_3 \),

\[
\overline{A^{(t)}_i(s, a_i)} \geq \frac{-\Delta}{16|A_i|}, \quad \forall a_i \in l_0^{s}.
\]

Thus, for \( t \geq \max\{T_2, T_3\} \), from the fact that \( \sum_{a_i} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) = 0 \), we have:

\[
0 = \sum_{a_i \in l_0^{s}} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) + \sum_{a_i \in l_0^{s}} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) + \sum_{a_i \in l_0^{s}} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i)
\]
\[
\geq \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) + \sum_{a_i \in B_0^{s}} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i)
\]
\[
+ \pi_i^{(t)}(a_i^+ | s) \overline{A^{(t)}_i(s, a_i)} + \sum_{a_i \in l_0^{s}} (1 - \gamma) \frac{\Delta}{16|A_i|} \pi_i^{(t)}(a_i^+ | s) A^{(t)}_i(s, a_i)
\]
\[
\geq \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) + |A_i| \pi_i^{(t)}(a_i | s) \frac{-\Delta}{16|A_i|}
\]
\[
+ \pi_i^{(t)}(a_i^+ | s) \frac{\Delta}{4} + |A_i| (1 - \gamma) \frac{\Delta}{16|A_i|} \pi_i^{(t)}(a_i^+ | s) \frac{-1}{1 - \gamma}
\]
\[
\geq \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) + \pi_i^{(t)}(a_i^+ | s) \frac{\Delta}{8}
\]
\[
\implies \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i) < 0.
\]

Thus for \( t \geq \max\{T_2, T_3\} \),

\[
\sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \theta^{(t+1)}_{s, a_i} = \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \theta^{(t)}_{s, a_i} + \frac{\eta}{1 - \gamma} \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) A^{(t)}_i(s, a_i)
\]
\[
< \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \theta^{(t)}_{s, a_i},
\]

which leads to the fact that \( \sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \theta^{(t)}_{s, a_i} \) is bounded from above. Further, from Lemma \( \boxed{D.4} \) \( \theta^{(t)}_{s, a_i} \) is bounded from below, thus the value

\[
\sum_{a_i \in l_0^{s} \setminus B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s)
\]

is bounded from above. However from Corollary \( \boxed{D.3} \)

\[
\lim_{t \to +\infty} \sum_{a_i \in l_0^{s}} \pi_i^{(t)}(a_i | s) \pi_i^{(t)}(a_i^+ | s) = +\infty.
\]

Thus

\[
\lim_{t \to +\infty} \sum_{a_i \in B_0^{s} (a_i^+) \setminus (a_i^+)} \pi_i^{(t)}(a_i | s) \pi_i^{(t)}(a_i^+ | s) = +\infty,
\]

which contradicts the fact that

\[
\pi_i^{(t)}(a_i | s) \leq \pi_i^{(t)}(a_i^+ | s), \quad \forall a_i \in B_0^{s} (a_i^+)
\]

and finishes the proof by contradiction. \( \Box \)
Lemma [D.6] directly implies asymptotic convergence for gradient play as state in Theorem 3.3.

D.1.2 Asymptotic convergence for natural gradient play

The asymptotic convergence for natural gradient play is easier to establish compared with gradient play.

From Lemma [D.9] and the assumption that $\phi(s,a)$ is upper-bounded, we know

$$\lim_{t \to +\infty} \sum_{a_i} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta A_i^{(t)}(s,a_i)}{1 - \gamma} \right) = 1, \forall s, i = 1, 2, \ldots, n.$$ 

Since

$$\sum_{a_i} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta A_i^{(t)}(s,a_i)}{1 - \gamma} \right) = 1 + \frac{\eta^2}{4(1 - \gamma)^2} \sum_{a_i} \pi_i^{(t)}(a_i|s) A_i^{(t)}(s,a_i)^2$$

$$
\Rightarrow \lim_{t \to +\infty} \sum_{a_i} \pi_i^{(t)}(a_i|s) A_i^{(t)}(s,a_i)^2 = 0
$$

$$\Rightarrow \lim_{t \to +\infty} \pi_i^{(t)}(a_i|s) A_i^{(t)}(s,a_i) = 0, \forall s, a_i, i = 1, 2, \ldots, n$$

$$\Rightarrow \lim_{t \to +\infty} \|\nabla_{\theta} \Phi(\theta^{(t)})\|_2 = 0$$

Similar to the proof for gradient play, from Assumption 3.2, we can conclude that $\pi^{(t)}$ converges to some stationary policy $\pi^{(\infty)}$, and we can define $Q_i^{(\infty)}(s,a_i), A_i^{(\infty)}(s,a_i)$ accordingly. Asymptotic convergence is equivalent to

$$I_{+}^{s} := \{ a_i : A_i^{(\infty)}(s,a_i) > 0 \} = \emptyset, \forall s, i = 1, 2, \ldots, n$$

We prove by contradiction. Suppose there exists $a_i^+ \in I_{+}^{s}$ such that $A_i^{(\infty)}(s,a_i^+) > 0$. From $\lim_{t \to +\infty} \pi_i^{(t)}(a_i|s) A_i^{(t)}(s,a_i) = 0$, we have that $\lim_{t \to +\infty} \pi_i^{(t)}(a_i^+|s) = 0$.

Select $a_0$ such that $\lim_{t \to +\infty} \pi_i^{(t)}(a_0|s) = 0$. From $\lim_{t \to +\infty} \pi_i^{(t)}(a_i|s) A_i^{(t)}(s,a_i) = 0$, we have that $\lim_{t \to +\infty} A_i^{(t)}(s,a_0) = 0$.

Thus there exists $\Delta > 0$ and $T$ such that for $t > T$,

$$A_i^{(t)}(s,a_i^+) > \Delta, A_i^{(t)}(s,a_0) < \frac{\Delta}{2}$$

Thus from natural gradient play scheme [S]

$$\frac{\pi_i^{(t)}(a_i^+|s)}{\pi_i^{(t)}(a_0^+|s)} = \pi_i^{(T)}(a_i^+|s) \exp \left( \frac{\eta}{1 - \gamma} \sum_{t=T}^{t-1} A_i^{(t)}(s,a_i^+) - A_i^{(t)}(s,a_0) \right) \geq \pi_i^{(T)}(a_i^+|s)$$

which contradict the fact that $\lim_{t \to +\infty} \frac{\pi_i^{(t)}(a_i^+|s)}{\pi_i^{(t)}(a_0^+|s)} = 0$, and thus completes the proof.

D.2 Proof of Lemma 3.4 and Lemma 3.5

Lemma D.7.

$$\text{NE-gap}_{i}(\theta) \leq \frac{1}{1 - \gamma} \max_{s,a_i} A_i^{(t)}(s,a_i), \quad \text{NE-gap}(\theta) \leq \frac{1}{1 - \gamma} \max_{s,a_i} A_i^{(t)}(s,a_i).$$

Proof. From performance difference lemma

$$J_i(\theta', \theta_{-i}) - J_i(\theta, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s,a_i} d_{\theta'}(s) \pi_{\theta'}(a_i|s) A_i^{(t)}(s,a_i)$$ (Lemma 3.1)
Thus we have that
\[
\text{NE-gap}_1(\theta) \leq \frac{1}{1 - \gamma} \max_{s,a_i} \overline{A}_i(s,a_i), \quad \text{NE-gap}(\theta) \leq \frac{1}{1 - \gamma} \max_{s,a_i} \overline{A}_i(s,a_i).
\]

\[\Box\]

Proof. (of Lemma 3.4) From Lemma D.7 we have that
\[
\text{NE-gap}_1(\theta) \leq \frac{1}{1 - \gamma} \max_{s,a_i} \overline{A}_i(s,a_i).
\]

Since
\[
\max_{a_i} \overline{A}_i(s,a_i) \leq \frac{1}{\sum_{a_i} \pi_{\theta,1}(a_i^*) |s\rangle \sum_{a_i} |\pi_{\theta,1}(a_i|s)\overline{A}_i(s,a_i)|}
\]
\[
\leq \frac{\sqrt{|A|}}{\sum_{a_i} \pi_{\theta,1}(a_i^*) |s\rangle \sum_{a_i} (\pi_{\theta,1}(a_i|s)\overline{A}_i(s,a_i))}
\]
\[
= \frac{\sqrt{|A|}}{\sum_{a_i} \pi_{\theta,1}(a_i^*) |s\rangle \sum_{a_i} (1 - \gamma d_{\theta}(s)\pi_{\theta,1}(a_i|s)\overline{A}_i(s,a_i))}
\]
\[
\leq (1 - \gamma) M(\theta) \sqrt{|A|} \|\nabla_{\theta_i} J_i(\theta)\|_2.
\]

Thus
\[
\text{NE-gap}_1(\theta) \leq \frac{1 - \gamma}{1 - \gamma} \max_{s,a_i} \overline{A}_i(s,a_i)
\]
\[
\frac{\sqrt{|A|} M(\theta)}{\|\nabla_{\theta_i} J_i(\theta)\|_2}
\]

\[\Box\]

