Learning Equilibria in Adversarial Team Markov Games: A Nonconvex-Hidden-Concave Min-Max Optimization Problem

Fivos Kalogiannis[⋆]
University of California, Irvine
Archimedes/Athena RC, Greece

Jingming Yan[⋆]
University of California, Irvine

Ioannis Panageas
University of California, Irvine
Archimedes/Athena RC, Greece

Abstract

We study the problem of learning a Nash equilibrium (NE) in Markov games which is a cornerstone in multi-agent reinforcement learning (MARL). In particular, we focus on infinite-horizon adversarial team Markov games (ATMGs) in which agents that share a common reward function compete against a single opponent, *the adversary*. These games unify two-player zero-sum Markov games and Markov potential games, resulting in a setting that encompasses both collaboration and competition. [65] provided an efficient equilibrium computation algorithm for ATMGs which presumes knowledge of the reward and transition functions and has no sample complexity guarantees. We contribute a learning algorithm that utilizes MARL policy gradient methods with iteration and sample complexity that is polynomial in the approximation error ϵ and the natural parameters of the ATMG, resolving the main caveats of the solution by [65]. It is worth noting that previously, the existence of learning algorithms for NE was known for Markov two-player zero-sum and potential games but not for ATMGs.

Seen through the lens of min-max optimization, computing a NE in these games consists a nonconvex–nonconcave saddle-point problem. Min-max optimization has received extensive study. Nevertheless, the case of nonconvex-nonconcave landscapes remains elusive: in full generality, finding saddle-points is computationally intractable [33]. We circumvent the aforementioned intractability by developing techniques that exploit the hidden structure of the objective function via a nonconvex–concave reformulation. However, this introduces the challenge of a feasibility set with coupled constraints. We tackle these challenges by establishing novel techniques for optimizing weakly-smooth nonconvex functions, extending the framework of [35].

1 Introduction

Multi-agent reinforcement learning (MARL) investigates behaviors of multiple interacting agents within a dynamic, shared environment where the actions of each agent not only impact their individual rewards but also the overall state of the system. MARL has introduced several practical techniques that have justifiably captured public interest in recent years, particularly in skill-intensive games like starcraft, go, chess, and poker [12, 97, 101, 79, 16, 15, 14, 86], where its empirical methods have

[⋆]Equal Contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2948

achieved super-human performance. More recently, MARL methods combined with large language models has excelled in the game of Diplomacy [6]. Despite these practical achievements, theoretical understanding of MARL has lagged behind its empirical successes.

Markov games (MGs) [95] is a rigorous and versatile mathematical structure that MARL employs to systematically formalize the strategic interactions in the dynamic settings [71]. These games extend Markov decision processes (MDPs) [88] to multiple agents, each making decisions and receiving rewards independently as the environment evolves. The joint decisions of the agents influence both individual rewards and the transition of the environment. MARL in general is occupied with leading the multi-agent system to a favorable outcome. Through the lens of game theory, the notion of a "favorable outcome" is formally defined through concepts like a Nash equilibrium and a (coarse) correlated equilibrium. Although computing Nash equilibria is generally computationally intractable—even in two-player games without states [28, 24]—it becomes tractable in fully cooperative settings like Markov potential games [114, 68] and is also tractable in competitive scenarios such as two-player zero-sum Markov games [27, 103, 17]. Recent advances [65] also show computational tractability in adversarial team Markov games (ATMGs)—a context that combines both cooperative and competitive dynamics among agents. More specifically, an infinite-horizon adversarial team Markov game (ATMG) is a Markov Game in which n team players, compete against one adversary. Each of the team players receives the same reward and is equal to minus the reward of the adversary. ATMGs generalize both Markov zero-sum and potential games; the former can be viewed as ATMGs with $n = 1$, the latter by choosing the adversary to be dummy (having one action).

Nash equilibrium computation in ATMGs naturally leads to a min-max optimization problem. Min-max optimization has been deeply explored across game theory, optimization, and machine learning. The past decade it has witnessed a proliferation of min-max optimization applications, notably in areas like generative adversarial networks (GANs) [49], robust machine learning [73], and adversarial training [50]. In these applications, the optimization objectives often involve nonconvex–nonconcave functions which pose substantial challenges. Typically, the aim is to approximate saddle-points of $f(\boldsymbol{x}, \boldsymbol{y})$. In normal form games, these points correspond to Nash equilibria. This correspondence also holds true for MGs due to the gradient domination property [3]. Although we cannot aspire to cover the vast quantity of works in MARL and optimization, we select some representative works that we defer to Appendix A due to space constraints.