Proof. (of Lemma 3.5) From performance difference lemma, let \(\theta' := (\theta^*_i, \theta^*_j)\)
\[
J_i(\theta^*_i, \theta^*_j) - J_i(\theta^*_i, \theta^*_j) = \frac{1}{1 - \gamma} \sum_{s,a_i} d_{\theta'}(s) \pi_{\theta'}(a_i|s) \overline{A}^\theta(s,a_i)
\]

Select \(a^*_i(s) \in \arg\max_{a_i} \overline{A}^\theta(s,a_i)\) and set:
\[
\pi_{\theta'}(a_i|s) = 1\{a_i = a^*_i(s)\},
\]
then
\[
J_i(\theta^*_i, \theta^*_j) - J_i(\theta^*_i, \theta^*_j) = \frac{1}{1 - \gamma} \sum_{s,a_i} d_{\theta'}(s) \pi_{\theta'}(a_i|s) \overline{A}^\theta(s,a_i)
\]
\[
= \frac{1}{1 - \gamma} \sum_{a_i} d_{\theta}(s) \max_{a_i} \overline{A}^\theta(s,a_i) \geq 0.
\]

22
Since \( \theta^* \) is a NE,
\[
\max A^*_i(s, a_i) = 0, \; \forall \; s, \; \forall \; i.
\]
Let \( \Delta := \min_{a_i} \min \Phi(s, a_i) \). Since \( \sum_{a_i} \pi_{\theta^*}(a_i | s) A^*_i(s, a_i) = 0 \)
\[
\Rightarrow 0 = \sum_{a_i \in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s) A^*_i(s, a_i) + \sum_{a_i \not\in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s) A^*_i(s, a_i)
\]
\[
\leq -\Delta \sum_{a_i \not\in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s)
\]
\[
\Rightarrow \sum_{a_i \not\in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s) = 0
\]
\[
\Rightarrow \sum_{a_i \in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s) = 1
\]
\[
\Rightarrow \sum_{a_i \not\in \text{argmax}_{a_i}} \pi_{\theta^*}(a_i | s) = 1
\]
\[
\Rightarrow c(\theta^*) = 1 \quad \square
\]

### D.3 Proof of Theorem 3.6

**Theorem D.8.** (Theorem 3.6 restated) Under Assumption 2.3, 2.4 and 3.2, for \( \eta \leq \frac{(1-\gamma)^2}{6n} \) running gradient play [6] for \( T \) steps will guarantee:

\[
\frac{\sum_{t=0}^{T-1} \text{NE-gap}(\theta(t))^2}{T} \leq \frac{2 \max_i |A_i| (\phi_{\text{max}} - \phi_{\text{min}}) M^2}{(1-\gamma)c^2 \eta T},
\]

where \( c = \inf_{t \leq T} c(\theta(t)) \), from Lemma 3.5 and Theorem 3.3, \( c > 0 \). \( M \) is defined as in [5].

**Proof.** From Lemma F.1, \( \Phi \) is \( \beta \)-smooth with \( \beta = \frac{6n}{(1-\gamma)^2} \), we have that:

\[
\Phi(\theta(t+1)) - \Phi(\theta(t)) \geq \left( \nabla \Phi(\theta(t)), \theta(t+1) - \theta(t) \right) - \frac{\beta}{2} \|\theta(t+1) - \theta(t)\|^2
\]

\[
= (\eta - \frac{\beta \eta^2}{2}) \|\nabla \Phi(\theta(t))\|^2
\]

\[
\geq \frac{\eta}{2} \|\nabla \Phi(\theta(t))\|^2
\]

Summing over \( t \) we get:

\[
\phi_{\text{max}} - \phi_{\text{min}} \geq \frac{1}{\gamma} \Phi(\theta(T)) - \Phi(\theta(0)) \geq \frac{\eta}{2} \sum_{t=0}^{T-1} \|\nabla \Phi(\theta(t))\|^2
\]

From Theorem 3.3 we have that

\[
\|\nabla \Phi(\theta(t))\| \geq \frac{c}{M \sqrt{\max_i |A_i|} \text{NE-gap}(\theta(t))}
\]

Thus

\[
\frac{1}{T} \sum_{t=0}^{T-1} \text{NE-gap}(\theta(t))^2 \leq \frac{2 \max_i |A_i| M^2 (\phi_{\text{max}} - \phi_{\text{min}})}{(1-\gamma)c^2 \eta T}
\]

which completes the proof. \( \square \)
D.4 Proof of Theorem 3.7

Lemma D.9. For \( \eta \leq \frac{(1-\gamma)^2}{2n(\theta_{\max} - \theta_{\min})} \), running scheme \([8]\) will guarantee that

\[
\Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \frac{1}{\eta} \sum_{i=1}^{n} \sum_{a} d(t+1)(s) \log Z_t^{i,s},
\]

where \( Z_t^{i,s} \) is defined by

\[
Z_t^{i,s} := \sum_{a_i} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta A^{(t)}_{i,\phi}(s,a_i)}{1-\gamma} \right).
\]

Proof. From performance difference lemma we have that

\[
\Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) = \frac{1}{1-\gamma} \sum_{s} d(t+1)(s) \sum_{a} \left( \pi^{(t+1)}(a|s) - \pi^{(t)}(a|s) \right) A^{(t)}_{\phi}(s,a).
\]

We define

\[
\overline{A}^{(t)}_{i,\phi}(s,a_i) := \sum_{a_{i-1}} \prod_{j=1}^{i-1} \pi_j^{(t+1)}(a_j|s) \prod_{j=i+1}^{n} \pi_j^{(t)}(a_j|s) A^{(t)}_{\phi}(s,a_i, a_{i-1}).
\]

Then

\[
\Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) = \frac{1}{1-\gamma} \sum_{s} d(t+1)(s) \sum_{a} \left( \pi^{(t+1)}(a|s) - \pi^{(t)}(a|s) \right) A^{(t)}_{\phi}(s,a)
\]

\[
= \frac{1}{1-\gamma} \sum_{s} d(t+1)(s) \sum_{a} \left( \prod_{j=1}^{i-1} \pi_j^{(t+1)}(a_j|s) \prod_{j=i+1}^{n} \pi_j^{(t)}(a_j|s) - \prod_{j=1}^{i-1} \pi_j^{(t+1)}(a_j|s) \prod_{j=i+1}^{n} \pi_j^{(t)}(a_j|s) \right) A^{(t)}_{i,\phi}(s,a_i)
\]

\[
= \frac{1}{1-\gamma} \sum_{s} d(t+1)(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) \overline{A}^{(t)}_{i,\phi}(s,a_i)
\]

\[
\text{Part A}
\]

\[
+ \frac{1}{1-\gamma} \sum_{s} d(t+1)(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) \left( \overline{A}^{(t)}_{i,\phi}(s,a_i) - \overline{A}^{(t)}_{i,\phi}(s,a_i) \right)
\]

\[
\text{Part B}
\]

From scheme \([8]\),

\[
\overline{A}^{(t)}_{i,\phi}(s,a_i) = \frac{1-\gamma}{\eta} \left( \log \left( \frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} \right) + \log \left( Z_t^{i,s} \right) \right)
\]

Substitute this into Part A, we have

\[
\text{Part A} = \frac{1}{\eta} \sum_{s} d(t+1)(s) \sum_{a_i} \overline{A}^{(t)}_{i,\phi}(s,a_i)
\]

\[
= \frac{1}{\eta} \sum_{s} d(t+1)(s) \sum_{a_i} \pi_i^{(t+1)}(a_i|s) \left( \log \left( \frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} \right) + \log \left( Z_t^{i,s} \right) \right)
\]

\[24\]
Thus, when which completes the proof.

\[
\begin{align*}
|\text{Part B}| & \leq \frac{1}{1-\gamma} \sum_{i=1}^{n} d^{(t+1)}(s) \sum_{i=1}^{n} \sum_{a_i} \left| \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right| \left\| \overline{A}^{(t)}_{i,\phi}(s, a_i) - \overline{A}^{(t)}_{i,\phi}(s, a_i) \right\|_1 \\
& \leq \frac{\phi_{\max} - \phi_{\min}}{(1-\gamma)^2} \sum_{i=1}^{n} \sum_{a_i} d^{(t+1)}(s) \sum_{i=1}^{n} \left| \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right| \sum_{j=1}^{n} \left\| \pi_j^{(t+1)} - \pi_j^{(t)} \right\|_1 \\
& \leq \frac{\phi_{\max} - \phi_{\min}}{(1-\gamma)^2} \sum_{i=1}^{n} \sum_{a_i} d^{(t+1)}(s) \left( \sum_{i=1}^{n} \left\| \pi_i^{(t+1)} - \pi_i^{(t)} \right\|_1 \right)^2 \\
& \leq \frac{n(\phi_{\max} - \phi_{\min})}{(1-\gamma)^2} \sum_{i=1}^{n} d^{(t+1)}(s) \sum_{a_i} \left| \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right| \\
& \leq \frac{2n(\phi_{\max} - \phi_{\min})}{(1-\gamma)^2} \sum_{i=1}^{n} d^{(t+1)}(s) \sum_{i=1}^{n} \text{KL}(\pi_i^{(t+1)} || \pi_i^{(t)}) \quad \text{( Pinsker's inequality) }
\end{align*}
\]

Thus, when \( \eta \leq \frac{(1-\gamma)^2}{2n(\phi_{\max} - \phi_{\min})} \), we have that

\[
\Phi(\hat{\theta}^{(t+1)}) - \Phi(\theta^{(t)}) = \text{Part A} + \text{Part B} \\
\geq \left( \frac{1}{\eta} - \frac{2n(\phi_{\max} - \phi_{\min})}{(1-\gamma)^2} \right) \sum_{i=1}^{n} \sum_{s} d^{(t+1)}(s) \text{KL}(\pi_i^{(t+1)} || \pi_i^{(t)}) + \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s} d^{(t+1)}(s) \log \left( Z_t^{i,s} \right)
\]

which completes the proof.

**Lemma D.10.** For \( \eta \leq (1-\gamma)^2 \)

\[
\sum_{i=1}^{n} \sum_{s} d^{(t+1)}(s) \log Z_t^{i,s} \geq \frac{\exp^2}{3M} \text{NE-gap}(\theta^{(t)})^2
\]

**Proof.** From Lemma D.7, we have that \( \text{NE-gap}(\theta) \leq \frac{1}{1-\gamma} \max_i \max_{s,a_i} \overline{A}_i^{(t)}(s, a_i) \). On the other hand,

\[
\begin{align*}
Z_t^{i,s} &= \sum_{a_i} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta \overline{A}_i^{(t)}(s, a_i)}{1-\gamma} \right) \\
&= \sum_{a_i \notin \text{argmax}_{a_i} \overline{Q}_i^{(t)}(s, a_i)} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta \overline{A}_i^{(t)}(s, a_i)}{1-\gamma} \right) + \sum_{a_i \notin \text{argmax}_{a_i} \overline{Q}_i^{(t)}(s, a_i)} \pi_i^{(t)}(a_i|s) \exp \left( \frac{\eta \max_{a_i} \overline{A}_i^{(t)}(s, a_i)}{1-\gamma} \right)
\end{align*}
\]

25
Theorem D.11. (Theorem 3.7 restate) Under Assumption 2.3, 2.4 and 3.2, for
\[
\sum_{t=0}^{T-1} \frac{\text{NE-gap}(\theta^{(t)})^2}{T} \leq \frac{3(\phi_{\text{max}} - \phi_{\text{min}})M}{(1 - \gamma)cT},
\]
(23)
running natural gradient play \([5]\) will guarantee:
Proof. Combining Lemma [D.9] and [D.10] we have
\[
\Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s} d^{(t+1)}(s) \log Z_{i,s}^{t+1}
\]
\[
\geq \frac{c_\eta}{3M} \text{NE-gap}(\theta^{(t)})^2
\]
Summing over \( t \) we have
\[
\frac{\phi_{\max} - \phi_{\min}}{1 - \gamma} \geq \Phi(\theta^{(T)}) - \Phi(\theta^{(0)}) \geq \frac{c_\eta}{3M} \sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)})^2,
\]
thus
\[
\sum_{t=0}^{T-1} \frac{\text{NE-gap}(\theta^{(t)})^2}{T} \leq \frac{3M(\phi_{\max} - \phi_{\min})}{(1 - \gamma) c_\eta T},
\]
which completes the proof. \( \square \)

E Proof for log-barrier regularization

E.1 Proof of Theorem 4.2

Proof. (of Lemma 4.1) From \( \|\nabla \theta J_i(\theta)\|_2 \leq \frac{\lambda}{2} \) we have that
\[
\frac{\partial J_i(\theta)}{\partial \theta_{s,a_i}} = \frac{1}{1 - \gamma} d_\theta(s) \pi_{\theta_i}(a_i|s) \overline{A}_i^\theta(s,a_i) + \lambda - \lambda |A_i| \pi_{\theta_i}(a_i|s)
\]
\[
= \pi_{\theta_i}(a_i|s) \left( \frac{1}{1 - \gamma} d_\theta(s) \overline{A}_i^\theta(s,a_i) - \lambda |A_i| \right) + \lambda \leq \lambda
\]
\[
\Rightarrow \pi_{\theta_i}(a_i|s) \left( \frac{1}{1 - \gamma} d_\theta(s) \overline{A}_i^\theta(s,a_i) - \lambda |A_i| \right) \leq 0
\]
\[
\Rightarrow \frac{1}{1 - \gamma} d_\theta(s) \overline{A}_i^\theta(s,a_i) - \lambda |A_i| \leq 0
\]
\[
\Rightarrow \overline{A}_i^\theta(s,a_i) \leq \frac{\lambda |A_i|(1 - \gamma)}{d_\theta(s)} \leq \lambda |A_i|(1 - \gamma)M
\]

Thus,
\[
\text{NE-gap}_i(\theta) = \sup_{\theta_i^*} J_i(\theta_i^*, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} d_\theta(s) \pi_{\theta_i}(a_i|s) \overline{A}_i^\theta(s,a_i)
\]
\[
\leq \frac{1}{1 - \gamma} \sum_{s,a_i} d_\theta(s) \max_{\theta_{-i}} \overline{A}_i^\theta(s,a_i)
\]
\[
\leq \frac{1}{1 - \gamma} \sum_{s,a_i} d_\theta(s) |A_i| \pi_{\theta_i}(a_i|s) \overline{A}_i^\theta(s,a_i)
\]
\[
\leq \lambda |A_i| M.
\]

We now prove Theorem 4.2

Theorem E.1. (Theorem 4.2 restated) Under Assumption 2.3 and 2.4, for \( \eta \leq \frac{\lambda M \max |A_i|}{\lambda M \max |A_i| (1 - \gamma)^2} \), and \( \lambda = \frac{\lambda}{M \max |A_i|} \), set \( \theta^{(0)} \) be the uniform random policy, i.e., \( \theta^{(0)} = 0 \), then running gradient play [14] for \( T \) steps, where
\[
T \geq \frac{2 \max_i |A_i|^2 \phi_{\max} - \phi_{\min})M^2}{(1 - \gamma)\eta^2}
\]

27
\[ \sim O \left( \frac{n \max_i |A_i|^2 (\phi_{\max} - \phi_{\min}) M^2}{(1 - \gamma)^4 \epsilon^2} \right) \]

will guarantee that

\[ \min_{0 \leq t \leq T-1} \text{NE-gap}(\theta(t)) \leq \epsilon \]

**Proof.** From Lemma F.2, \( \tilde{\Phi} \) is \( \beta \)-smooth with \( \beta = \frac{6n}{(1-\gamma)^2} + 2 \lambda \max_i |A_i| \), we have that:

\[
\tilde{\Phi}(\theta(t+1)) - \tilde{\Phi}(\theta(t)) \geq \left( \nabla \tilde{\Phi}(\theta(t)), \theta(t+1) - \theta(t) \right) - \frac{\beta}{2} \| \theta(t+1) - \theta(t) \|^2 \\
= \left( \eta - \frac{\beta \eta^2}{2} \right) \| \nabla \Phi(\theta(t)) \|^2 \\
\geq \frac{\eta}{2} \| \nabla \Phi(\theta(t)) \|^2
\]

For \( \theta(0) = 0 \), summing over \( t \) we get:

\[
\frac{\phi_{\max} - \phi_{\min}}{1 - \gamma} \geq \tilde{\Phi}(\theta(T)) - \tilde{\Phi}(\theta(0)) \geq \frac{\eta}{2} \sum_{t=0}^{T-1} \| \nabla \Phi(\theta(t)) \|^2
\]

Thus,

\[
\min_{0 \leq t \leq T-1} \| \nabla \tilde{\Phi}(\theta(t)) \| \leq \frac{2(\phi_{\max} - \phi_{\min})}{(1 - \gamma) \eta T}.
\]

Thus for

\[
T \geq \frac{2(\phi_{\max} - \phi_{\min})}{(1 - \gamma) \eta \lambda^2} = \frac{2 \max_i |A_i|^2 (\phi_{\max} - \phi_{\min}) M^2}{(1 - \gamma) \eta \epsilon^2},
\]

it can be guaranteed that

\[
\min_{0 \leq t \leq T-1} \| \nabla \tilde{\Phi}(\theta(t)) \| \leq \lambda = \frac{\epsilon}{\max_i |A_i| M}.
\]

Then applying Lemma 4.1 completes the proof.