This paper aims to develop learning methods to approximate Nash equilibria in team Markov games by using only individual rewards and state observations as feedback, addressing the following question and answering one of the main caveats of the solutions provided in [65]:

Is it possible for agents to efficiently learn Nash equilibria in adversarial team Markov games, having only access to trajectory roll-out samples and (almost**) no communication, i.e., independently?* ($\star$)

1.1 Our Contributions

Let us provide some context before stating our main results. An infinite-horizon adversarial team Markov game (ATMG) is characterized by a finite state-space $\mathcal{S}$, n team players, each equipped with a finite action-space $\mathcal{A}_i,\ i \in \{1, \ldots, n\}$, and one adversary with a finite action-space $\mathcal{B}$. Each of the team players receives the same reward which is equal to minus the reward of the adversary. The adversary's value function is defined as the discounted expected sum of their rewards, where the discount factor is $\gamma \in [0, 1)$. An approximate Nash equilibrium is a product distribution over policy space such that no agent can improve their value by unilaterally deviating. We propose a learning algorithm that has both iteration and sample complexity polynomial in the parameters of the Markov Game and returns approximate Nash equilibria.

Theorem 1.1 (Informal Version of Theorem 3.3)**.** *There is a learning algorithm (*ISPNG*) that uses bandit feedback and guarantees convergence to an ϵ-approximate Nash equilibrium in adversarial*

*We say "almost" as the agents need to take turns in updating their policies instead of making updates simultaneously. Nevertheless, the learning dynamics remain uncoupled.

team Markov games, the sample and iteration complexities of which are

$$\mathsf{poly}\left(\frac{1}{\epsilon}, |\mathcal{S}|, \sum_{k=1}^{n} |\mathcal{A}_i| + |\mathcal{B}|, \frac{1}{1-\gamma}\right).$$

We deem noteworthy that our algorithm manages to compute a Nash equilibrium in a Markov game, which combines opposing and shared agent interests, by only using a number of iterations and samples that is polynomial in the approximation error and the description of the game. Further, it manages to beat the *curse of multi-agents* [62]—*i.e.,* its iteration and sample complexity depends on $\sum_{k=1}^{n} |\mathcal{A}_i|$ instead of $\prod_{k=1}^{n} |\mathcal{A}_i|$.

In order to achieve the latter contribution, we acquired convergence guarantees for stochastic projected gradient descent in nonconvex functions when the gradient is Hölder-continuous—a notion of continuity weaker than that of Lipschitz. Finally, we contribute a general result that guarantees convergence to a saddle-point in functions that are nonconvex–hidden-strongly-concave.

1.2 Technical Overview

The problem of computing an approximate Nash equilibrium in an adversarial team Markov game boils down to computing an approximate saddle-point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of the adversary's value function $V(\boldsymbol{x}, \boldsymbol{y})$; see Definition 2.3. The variables $\boldsymbol{x}$ denote the policies of the team, each member of which aims to individually minimize V. Moreover, $\boldsymbol{y}$ denotes the policy of the adversary who aims to maximize V. The equivalence between saddle-points and equilibria is due to (i) the game being *zero-sum* between the team and the adversary and (ii) the *gradient domination property* (see Lemma C.7) that holds per player, and has already been established in prior works [3, 68, 114]. In words, gradient domination in our setting implies that any approximate first-order stationary policy is also an approximate best response for that player.

The problem of computing an approximate saddle-point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of the objective $V(\boldsymbol{x}, \boldsymbol{y})$ poses computational challenges due to its nonconvex–nonconcave nature. Previous work [65] showed that one can compute an approximate saddle-point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of V, by first obtaining an approximate stationary point $\boldsymbol{x}^*$ of $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}} V(\boldsymbol{x}, \boldsymbol{y})$ through a Moreau envelope argument and then extending it to $(\boldsymbol{x}^*, \boldsymbol{y}^*)$. The proof of extendibility uses involved arguments that utilize the Lagrange multipliers of a carefully chosen nonlinear program (for the stationary point $\boldsymbol{x}^*$), while the computation of $\boldsymbol{y}^*$ requires solving another linear program. It is worth noting that the aforementioned linear program presumes access to the full description of the reward function and the transition model of the underlying Markov game when the team plays policy $\boldsymbol{x}^*$. This fact prevents the possibility of casting this approach into a learning algorithm.