**E.2 Proof of Theorem 4.3**

For notational simplicity, we define the following variables:

\[
f_i^{(t)}(s, a_i) := \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) + \frac{\lambda}{d^{(t)}(s) \pi_i^{(t)}(a_i|s)} - \frac{\lambda |A_i|}{d^{(t)}(s)}
\]

\[
Z_i^{t,s} := \sum_{a_i} \pi_i^{(t)}(a_i|s) \exp \left( \eta f_i^{(t)}(s, a_i) \right)
\]

\[
\Delta_i^{(t)}(s, a_i) := \frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} - 1
\]

**Lemma E.2.**

\[
\sum_{a_i} \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i) = 0
\]

\[
\sum_{a_i} \pi_i^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) = 0
\]

\[
Z_i^{t,s} \geq 1
\]
Proof. From the definition of $f_i^{(t)}(s,a_i),\Delta_i^{(t)}(s,a_i),$

$$
\sum_{a_i} \pi_i^{(t)}(a_i|s)f_i^{(t)}(s,a_i) = \frac{1}{1 - \gamma} \sum_{a_i} \pi_i^{(t)}(a_i|s)A_i^{(t)}(s,a_i) + \lambda \sum_{a_i} \pi_i^{(t)}(a_i|s)\frac{1}{d^{(t)}(s)\pi^{(t)}(a_i|s)} - \frac{\lambda|A_i|}{d^{(t)}(s)} \sum_{a_i} \pi_i^{(t)}(a_i|s)
$$

$$
= \frac{1}{d^{(t)}(s)}(\lambda|A_i| - \lambda|A_i|) = 0
$$

$$
\sum_{a_i} \pi_i^{(t)}(a_i|s)\Delta_i^{(t)}(s,a_i) = \sum_{a_i} \pi_i^{(t)}(a_i|s)\left(\frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} - 1\right)
$$

$$
= \sum_{a_i} \pi_i^{(t+1)}(a_i|s) - \sum_{a_i} \pi_i^{(t)}(a_i|s) = 1 - 1 = 0
$$

Using the fact that $e^x \geq 1 + x,$

$$Z_t^{s,a} = \sum_{a_i} \pi_i^{(t)}(a_i|s)\exp\left(\eta f_i^{(t)}(s,a_i)\right)
$$

$$
\geq \sum_{a_i} \pi_i^{(t)}(a_i|s)\left(1 + \eta f_i^{(t)}(s,a_i)\right)
$$

$$
\geq \sum_{a_i} \pi_i^{(t)}(a_i|s) + \eta \sum_{a_i} \pi_i^{(t)}(a_i|s)f_i^{(t)}(s,a_i) \geq 1.
$$

Lemma E.3. For $\eta \leq \frac{1}{15\left(\frac{1}{1 - \gamma^2} + \lambda|A_i|M\right)},\theta^{(0)} = 0,$ running scheme \[16\] will guarantee that

$$\pi_i^{(t)}(a_i|s) \geq \frac{\lambda}{4\left(\lambda|A_i|M + \frac{1}{1 - \gamma^2}\right)}.$$

Proof. We will prove by induction. For $\theta^{(0)} = 0$ apparently $\pi^{(0)}$ satisfies the lower bound. Suppose that

$$\pi_i^{(t)}(a_i|s) \geq \frac{\lambda}{4\left(\lambda|A_i|M + \frac{1}{1 - \gamma^2}\right)},$$

then from the definition of $f_i^{(t)}(s,a)$ we have that

$$-rac{1}{(1 - \gamma)^2} - \lambda|A_i|M \leq f_i^{(t)}(s,a_i) \leq 5 \left(-\frac{1}{(1 - \gamma)^2} + \lambda|A_i|M\right)$$

Thus

$$-\frac{1}{15} \leq \eta f_i^{(t)}(s,a_i) \leq \frac{1}{3},$$

which leads to the fact that

$$Z_t^{s,a} = \sum_{a_i} \pi_i^{(t)}(a_i|s)\exp\left(\eta f_i^{(t)}(s,a_i)\right)
$$

$$\leq \sum_{a_i} \pi_i^{(t)}(a_i|s)\left(1 + (\eta f_i^{(t)}(s,a_i)) + (\eta f_i^{(t)}(s,a_i))^2\right) (e^x \leq 1 + x + x^2, \text{for } -\frac{1}{15} \leq x \leq \frac{1}{3})$$

$$= 1 + \sum_{a_i} \pi_i^{(t)}(a_i|s)\left(\eta f_i^{(t)}(s,a_i)\right)^2
$$

$$\leq 1 + \frac{1}{3^2} = \frac{10}{9}. \quad (24)
$$

Thus we have that

$$\frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} = \frac{\exp\left(\eta f_i^{(t)}(s,a_i)\right)}{Z_t^{s,a}} \geq \frac{1 + \eta f_i^{(t)}(s,a_i)}{Z_t^{s,a}} \geq \frac{1 - \frac{1}{15}}{\frac{10}{9}} = \frac{21}{25}.$$
Thus, for $a_i$ such that $\pi^{(t)}_i(a_i|s) \geq \frac{\lambda}{3(\lambda|A_i|M + \frac{1}{(1-\gamma)^2})}$, we have

$$
\pi^{(t+1)}_i(a_i|s) \geq \frac{21}{25} \frac{\lambda}{3(\lambda|A_i|M + \frac{1}{(1-\gamma)^2})} \geq \frac{\lambda}{4(\lambda|A_i|M + \frac{1}{(1-\gamma)^2})}.
$$

On the other hand, for $a_i$ such that $\pi^{(t)}_i(a_i|s) < \frac{\lambda}{3(\lambda|A_i|M + \frac{1}{(1-\gamma)^2})}$, we have

$$
f^{(t)}_i(s,a_i) \geq -\frac{1}{(1-\gamma)^2} - \lambda|A_i|M + 3 \left(\frac{1}{(1-\gamma)^2} + \lambda|A_i|M\right) = 2 \left(\frac{1}{(1-\gamma)^2} + \lambda|A_i|M\right),
$$

then according to the induction assumption, we have

$$
\pi^{(t+1)}_i(a_i|s) = \exp \left(\eta f^{(t)}_i(s,a_i)\right) \geq \frac{1 + \eta f^{(t)}_i(s,a_i)}{Z^{t,s}_i} \geq \frac{1 + 25\eta \left(\frac{1}{(1-\gamma)^2} + \lambda|A_i|M\right)}{1 + \frac{5}{3} \eta \left(\frac{1}{(1-\gamma)^2} + \lambda|A_i|M\right)} \geq 1,
$$

Thus

$$
\frac{\pi^{(t+1)}_i(a_i|s)}{\pi^{(t)}_i(a_i|s)} = \frac{\exp \left(\eta f^{(t)}_i(s,a_i)\right)}{Z^{t,s}_i} \geq \frac{1 + \eta f^{(t)}_i(s,a_i)}{Z^{t,s}_i} \geq \frac{1 + \frac{25}{12}}{1 + \frac{5}{3} \eta \left(\frac{1}{(1-\gamma)^2} + \lambda|A_i|M\right)} \geq 4\left(1 + \frac{1}{25}\right) \geq \frac{4}{5},
$$

then according to the induction assumption, we have

$$
\pi^{(t+1)}_i(a_i|s) \geq \frac{\lambda}{4(\lambda|A_i|M + \frac{1}{(1-\gamma)^2})},
$$

which completes the proof of the lemma.

□

**Corollary E.4.** Under the condition of Lemma E.3 running (16) will guarantee that

$$
-\frac{1}{15} \leq \eta f^{(t)}_i(s,a_i) \leq \frac{1}{3}, \quad Z^{t,s}_i \leq \frac{10}{9}, \quad -\frac{1}{5} \leq \Delta^{(t)}_i(s,a_i) \leq \frac{1}{2}.
$$

**Proof.** The first two inequalities are proved in the proof of Lemma E.3; we only need to show $-\frac{1}{5} \leq \Delta^{(t)}_i(s,a_i) \leq \frac{1}{2}$. In the proof of Lemma E.3, we have already shown that

$$
\frac{\pi^{(t+1)}_i(a_i|s)}{\pi^{(t)}_i(a_i|s)} = \frac{\exp \left(\eta f^{(t)}_i(s,a_i)\right)}{Z^{t,s}_i} \geq \frac{1 + \eta f^{(t)}_i(s,a_i)}{Z^{t,s}_i} \geq \frac{1 + \frac{1}{15}}{1 + \frac{1}{25}} = \frac{21}{25} \geq \frac{4}{5},
$$

thus

$$
\Delta^{(t)}_i(s,a_i) \geq \frac{4}{5} - 1 = -\frac{1}{5}.
$$

On the other hand,

$$
\frac{\pi^{(t+1)}_i(a_i|s)}{\pi^{(t)}_i(a_i|s)} = \frac{\exp \left(\eta f^{(t)}_i(s,a_i)\right)}{Z^{t,s}_i} \leq \exp \left(\eta f^{(t)}_i(s,a_i)\right) \leq \exp \left(\frac{1}{3}\right) \leq \frac{3}{2},
$$

thus

$$
\Delta^{(t)}_i(s,a_i) \leq \frac{3}{2} - 1 = \frac{1}{2},
$$

which completes the proof of the corollary. □
Lemma E.5.

\[ \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \left( \frac{1}{2\eta} - 4\lambda \max_i |A_i| M^2 - \frac{4M}{(1-\gamma)^3} \right) \sum_{i=1}^{n} \sum_{s,a_i} d_i(s) \pi_i^{(t)}(a_i|s) \Delta_i^{(t)}(s,a_i)^2 \]

Proof. Let \( \tilde{\theta}^{i,(t)} \) be defined as:

\[ \tilde{\theta}^{i,(t)} := (\theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_i^{(t+1)}, \ldots, \theta_n^{(t+1)}) \]

Then we have that

\[ \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) = \sum_{i=1}^{n} \Phi(\tilde{\theta}^{i,(t)}) - \Phi(\tilde{\theta}^{i+1,(t)}) \]

\[ = \sum_{i=1}^{n} J_i(\tilde{\theta}^{i,(t)}) - J_i(\tilde{\theta}^{i+1,(t)}) \]

\[ = \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \pi_i^{(t+1)}(a_i|s) A_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) \] (Lemma B.1)

\[ = \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) \]

Thus

\[ \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) = \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) + \lambda \sum_{i} \sum_{s,a_i} \log \left( \frac{\pi_i^{(t+1)}(a_i|s)}{\pi_i^{(t)}(a_i|s)} \right) \]

\[ = \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) + \lambda \sum_{i} \sum_{s,a_i} \log \left( 1 + \Delta_i^{(t)}(s,a_i) \right) \]

\[ = \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) + \lambda \sum_{i} \sum_{s,a_i} \log \left( 1 + \Delta_i^{(t)}(s,a_i) \right) \]

\[ + \lambda \sum_{i} \sum_{s,a_i} \log \left( 1 + \Delta_i^{(t)}(s,a_i) \right) - \Delta_i^{(t)}(s,a_i) \] \hspace{1cm} \text{Part A}

\[ + \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) - Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) \] \hspace{1cm} \text{Part B}

\[ + \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) \] \hspace{1cm} \text{Part C}

\[ + \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} d_{i,(t)}(s) \sum_{a_i} \left( \pi_i^{(t+1)}(a_i|s) - \pi_i^{(t)}(a_i|s) \right) Q_i^{\tilde{\theta}^{i+1,(t)}}(s,a_i) . \] \hspace{1cm} \text{Part D
We will now bound each part separately. We first get a lower bound for part A:

\[ \text{Part A} = \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \left( \pi^{(t+1)}(a_i|s) - \pi^{(t)}(a_i|s) \right) \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) \right) + \frac{\lambda}{d^{(t)}(s)} \Delta_i^{(t)}(s, a_i) \]

\[ = \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \left[ \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \right] \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) \right) + \frac{\lambda}{d^{(t)}(s)} \Delta_i^{(t)}(s, a_i) \]

\[ = \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) \right) + \frac{\lambda}{d^{(t)}(s)} \pi^{(t)}(a_i|s) - \frac{\lambda |A_i|}{d^{(t)}(s)} \Delta_i^{(t)}(s, a_i) \]

\[ = \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) f^{(t)}(s, a_i) \]

\[ = \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \left[ \sum_{s,a_i} \frac{1}{\eta} \left( \log \frac{\pi^{(t+1)}(a_i|s)}{\pi^{(t)}(a_i|s)} + \log(Z_i^{(s)}) \right) \right] \]

\[ = \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) + \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \log(Z_i^{(s)}) \sum_{a_i} \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \]

\[ = \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) \]

\[ = \frac{1}{\eta} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i) \left| \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) \right| \]

From the boundedness of \( \Delta_i^{(t)}(s, a_i) \) in Corollary E.4, we have that

\[ \left| \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) \right| \geq \frac{1}{2} \left| \Delta_i^{(t)}(s, a_i) \right| \]

Substitute this into the above inequalities, we get

\[ \text{Part A} \geq \frac{1}{2\eta} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i)^2 \]

Now we will give a lower bound for part B. Similarly, from the boundedness of \( \Delta_i^{(t)}(s, a_i) \) in Corollary E.4, we have that

\[ \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) - \Delta_i^{(t)}(s, a_i) \geq -\Delta_i^{(t)}(s, a_i)^2. \]

Thus

\[ \text{Part B} = \lambda \sum_{i} \sum_{s,a_i} \log \left( 1 + \Delta_i^{(t)}(s, a_i) \right) - \Delta_i^{(t)}(s, a_i) \geq -\lambda \sum_{i} \sum_{s,a_i} \Delta_i^{(t)}(s, a_i)^2. \]

Additionally, using Lemma E.3

\[ \text{Part B} \geq -\lambda \sum_{i} \sum_{s,a_i} \Delta_i^{(t)}(s, a_i)^2 \]

\[ \geq -4 \left( \lambda \max_i |A_i| M + \frac{1}{(1 - \gamma)^2} \right) \sum_{i} \sum_{s,a_i} \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i)^2 \]

\[ \geq -4M \left( \lambda \max_i |A_i| M + \frac{1}{(1 - \gamma)^2} \right) \sum_{i} \sum_{s,a_i} d^{(t)}(s) \pi^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i)^2. \]
We will now move on to bound the absolute value of Part C.

\[ |\text{Part C}| \leq \frac{1}{1 - \gamma} \sum_{i=1}^{n} \sum_{s} d^{(t)}(s) \sum_{a_i} \left| \pi_i^{(t+1)}(a_i | s) - \pi_i^{(t)}(a_i | s) \right| |Q_i^{\bar{g}^{t+1},(t)}(s, a_i) - Q_i^{\bar{g}^{t},(t)}(s, a_i)|. \]

Since

\[ |Q_i^{\bar{g}^{t+1},(t)}(s, a_i) - Q_i^{\bar{g}^{t},(t)}(s, a_i)| \leq \sum_{a_i, a_{-i}} |\pi_{\bar{g}_i^{t+1}}(a_{-i} | s) - \pi_{\bar{g}_i^{t}}(a_{-i} | s)| |Q_i^{(t)}(s, a_i)| \]

\[ + \sum_{a_{-i}} |\pi_{\bar{g}_i^{t+1}}(a_{-i} | s) - \pi_{\bar{g}_i^{t}}(a_{-i} | s)| \left| Q_i^{(t)}(s, a_i) \right| \]

\[ \leq \max_a |Q_i^{\bar{g}^{t+1},(t)}(s, a) - Q_i^{(t)}(s, a)| + \frac{1}{1 - \gamma} \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1. \]

From Lemma G.1

\[ \max_a |Q_i^{\bar{g}^{t+1},(t)}(s, a) - Q_i^{(t)}(s, a)| \leq \frac{1}{(1 - \gamma)^2} \max_s \| \pi_{\bar{g}_i^{t+1}} - \pi_{\bar{g}_i^{t}} \|_1 \]

\[ \leq \frac{1}{(1 - \gamma)^2} \max_s \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1 \]

Thus we have that

\[ |Q_i^{\bar{g}^{t+1},(t)}(s, a_i) - Q_i^{\bar{g}^{t},(t)}(s, a_i)| \leq \max_a |Q_i^{\bar{g}^{t+1},(t)}(s, a) - Q_i^{(t)}(s, a)| + \frac{1}{1 - \gamma} \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1 \]

\[ \leq \frac{2}{(1 - \gamma)^2} \max_s \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1. \]

Thus

\[ |\text{Part C}| \leq \frac{1}{1 - \gamma} \sum_{i=1}^{n} \sum_{s} d^{(t)}(s) \sum_{a_i} \left| \pi_i^{(t+1)}(a_i | s) - \pi_i^{(t)}(a_i | s) \right| \frac{2}{(1 - \gamma)^2} \max_s \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1 \]

\[ \leq \frac{2}{(1 - \gamma)^3} \left( \max_s \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1 \right)^2 \left( \sum_{i=1}^{n} \sum_{s} d^{(t)}(s) \sum_{a_i} \left| \pi_i^{(t+1)}(a_i | s) - \pi_i^{(t)}(a_i | s) \right| \right) \]

\[ \leq \frac{2}{(1 - \gamma)^3} \left( \max_s \sum_{j=1}^{n} \| \pi_{j,s}^{(t+1)} - \pi_{j,s}^{(t)} \|_1 \right)^2 \left( \sum_{j=1}^{n} \sum_{a_j} \pi_j^{(t)}(a_j | s) \Delta_j^{(t)}(s, a_j) \right)^2 \]

\[ = \sum_{j=1}^{n} \sum_{a_j} \pi_j^{(t)}(a_j | s) \Delta_j^{(t)}(s, a_j)^2 \]

33
Thus

\[
|\text{Part C}| \leq \frac{2n}{(1-\gamma)^2} \sum_{i=1}^{n} \sum_{a_{i}} \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(s, a_{i})^2
\]

\[
\leq \frac{2nM}{(1-\gamma)^2} \sum_{i=1}^{n} \sum_{a_{i}} d^{(t)}(s) \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(s, a_{i})^2
\]

Lastly, we will bound the absolute value of Part D.

\[
|\text{Part D}| = \left| \frac{1}{1-\gamma} \sum_{i=1}^{n} \sum_{s} (d_{\pi^{(t)}_i}(s) - d^{(t)}(s)) \sum_{a_{i}} \left( \pi^{(t+1)}_i(a_{i}|s) - \pi^{(t)}_i(a_{i}|s) \right) Q^{\pi^{(t+1)}_i}(s, a_{i}) \right|
\]

\[
\leq \frac{1}{(1-\gamma)^2} \sum_{i=1}^{n} \sum_{s} \left| d_{\pi^{(t)}_i}(s) - d^{(t)}(s) \right| \sum_{a_{i}} \left| \pi^{(t+1)}_i(a_{i}|s) - \pi^{(t)}_i(a_{i}|s) \right|
\]

\[
\leq \frac{1}{(1-\gamma)^2} \sum_{i=1}^{n} \max_{s} \|\pi^{(t)}_i\|_1 \sum_{s} \sum_{a_{i}} \left| d_{\pi^{(t)}_i}(s) - d^{(t)}(s) \right|.
\]