Our proposed (learning) algorithm bypasses the requirement for knowledge of the reward function and the transition model, and works under the bandit feedback framework. The first idea behind our algorithm is to consider the adversary's value function as a function F of the *adversary's* state-action visitation measure $\boldsymbol{\lambda}$, $F(\boldsymbol{x}, \boldsymbol{\lambda}) \coloneqq V(\boldsymbol{x}, \boldsymbol{y})$, and the addition of a regularizing term $-\frac{\nu}{2}\|\boldsymbol{\lambda}\|^2$ (ν can be thought of as a small positive scalar). As a result, the max function of the regularized value function, $\Phi^{\nu}(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x})} \left\{ F(\boldsymbol{x}, \boldsymbol{\lambda}) - \frac{\nu}{2}\|\boldsymbol{\lambda}\|^2 \right\}$, is differentiable, where $\Lambda(\boldsymbol{x}) \subseteq \Delta^{|\mathcal{S}||\mathcal{B}|}$ denotes the feasibility set of $\boldsymbol{\lambda}$ and depends on $\boldsymbol{x}$. Effectively, different policies, $\boldsymbol{x}$, for the team induce a different single agent Markov decision process for the adversary. The addition of the regularizer allows us to apply Danskin's theorem on a function with a unique maximizer circumventing the necessity of solving a linear program; one only needs to approach that unique solution. To the best of our knowledge, this is the first work introducing a function of $\boldsymbol{\lambda}$ as a regularizing term.

By reformulating the regularized value function using state-action visitation measure $\boldsymbol{\lambda}$, the problem boils down to learning an approximate saddle-point of a nonconvex–strongly-concave function with *coupled constraints*. Coupled constraints are a type of constraints that cannot be expressed as a Cartesian product (the main well-studied setting in min-max optimization [63]), *i.e.*, the feasibility set $\Lambda(\boldsymbol{x})$, depends on $\boldsymbol{x}$. The first challenge towards handling the coupled constraints is to argue that $\nabla\Phi^{\nu}$ is Hölder-continuous which is a notion of continuity weaker than Lipschitz continuity (see Definition 2.1). Specifically, in Theorem 3.2, we show that $\Phi^{\nu}(\boldsymbol{x})$ is weakly-smooth, or equivalently, $\nabla\Phi^{\nu}$ is Hölder-continuous. It seems unlikely that we could use Moreau envelope techniques to prove convergence of stochastic projected gradient descent on a weakly-smooth function. The next

https://doi.org/10.52202/079017-2948

step of our proof is to transfer the weakly-smooth nonconvex optimization problem into a smooth optimization problem with inexact gradient oracles, extending the techniques from [35] to nonconvex and constrained settings. Since we only allow each player to observe the reward they received and not the action chosen by the other players (including the adversary), one last challenge we have to deal with is the inability to estimate the state-action visitation measure $\boldsymbol{\lambda}$ of the adversary, making the gradient inexact when computing $\nabla\Phi^{\nu}(x)$ in both deterministic and stochastic settings.

2 Preliminaries

Starting, we will introduce the notation conventions we use and split the rest of the preliminaries into two subsections. Section 2.1 provides necessary definitions whereas Section 2.2 deals with the preliminaries of (adversarial team) Markov games and the notion of Nash equilibrium.

Notation. We denote $[n] \coloneqq \{1, \dots, n\}$. We use superscripts to denote the (discrete) time index, and subscripts to index the players. We use boldface for vectors and matrices; scalars will be denoted by lightface variables. We define $\|\cdot\|_2, \|\cdot\|_1, \|\cdot\|_\infty$ to be the ℓ_2-norm, the ℓ_1-norm and the ℓ_∞ norm respectively. The simplex of probability vectors supported on a finite set $\mathcal{A}$ is noted as $\Delta(\mathcal{A})$. Unless specified otherwise, we denote $\|\cdot\|_2$ by $\|\cdot\|$. $\mathrm{Diam}_{\mathcal{X}}$ denotes the diameter of a compact set $\mathcal{X}$ in ℓ_2-distance. For simplicity in the exposition, we may sometimes use the $O(\cdot)$ notation to suppress dependencies that are polynomial in the natural parameters of the problem and $\tilde{O}(\cdot)$ to further hide logarithmic factors; precise statements are given in the Appendix. For the convenience of the reader, a comprehensive overview of our notation is given in Table 1.