From Corollary G.3

\[
\sum_{s} \left| d_{\pi^{(t)}_i}(s) - d^{(t)}(s) \right| \leq \frac{1}{1-\gamma} \max_{s} \left\| \pi^{(t)}_i(a_{i}|s) - \pi^{(t)}(a_{i}|s) \right\|_1
\]

\[
\leq \frac{1}{1-\gamma} \max_{s} \sum_{i=1}^{n} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1.
\]

Thus

\[
|\text{Part D}| \leq \frac{1}{(1-\gamma)^2} \sum_{i=1}^{n} \max_{s} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1 \sum_{s} \sum_{a_{i}} \left| d_{\pi^{(t)}_i}(s) - d^{(t)}(s) \right|
\]

\[
\leq \frac{1}{(1-\gamma)^2} \left( \sum_{i=1}^{n} \max_{s} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1 \right) \left( \max_{s} \sum_{i=1}^{n} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1 \right)
\]

\[
\leq \frac{1}{(1-\gamma)^2} \left( \sum_{i=1}^{n} \max_{s} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1 \right)^2.
\]

From Cauchy-Schwarz inequality

\[
\left( \sum_{i=1}^{n} \max_{s} \|\pi^{(t+1)}_i - \pi^{(t)}_i\|_1 \right)^2 \leq \sum_{i=1}^{n} \max_{s} \left( \sum_{a_{i}} \left| \pi^{(t+1)}_i(a_{i}|s) - \pi^{(t)}_i(a_{i}|s) \right| \Delta^{(t)}_i(a_{i}|s) \right)^2
\]

\[
= \sum_{i=1}^{n} \max_{s} \left( \sum_{a_{i}} \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(a_{i}|s) \right)^2
\]

\[
\leq \sum_{i=1}^{n} \max_{s} \left( \sum_{a_{i}} \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(a_{i}|s) \right)^2
\]

\[
\leq \sum_{i=1}^{n} \sum_{a_{i}} \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(a_{i}|s)^2
\]

\[
\leq \sum_{i=1}^{n} \sum_{a_{i}} \pi^{(t)}_i(a_{i}|s) \Delta^{(t)}_i(a_{i}|s)^2.
\]
Thus
\[ |\text{Part D}| \leq \frac{n}{(1 - \gamma)^3} \sum_{i=1}^{n} \sum_{s,a_i} \pi_i(t)(a_i|s)\Delta_i(t)(a_i|s)^2 \]
\[ \leq \frac{nM}{(1 - \gamma)^3} \sum_{i=1}^{n} \sum_{s,a_i} d_i(t)(s)\pi_i(t)(a_i|s)\Delta_i(t)(a_i|s)^2, \]

Combining the bounds on Part A, B, C, D we get
\[ \Phi(\theta^{t+1}) - \Phi(\theta^t) = \text{Part A} + \text{Part B} + \text{Part C} + \text{Part D} \]
\[ \geq \left( \frac{1}{2q} - 4\lambda \max_i |A_i| M^2 - \frac{4M}{(1 - \gamma)^3} - \frac{3nM}{(1 - \gamma)^3} \right) \sum_{i=1}^{n} \sum_{s,a_i} d_i(t)(s)\pi_i(t)(a_i|s)\Delta_i(t)(a_i|s)^2, \]

which completes the proof.

\[ \square \]

**Lemma E.6.** Under the condition as in Lemma E.3
\[ \sum_{i=1}^{n} \sum_{s,a_i} d_i(t)(s)\pi_i(t)(a_i|s)\Delta_i(t)(s,a_i)^2 \geq \frac{\eta^2}{9} \sum_{i=1}^{n} \sum_{s,a_i} d_i(t)(s)\pi_i(t)(a_i|s)f_i(t)(s,a_i)^2 \]

**Proof.** Recall from the definition of \( \Delta_i(t)(s,a_i) \):
\[ \Delta_i(t)(s,a_i) = \frac{\exp(\eta f_i(t)(s,a_i))}{Z_i^{t,s}} - 1. \]

Thus
\[ \sum_{a_i} \pi_i(t)(a_i|s)\Delta_i(t)(s,a_i)^2 = \frac{1}{(Z_i^{t,s})^2} \sum_{a_i} \pi_i(t)(a_i|s) \left( \exp(\eta f_i(t)(s,a_i)) - Z_i^{t,s} \right)^2 \]
\[ = \frac{1}{(Z_i^{t,s})^2} \left[ \sum_{a_i} \pi_i(t)(a_i|s) \left( \exp(\eta f_i(t)(s,a_i)) - 1 \right)^2 \right. \]
\[ - 2 \sum_{a_i} \pi_i(t)(a_i|s) \left( \exp(\eta f_i(t)(s,a_i)) - 1 \right) \left( Z_i^{t,s} - 1 \right) + \left. \sum_{a_i} \pi_i(t)(a_i|s) \left( Z_i^{t,s} - 1 \right)^2 \right] \]
\[ = \frac{1}{(Z_i^{t,s})^2} \left[ \sum_{a_i} \pi_i(t)(a_i|s) \left( \exp(\eta f_i(t)(s,a_i)) - 1 \right)^2 - \left( Z_i^{t,s} - 1 \right)^2 \right] \]

Since \(|e^x - 1| \geq \frac{|x|^2}{2} \) for \( x \geq -1 \), we have that
\[ \sum_{a_i} \pi_i(t)(a_i|s) \left( \exp(\eta f_i(t)(s,a_i)) - 1 \right)^2 \geq \sum_{a_i} \pi_i(t)(a_i|s) \left( \frac{\eta}{2} f_i(t)(s,a_i) \right)^2 \]
\[ \geq \frac{\eta^2}{4} \sum_{a_i} \pi_i(t)(a_i|s)f_i(t)(s,a_i)^2 \]

Additionally, as is proved in Lemma E.3
\[ Z_i^{t,s} = \sum_{a_i} \pi_i(t)(a_i|s) \exp(\eta f_i(t)(s,a_i)) \]
\[ \leq \sum_{a_i} \pi_i(t)(a_i|s) \left( 1 + (\eta f_i(t)(s,a_i)) + (\eta f_i(t)(s,a_i))^2 \right) \]
\[ \leq 1 + \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2. \]

Thus
\[
\sum_{a_i} \pi_i^{(t)}(a_i | s) \Delta_i^{(t)}(s, a_i)^2 = \frac{1}{\left( \frac{\eta^2}{4} \right)^2} \left[ \sum_{a_i} \pi_i^{(t)}(a_i | s) \left( \exp \left( \eta f_i^{(t)}(s, a_i) \right) - 1 \right)^2 - \left( Z_i^{t,s} - 1 \right)^2 \right] 
\geq \frac{1}{\left( \frac{\eta^2}{4} \right)^2} \left[ \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 - \left( \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right)^2 \right] 
= \frac{1}{\left( \frac{\eta^2}{4} \right)^2} \left( \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right) \left( \frac{1}{4} - \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right). 
\]

From Corollary E.4
\[- \frac{1}{15} \leq \eta f_i^{(t)}(s, a_i) \leq \frac{1}{3} \]
\[ \implies \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \leq \frac{1}{9}. \]

Thus
\[
\sum_{a_i} \pi_i^{(t)}(a_i | s) \Delta_i^{(t)}(s, a_i)^2 \frac{1}{\left( \frac{\eta^2}{4} \right)^2} \left( \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right) \left( \frac{1}{4} - \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right) 
\geq \frac{1}{\left( \frac{\eta^2}{4} \right)^2} \left( \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right) \left( \frac{1}{4} - \frac{1}{9} \right) 
\geq \left( \frac{9}{10} \right)^2 \left( \frac{1}{4} - \frac{1}{9} \right) \left( \eta^2 \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \right) 
\geq \frac{\eta^2}{9} \sum_{a_i} \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2, 
\]

Thus
\[
\sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi_i^{(t)}(a_i | s) \Delta_i^{(t)}(s, a_i)^2 \geq \frac{\eta^2}{9} \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 
\]
which completes the proof.

Lemma E.7.
\[
\text{NE-gap}(\theta^{(t)}) \leq \sum_{i=1}^{n} \sum_{s,a_i} d^{(t)}(s) \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 \leq \lambda \max_i |A_i| M, 
\]
where \( M = \sup_{\pi} \max_s \frac{1}{\pi_i^{(t)}(s | A_i)}. \)

Proof. We will now prove the lemma.
\[
d^{(t)}(s) \pi_i^{(t)}(a_i | s) f_i^{(t)}(s, a_i)^2 = d^{(t)}(s) \pi_i^{(t)}(a_i | s) \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) + \lambda \frac{1}{d^{(t)}(s) \pi_i^{(t)}(a_i | s) - \frac{\lambda |A_i|}{d^{(t)}(s)}} \right)^2 
= d^{(t)}(s) \pi_i^{(t)}(a_i | s) \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) - \frac{\lambda |A_i|}{d^{(t)}(s) \pi_i^{(t)}(a_i | s)} \right)^2 + \frac{\lambda^2}{d^{(t)}(s) \pi_i^{(t)}(a_i | s)} + 2\lambda \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) - \frac{\lambda |A_i|}{d^{(t)}(s) \pi_i^{(t)}(a_i | s)} \right) 
\geq 4\lambda \left( \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) - \frac{\lambda |A_i|}{d^{(t)}(s) \pi_i^{(t)}(a_i | s)} \right). 
\]
we have
\[ \frac{1}{1 - \gamma} A_i^{(t)}(s, a_i) \leq \frac{\sum d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2}{4\lambda} + \lambda |A_i| \sum_{a_i} d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2 + \lambda \max_i |A_i| M. \]

Thus from Lemma [D.7]
\[ \text{NE-gap}_i(\theta^{(t)}) \leq \frac{1}{1 - \gamma} \max_{s, a_i} A_i^{(t)}(s, a_i) \leq \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2 + \lambda \max_i |A_i| M, \]
which completes the proof.