2.1 Basic Definitions and Facts

We commence this subsection by introducing a number of concepts and statements of mathematical analysis and optimization. We define Hölder continuity and the notion of a stationary point in constrained minimization and min-max optimization.

The notion of Hölder continuity of the gradient is a weaker notion of Lipschitz gradient continuity.

Definition 2.1 (p-Hölder continuous gradient)**.** *A function $\phi : \mathbb{R}^d \to \mathbb{R}$ is said to have a (ℓ_p, p)-Hölder continuous gradient if for every $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, it holds that:*

$$\|\nabla\phi(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z}')\|_2 \leq \ell_p \|\boldsymbol{z} - \boldsymbol{z}'\|_2^p.$$

When $p = 1$, we retrieve the definition of an ℓ-smooth function.

Throughout, following standard conventions, we will refer to functions for which the gradient is p-Hölder continuous with a $p < 1$ as *weakly-smooth*. We state the notions of first-order stationarity relevant to our work.

Definition 2.2 (ϵ-FOSP)**.** *In the context of the constrained minimization problem $\min_{\boldsymbol{z}\in\mathcal{Z}} \phi(\boldsymbol{z})$, a point $\boldsymbol{z} \in \mathcal{Z}$ is said to be an ϵ-approximate stationary point if,*

$$\langle -\nabla_{\boldsymbol{z}}\phi(\boldsymbol{z}), \boldsymbol{z}' - \boldsymbol{z}\rangle \leq \epsilon, \quad \forall \boldsymbol{z}' \in \mathcal{Z}.$$

Similarly, we will define an ϵ-approximate saddle-point for the constrained min-max optimization problem $\min_{\mathcal{X}} \max_{\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$.

Definition 2.3 (ϵ-SP)**.** *Let a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. A point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ is said to be an ϵ-approximate saddle-point (or ϵ-FOSP for the min-max problem) if,*

$$-\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})^\top (\boldsymbol{x}' - \boldsymbol{x}) \leq \epsilon, \forall \boldsymbol{x}' \in \mathcal{X};$$
$$\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})^\top (\boldsymbol{y}' - \boldsymbol{y}) \leq \epsilon, \forall \boldsymbol{y}' \in \mathcal{Y}.$$

2.2 Adversarial Team Markov Games

An adversarial team Markov game is the Markov game extension of normal-form adversarial team games [98]. The game takes place in an infinite-horizon discounted setting where a team of identically-interested players compete against one adversarial player, the *adversary*. We can formally define an adversarial team Markov game as a tuple $\Gamma(\mathcal{S}, [n+1], \mathcal{A}, \mathcal{B}, r, \mathbb{P}, \gamma, \boldsymbol{\rho})$, where:

- $\mathcal{S}$ is the finite set of states, or *state-space*, with cardinality $S \coloneqq |\mathcal{S}|$;
- $[n+1]$ is the set of players, with the first n players belonging to the team and the last one being the adversary;
- $\mathcal{A} = \bigtimes_{i=1}^{n} \mathcal{A}_i$ is the finite set of the team's joint actions (or, team's *action-space*), while $\mathcal{A}_i$ is the i-th player's *action-space*; respectively $\mathcal{B}$ is the adversary's action-space; further, $A \coloneqq \max_{i\in[n]} |\mathcal{A}_i|$ and $B \coloneqq |\mathcal{B}|$;
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0,1]$ is the adversary's reward function;
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is transition probability function;
- $\gamma \in [0,1)$ is the discount factor;
- $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ is the initial state distribution. We assume that $\boldsymbol{\rho}$ is of full-support, $\rho(s) > 0, \forall s \in \mathcal{S}$.

Every team player $i \in [n]$ gets the same reward and the sum of team players' rewards are equal to the adversary's loss, *i.e.*, $\sum_{i=1}^{n} r_i(s, \boldsymbol{a}, b) = -r(s, \boldsymbol{a}, b)$.

2.2.1 Policies, Value Function, and Visitation Measures

In this part, we describe policy classes, the value function, and the state-action visitation measures. All of these notions are indispensable for our analysis.

Policy Definitions. For any agent i, a *stationary* policy $\boldsymbol{\pi}_i$ is defined as a mapping from any given state to a probability distribution over possible actions, where $\boldsymbol{\pi}_i : \mathcal{S} \ni s \mapsto \boldsymbol{\pi}_i(\cdot|s) \in \Delta(\mathcal{A}_i)$. A policy $\boldsymbol{\pi}_i$ is described as *deterministic* when, for any state, it selects a particular action with probability of 1. To simplify, we denote the policy spaces for the team and the adversary as $\Pi_{\text{team}} : \mathcal{S} \to \Delta(\mathcal{A})$ and $\Pi_{\text{adv}} : \mathcal{S} \to \Delta(\mathcal{B})$, respectively. Additionally, the combined policy space for all participants can be represented as $\Pi : \mathcal{S} \to \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$.

Direct Policy Parametrization. In the context of our work, we assume the strategy of *direct policy representation* for all players. Specifically, for each player i within the set $[n]$, the policy space $\mathcal{X}_i$ is defined as $\Delta(\mathcal{A}_i)^S$, with $\boldsymbol{\pi}_i = \boldsymbol{x}_i$, such that the probability of choosing action a in state s, $x_{i,s,a}$, equals $\boldsymbol{\pi}_i(a|s)$. By the usual game-theoretic convention, $\boldsymbol{\pi}_{-i}$ denotes the policy of all agents apart from i. For the adversary, $\mathcal{Y}$ is set as $\Delta(\mathcal{B})^S$, with $\boldsymbol{\pi}_{\text{adv}} = \boldsymbol{y}$, so that $y_{s,a} = \boldsymbol{\pi}_{\text{adv}}(a|s)$.

Having defined policies, we can introduce some standard shortcut notations such as $r(s, \boldsymbol{x}, \boldsymbol{y}) \coloneqq \mathbb{E}_{(\boldsymbol{a}\, b)\sim(\boldsymbol{x},\boldsymbol{y})}[r(s, \boldsymbol{a}, b)]$, and the vectors $\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}, \boldsymbol{r}(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{|\mathcal{S}|}$ with $\boldsymbol{r}(\boldsymbol{x}) \coloneqq [\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{x}}[r(s, \boldsymbol{a}, b)]]_{s,b}$ and $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \left[\mathbb{E}_{(\boldsymbol{a},b)\sim\boldsymbol{x}}[r(s, \boldsymbol{a}, b)]\right]_s$. Further, we define $\mathbb{P}(s'|s, \boldsymbol{x}, \boldsymbol{y})$ as $\mathbb{P}(s'|s, \boldsymbol{x}, \boldsymbol{y}) \coloneqq \mathbb{E}_{(\boldsymbol{a},b)\sim(\boldsymbol{x},\boldsymbol{y})}[\mathbb{P}(s'|s, \boldsymbol{a}, b)]$ and the vector $\mathbb{P}(s, \boldsymbol{x}, b) \in \Delta(\mathcal{S})$ with $\mathbb{P}(s, \boldsymbol{x}, \boldsymbol{y}) \coloneqq \left[\mathbb{E}_{(\boldsymbol{a},b)\sim(\boldsymbol{x},\boldsymbol{y})}[\mathbb{P}(s'|s, \boldsymbol{a}, b)]\right]_{s'}$.

The Value Function. The *value function* V_s, for a given state $s \in \mathcal{S}$, is defined as the adversary's expected total discounted reward over time under a combined policy $(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})$ from the policy space Π, with $\boldsymbol{x} = \boldsymbol{\pi}_{\text{team}}$ being the aggregation of policies $(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$. Formally, this is represented as

$$V_s(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, \boldsymbol{a}_h, b_h) \,\middle|\, s_0 = s\right],$$

where the expected value is calculated over the distribution of trajectories generated by the policies $\boldsymbol{x}$ and $\boldsymbol{y}$. If the initial state is instead sampled from a distribution $\boldsymbol{\rho}$, the value function is expressed as $V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{s\sim\boldsymbol{\rho}}\left[V_s(\boldsymbol{x}, \boldsymbol{y})\right]$.