We are now ready to prove Theorem 4.3.

**Theorem E.8.** (Theorem 4.3 restated) Under Assumption 2.3 and 2.4, for
\[ \eta \leq \min \left\{ \frac{1}{15(\frac{1}{1 - \gamma} + \lambda |A_i|M)} \eta T, \frac{1}{4(4\lambda \max_i |A_i| M^2 + \frac{M}{|A_i|} \gamma^2 + \frac{2M}{|A_i| \gamma^2})} \right\}, \]
running NPG scheme (10) will guarantee that
\[ \frac{\sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)})}{T} \leq \frac{9}{\eta \lambda T} \left( \Phi(\theta^{(T)}) - \Phi(\theta^{(0)}) \right) + \lambda \max_i |A_i| M, \]
where \( M \) is defined as in Assumption 2.3. Further, by setting \( \lambda = \frac{\eta T}{\max_i |A_i| M} \), \( \theta^{(0)} = \mathbf{0} \), then for any
\[ T \geq \frac{36 \max_i |A_i|}{\eta \epsilon^2} \left( \Phi(\theta^{(T)}) - \Phi(\theta^{(0)}) \right) M^2 \]
\[ \sim O \left( \frac{n \max_i |A_i| (\phi_{\max} - \phi_{\min}) M^2}{(1 - \gamma)^2 \epsilon^2} \right), \]
we have
\[ \frac{\sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)})}{T} \leq \epsilon. \]

**Proof.** From Lemma E.5 we have that for \( \eta \leq \min \left\{ \frac{1}{15(\frac{1}{1 - \gamma} + \lambda |A_i|M)} \eta T, \frac{1}{4(4\lambda \max_i |A_i| M^2 + \frac{M}{|A_i|} \gamma^2 + \frac{2M}{|A_i| \gamma^2})} \right\} \),
\[ \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \frac{1}{4\eta} \sum_{i=1}^{n} \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i)^2. \]

From Lemma E.6
\[ \Phi(\theta^{(t+1)}) - \Phi(\theta^{(t)}) \geq \frac{1}{4\eta} \sum_{i=1}^{n} \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) \Delta_i^{(t)}(s, a_i)^2 \]
\[ \geq \eta \frac{1}{36} \sum_{i=1}^{n} \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2 \]
Thus by telescoping we have
\[ \frac{\sum_{t=0}^{T-1} \sum_{i=1}^{n} \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2}{T} \leq \frac{36 \left( \Phi(\theta^{(T)}) - \Phi(\theta^{(0)}) \right)}{\eta T}. \]

From Lemma E.7
\[ \frac{\sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)})}{T} \leq \frac{1}{4\lambda} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \sum_{s, a_i} d_i(t) \pi_i^{(t)}(a_i|s) f_i^{(t)}(s, a_i)^2 + \lambda \max_i |A_i| M \]
\[ \leq \frac{9 \left( \Phi(\theta^{(T)}) - \Phi(\theta^{(0)}) \right)}{\eta \lambda T} + \lambda \max_i |A_i| M. \]
Specifically, set $\lambda = \frac{\max_i |A_i| M}{2 \lambda}$ and $\theta^{(0)} = 0$, then for any $T \geq \frac{18(\phi_{\max} - \phi_{\min})M}{(1 - \gamma)\eta \lambda}$

$$T \geq \frac{36 \max_i |A_i| (\phi_{\max} - \phi_{\min}) M^2}{(1 - \gamma)\eta \epsilon}$$

we have

$$\sum_{t=0}^{T-1} \text{NE-gap}(\theta^{(t)}) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which completes the proof.

### F Smoothness Proofs

**Lemma F.1.**

$$\|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_2 \leq \frac{6n}{(1 - \gamma)^3} \|\theta' - \theta\|_2$$

**Proof.** From Lemma F.3 we have that

$$\|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_2^2 = \sum_{i=1}^{n} \|\nabla_{\theta_i} \Phi(\theta') - \nabla_{\theta_i} \Phi(\theta)\|_2^2$$

$$\leq \sum_{i=1}^{n} \left( \frac{6}{(1 - \gamma)^3} \sum_{i=1}^{n} \|\theta'_i - \theta_i\|_2 \right)^2$$

$$= \frac{36n}{(1 - \gamma)^6} \left( \sum_{i=1}^{n} \|\theta'_i - \theta_i\|_2 \right)^2$$

$$\leq \frac{36n^2}{(1 - \gamma)^6} \sum_{i=1}^{n} \|\theta'_i - \theta_i\|^2_2$$

$$= \frac{36n^2}{(1 - \gamma)^6} \|\theta' - \theta\|_2^2,$$

thus

$$\|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_2 \leq \frac{6n}{(1 - \gamma)^3} \|\theta' - \theta\|_2$$

\[\square\]

**Lemma F.2.**

$$\|\nabla_\theta \tilde{\Phi}(\theta') - \nabla_\theta \tilde{\Phi}(\theta)\|_2 \leq \left( \frac{6n}{(1 - \gamma)^3} + 2\lambda \max_i |A_i| \right) \|\theta' - \theta\|_2$$

**Proof.** Since

$$\frac{\partial}{\partial \theta_{s,a_i}} \sum_{i=1}^{n} \sum_{s,a_i} \log \pi_{\theta_i}(a_i|s) = 1 - |A_i| \pi_{\theta_i}(a_i|s)$$

we have that

$$\|\nabla_\theta \left( \sum_{i=1}^{n} \sum_{s,a_i} \log \pi_{\theta'_i}(a_i|s) - \sum_{i=1}^{n} \sum_{s,a_i} \log \pi_{\theta_i}(a_i|s) \right)\|_2^2 = n |A_i|^2 \sum_s \sum_{a_i} \left( \pi_{\theta'_i}(a_i|s) - \pi_{\theta_i}(a_i|s) \right)^2$$

38
\[ \leq \sum_{i=1}^{n} |A_i|^2 \sum_{s} ||\pi_{\theta',s} - \pi_{\theta,s}||^2 \]
\[ \leq 4 \sum_{i=1}^{n} |A_i|^2 \sum_{s} ||\theta',s - \theta,s||^2 \quad \text{(Corollary G.6)} \]
\[ \leq 4 \max_{i} |A_i|^2 \|\theta' - \theta\|^2 \]

Thus
\[ \left\| \nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta) \right\|_2 \leq \left\| \nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta) \right\|_2 + \lambda \left\| \nabla_\theta \left( \sum_{i=1}^{n} \sum_{s,a} \log \pi_{\theta',(a_i|s)} - \sum_{i=1}^{n} \sum_{s,a} \log \pi_{\theta,(a_i|s)} \right) \right\|_2 \]
\[ \leq \left( \frac{6n}{(1 - \gamma)^3} + 2\lambda \max_{i} |A_i| \right) \|\theta' - \theta\|_2. \]

Lemma F.3.
\[ \|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_1 \leq \frac{6}{(1 - \gamma)^3} \sum_{i=1}^{n} \|\theta'_i - \theta_i\|_2 \]

Proof.
\[ \|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_1 = \frac{1}{1 - \gamma} \sum_{s,a} \left| d_{\theta'}(s)\pi_{\theta',(a_i|s)}A_{\theta'}^D(s,a_i) - d_{\theta}(s)\pi_{\theta,(a_i|s)}A_{\theta}^D(s,a_i) \right| \]
\[ \leq \frac{1}{1 - \gamma} \sum_{s,a} \left| d_{\theta'}(s)\pi_{\theta',(a_i|s)}\sum_{a_{-i}}\pi_{\theta',(a_{-i}|s)}A_{\theta'}^D(s,a_{-i}) - d_{\theta}(s)\pi_{\theta,(a_i|s)}\sum_{a_{-i}}\pi_{\theta,(a_{-i}|s)}A_{\theta}^D(s,a_{-i}) \right| \]
\[ \leq \frac{1}{1 - \gamma} \left( \sum_{s,a} \left| \sum_{a_{-i}} \left[ d_{\theta'}(s)\pi_{\theta',(a|s)} - d_{\theta}(s)\pi_{\theta,(a|s)} \right] A_{\theta'}^D(s,a) \right| + \sum_{s,a} \left| d_{\theta}(s)\pi_{\theta,(a|s)} A_{\theta}^D(s,a) \right| \right) \]
\[ \leq \frac{1}{1 - \gamma} \left( \frac{1}{1 - \gamma} \sum_{s,a} \left| \sum_{a_{-i}} \left[ d_{\theta'}(s)\pi_{\theta',(a|s)} - d_{\theta}(s)\pi_{\theta,(a|s)} \right] A_{\theta'}^D(s,a) \right| + \left( \sum_{s,a} \left| A_{\theta'}^D(s,a) - A_{\theta}^D(s,a) \right| \right) \right) \]