Visitation Measures. The important quantity of state-action visitation measures, or the expected discounted sum of visitations of a state-action pair.

Definition 2.4 (State-Action Visit. Measure)**.** *For any initial distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$, transition matrix $\mathbb{P}$, a team policy $\boldsymbol{x}$, and a policy $\boldsymbol{y} \in \mathcal{Y}$, we define the station-action visitation measure of the adversary $\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})$ as follows:*

$$\lambda_{s,b}(\boldsymbol{y}; \boldsymbol{x}) \coloneqq \sum_{h=0}^{\infty} \gamma^h \, \mathbb{P}(s_h = s, b_h = b | \boldsymbol{x}, \boldsymbol{y}, s_0 \sim \boldsymbol{\rho}).$$

https://doi.org/10.52202/079017-2948

Where $\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x})$ denotes the $(s,b)^{th}$ entry of $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$.

As we will further discuss in the appendix (Appendix C.1), the correspondence between $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ is "1–1" for a fixed team policy $\boldsymbol{x}$. This property is crucial for our contributions.

Reformulation of the Value Function. A key property of the value function $V_{\boldsymbol{\rho}}$ is that it can be rewritten as a concave function of the state-action visitation measure:

$$V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{r}(\boldsymbol{x})^{\top}\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}).$$

Definition 2.5 (ϵ-NE)**.** *A product policy $\left(\boldsymbol{x}^*,\boldsymbol{y}^*\right) \in \mathcal{X}\times\mathcal{Y}$ is called an ϵ-approximate Nash equilibrium for an $\epsilon \geq 0$, when*

$$V_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^*) \leq V_{\boldsymbol{\rho}}((\boldsymbol{x}_i',\boldsymbol{x}_{-i}^*),\boldsymbol{y}^*) + \epsilon, \forall \boldsymbol{x}_i' \in \mathcal{X}_i, \forall i \in [n];$$

and

$$V_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^*) \geq V_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}') - \epsilon, \qquad \forall \boldsymbol{y}' \in \mathcal{Y}.$$

2.2.2 The Gradient and Visitation Measure Estimators.

An essential element that led to the development of policy gradient methods is the policy gradient theorem [104]. Notably, it has enabled the design of finite-sample gradient estimators. This technique fits well into the *MARL independent learning protocol* [27]. After all agents have proposed their policy, the MDP is run to acquire batches of trajectories from which all agents will observe the chain's state and their individual reward. These samples are utilized to estimate gradients.

The team agents implement a batch version of the REINFORCE estimator whose definition is deferred to the Appendix C.6.1. As for the estimators that the adversary utilizes, we define the state-action visitation measure estimator and their gradient estimator closely following [113].

Definition 2.6 (State-Action Visitation Measure Estimator)**.** *Let $\boldsymbol{e}_{s,b}$ be the standard basis for the $(s,b)^{th}$ entry. Let $\tau = (s_0,b_0,s_1,b_1,\cdots,s_{H-1},b_{H-1})$ denote a trajectory with length H sampled under initial distribution $\boldsymbol{\rho}$ and policy $\boldsymbol{y}$ We define the estimator for $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$ with the trajectory τ as the following*

$$\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y}) := \sum_{h=0}^{H-1} \gamma^h \cdot \boldsymbol{e}_{s_h,b_h}.$$

By applying policy gradient theorem [104] along with the chain-rule, the gradient estimator for a value-function that is nonlinear in $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$, is computed by the following estimator [113].

Definition 2.7 (Gradient Estimator)**.** *Let $\tau = (s_0,b_0,s_1,b_1,\cdots,s_{H-1},b_{H-1})$ denote a trajectory with length H sampled under initial distribution $\boldsymbol{\rho}$ and policy $\boldsymbol{y}$. Let $F(\boldsymbol{\lambda}(\boldsymbol{y}))$ be the value function of the MDP w.r.t. $\boldsymbol{\lambda}(\boldsymbol{y})$ and $\boldsymbol{u} := \nabla_{\lambda}F(\boldsymbol{\lambda}(\boldsymbol{y}))$. The estimator for gradient $\nabla_{\boldsymbol{y}}F(\boldsymbol{\lambda}(\boldsymbol{y}))$ using the sampled trajectory τ is defined as*

$$\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y};\boldsymbol{u}) := \sum_{h=0}^{H-1} \gamma^h \cdot \boldsymbol{u}(s_h,b_h) \cdot \left(\sum_{h'=0}^{h} \nabla_{\boldsymbol{y}} \log \boldsymbol{y}(b_{h'}|s_{h'})\right).$$