From Lemma G.1 and Corollary G.3, we have that
\[ \|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_1 \leq \frac{3}{(1 - \gamma)^3} \max_{s} \left| \pi_{\theta',s} - \pi_{\theta,s} \right|_1 \]
\[ \leq \frac{3}{(1 - \gamma)^3} \max_{s} \sum_{s} \left| \pi_{\theta',s} - \pi_{\theta,s} \right|_1 \]

From Corollary G.6, we have that
\[ \|\nabla_\theta \Phi(\theta') - \nabla_\theta \Phi(\theta)\|_1 \leq \frac{3}{(1 - \gamma)^3} \max_{s} \sum_{s} \left| \pi_{\theta',s} - \pi_{\theta,s} \right|_1 \]
\[ \leq \frac{6}{(1 - \gamma)^3} \sum_{i=1}^{n} \|\theta_i' - \theta_i\|_2. \]

G Some Useful Lemmas

Lemma G.1.
\[ \left| Q^\theta(s,a) - Q'^\theta(s,a) \right| \leq \frac{1}{(1 - \gamma)^2} \max_{s} \left| \pi_{\theta,s} - \pi_{\theta,s} \right|_1 \]
and thus
\[
\| V^{\theta'}(s) - V^{\theta}(s) \| \leq \frac{1}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1,
\]

and thus
\[
\| A^{\theta'}(s, a) - A^{\theta}(s, a) \| \leq \frac{2}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]

Proof. From performance difference lemma we have that
\[
| Q^{\theta'}(s, a) - Q^{\theta}(s, a) | = \sum_{t=1}^{\infty} \sum_{s'} \gamma^t \sum_{s'} \Pr^{\theta'}(s(t) = s'|s(0) = s, a(0) = a) \sum_{a'} (\pi_{\theta'}(a'|s') - \pi_{\theta}(a'|s')) Q^{\theta}(s', a')
\]
\[
\leq \sum_{t=1}^{\infty} \sum_{s'} \gamma^t \sum_{s'} \Pr^{\theta'}(s(t) = s'|s(0) = s, a(0) = a) \sum_{a'} |\pi_{\theta'}(a'|s') - \pi_{\theta}(a'|s')| | Q^{\theta}(s', a') |
\]
\[
\leq \sum_{t=1}^{\infty} \gamma^t \max_{s'} \sum_{a'} |\pi_{\theta'}(a'|s') - \pi_{\theta}(a'|s')| \frac{1}{1 - \gamma} = \frac{1}{(1 - \gamma)^2} \max_s \sum_{a} |\pi_{\theta'}(a|s) - \pi_{\theta}(a|s)|
\]
\[
= \frac{1}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]

Same argument also holds for \( | V^{\theta'}(s) - V^{\theta}(s) | \), and thus
\[
\| A^{\theta'}(s, a) - A^{\theta}(s, a) \| \leq | Q^{\theta'}(s, a) - Q^{\theta}(s, a) | + | V^{\theta'}(s) - V^{\theta}(s) | \leq \frac{2}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]

Lemma G.2.
\[
\frac{1}{1 - \gamma} \sum_{s, a} (d_{\theta'}(s) \pi_{\theta'}(a|s) - d_{\theta}(s) \pi_{\theta}(a|s)) r(s, a) \leq \frac{1}{(1 - \gamma)^2} \| r \|_\infty \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1,
\]

where \( \| r \|_\infty = \max_{s, a} | r(s, a) | \).

Proof. For any reward function \( r(s, a) \), we can define its value function \( V^{\theta}(s) \) and \( Q^{\theta}(s, a) \) correspondingly. Using performance difference lemma we have that
\[
\frac{1}{1 - \gamma} \sum_{s, a} (d_{\theta'}(s) \pi_{\theta'}(a|s) - d_{\theta}(s) \pi_{\theta}(a|s)) r(s, a) = \sum_{s} \rho(s) (V^{\theta'}(s) - V^{\theta}(s))
\]
\[
= \frac{1}{1 - \gamma} \sum_{s} d_{\theta'}(s) \pi_{\theta'}(a|s) - \pi_{\theta}(a|s) Q^{\theta}(s, a)
\]
\[
\leq \frac{1}{(1 - \gamma)^2} \| r \|_\infty \sum_{s} d_{\theta'}(s) \| \pi_{\theta'} - \pi_{\theta} \|_1
\]
\[
\leq \frac{1}{(1 - \gamma)^2} \| r \|_\infty \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]

We have the following two corollaries for Lemma G.2.

Corollary G.3.
\[
\frac{1}{1 - \gamma} \sum_{s} | d_{\theta'}(s) - d_{\theta}(s) | \leq \frac{1}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]

Proof.
\[
\frac{1}{1 - \gamma} \sum_{s} | d_{\theta'}(s) - d_{\theta}(s) | = \max_{-1 \leq r(s) \leq 1} \frac{1}{1 - \gamma} \sum_{s, a} (d_{\theta'}(s) \pi_{\theta'}(a|s) - d_{\theta}(s) \pi_{\theta}(a|s)) r(s)
\]
\[
\leq \frac{1}{(1 - \gamma)^2} \max_s \| \pi_{\theta'} - \pi_{\theta} \|_1.
\]
Corollary G.4.

\[ \frac{1}{1 - \gamma} \sum_s |d_{\psi'}(s)\pi_{\psi'}(a|s) - d_{\phi}(s)\pi_{\phi}(a|s)| \leq \frac{1}{(1 - \gamma)^2} \max_s \|\pi_{\psi'} - \pi_{\phi}\|_1. \]

Proof.

\[ \frac{1}{1 - \gamma} \sum_s |d_{\psi'}(s)\pi_{\psi'}(a|s) - d_{\phi}(s)\pi_{\phi}(a|s)| = \max_{-1 \leq r(s,a) \leq 1} \frac{1}{1 - \gamma} \sum_{s,a} (d_{\psi'}(s)\pi_{\psi'}(a|s) - d_{\phi}(s)\pi_{\phi}(a|s))r(s,a) \]

\[ \leq \frac{1}{(1 - \gamma)^2} \max_s \|\pi_{\psi'} - \pi_{\phi}\|_1. \]

Lemma G.5.

\[ \sum_{a_i} (\pi_{\psi'}(a_i|s) - \pi_{\phi}(a_i|s)) f(a_i) \leq 2\|f\|_\infty \|\theta_{i,s}' - \theta_{i,s}\|_2 \]

Proof. It suffices to show that

\[ \left\| \nabla_{\theta_{i,s}} \sum_{a_i} \pi_{\phi}(a_i|s)f(a_i) \right\|_2 \leq 2\|f\|_\infty, \forall \theta, \]

then by Lagrange mean value theorem,

\[ \sum_{a_i} (\pi_{\psi'}(a_i|s) - \pi_{\phi}(a_i|s)) f(a_i) \leq \max_{t,\theta = \theta + (1 - t)\theta'} \left\| \nabla_{\theta_{i,s}} \sum_{a_i} \pi_{\phi}(a_i|s)f(a_i) \right\|_2 \|\theta_{i,s}' - \theta_{i,s}\|_2 \leq 2\|f\|_\infty \|\theta_{i,s}' - \theta_{i,s}\|_2. \]

Since

\[ \frac{\partial}{\partial \theta_{a_i,s}} \pi_{\phi}(a_i|s)f(a_i) = \pi_{\phi}(a_i|s)(f(a_i) - \bar{f}), \text{ where } \bar{f} = \sum_{a_i} \pi_{\phi}(a_i|s)f(a_i), \]

we have

\[ \left\| \nabla_{\theta_{i,s}} \sum_{a_i} \pi_{\phi}(a_i|s)f(a_i) \right\|_2^2 \leq \sum_{a_i} \pi_{\phi}(a_i|s)^2 (f(a_i) - \bar{f})^2 \leq \sum_{a_i} \pi_{\phi}(a_i|s)^2 (2\|f\|_\infty)^2 \leq 4\|f\|_\infty^2, \]

which completes the proof.

Corollary G.6. (of Lemma G.5)

\[ \|\pi_{\psi'}(a_i|s) - \pi_{\phi}(a_i|s)\|_1 \leq 2\|\theta_{i,s}' - \theta_{i,s}\|_2 \leq 2\|\theta_i' - \theta_i\|_2 \]

Proof.

\[ \|\pi_{\psi'}(a_i|s) - \pi_{\phi}(a_i|s)\|_1 = \max_{f: \|f\|_\infty \leq 1} \sum_{a_i} (\pi_{\psi'}(a_i|s) - \pi_{\phi}(a_i|s)) f(a_i) \leq 2\|\theta_{i,s}' - \theta_{i,s}\|_2. \]

\[ 41 \]