Sufficient Exploration. A standard, while rather naive, technique of bounding the variance of the REINFORCE gradient estimator is using ζ-greedy policy parametrization. Effectively, every action in a player's dispose is played with a probability of at least ζ. For our convenience, we ensure sufficient exploration by a ζ-*truncated simplex* approach. Moreover, for a given feasibility set $\mathcal{X}$, we denote $\mathcal{X}^{\zeta}$ to be the ζ-truncated feasibility set.

3 Main Results

We present our main results in two different subsections. In Section 3.1 we manage to attain guarantees for convergence to an approximate stationary-point to constrained nonconvex optimization with an stochastic inexact gradient oracle— we do so by extending previous results of [35]. While in Section 3.2, we apply the latter results along with RL techniques in order to design the first learning algorithm that computes a Nash equilibrium in ATMGs.

3.1 Stochastic Weakly-Smooth Nonconvex Optimization with Inexact Gradients

In this subsection we prove that projected gradient descent with a stohcastic inexact gradient oracle converges to an ϵ-FOSP in nonconvex functions with Hölder continuous gradients. We will use this key result in subsequent sections. We begin by defining the inexact gradient oracle and its stochastic version.

Definition 3.1 (Inexact Gradient Oracle)**.** *Let a differentiable function $\phi(\boldsymbol{z})$ and its gradient $\nabla\phi(\boldsymbol{z})$. We call the vector-valued function $\boldsymbol{g}(\boldsymbol{z})$ a ϑ-inexact gradient oracle if,*

$$\|\boldsymbol{g}(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z})\| \leq \vartheta, \quad \forall \boldsymbol{z}.$$

Further, given a random variable ξ in some sample space Ξ, we define a stochastic inexact gradient oracle $G : \mathcal{Z} \times \Xi \to \mathbb{R}^d$. We assume that the expected value of this oracle will be equal to a ϑ-inexact gradient oracle $\boldsymbol{g}(\boldsymbol{z})$. Additionally to being unbiased (with respect to a ϑ-inexact gradient oracle), we assume its variance to be bounded.

Assumption 3.1 (Unbiased and Bounded Variance)**.** For a variance parameter $\sigma^2 > 0$, the gradient oracle G, satisfies

$$\mathbb{E}_\xi[G(\boldsymbol{z}, \xi)] = \boldsymbol{g}(\boldsymbol{z}) \quad \text{and} \quad \mathbb{E}_\xi\left[\|G(\boldsymbol{z}, \xi) - \boldsymbol{g}(\boldsymbol{z})\|^2\right] \leq \sigma^2.$$

Following, we consider the simple update rule of *Mini-Batch Inexact Stochastic Projected Gradient Descent*, with a batch size $M > 0$ and $\hat{\boldsymbol{g}}^t = \frac{1}{M}\sum_{j=1}^{M} G\left(\boldsymbol{z}^t, \xi_j^t\right)$,

$$\boldsymbol{z}^{t+1} = \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^t - \eta\hat{\boldsymbol{g}}^t\right). \tag{Inexact Stoch-PGD}$$

We can now state our convergence Theorem for (Inexact Stoch-PGD) whose proof we defer to the appendix.

Theorem 3.1 (Convergence to ϵ-FOSP; Formally in Theorem B.1)**.** *Let $\phi : \mathcal{Z} \to \mathbb{R}$ be a Lipschitz continuous function with (ℓ_p, p)-Hölder continuous gradient and a desired accuracy ϵ. Also, let a stochastic inexact first-order oracle G satisfying Assumption 3.1. The update rule* (Inexact Stoch-PGD)*, with a step-size $\eta = \mathcal{O}\left(\epsilon^{\frac{1-p}{p}}\right)$, computes an ϵ-approximate stationary point after $T = O\left(\epsilon^{-\frac{1+p}{p}}\right)$ iterations.*

3.2 Learning Nash Equilibria in Adversarial Team Markov Games

In this subsection we state our contributed Algorithm 1, or ISPNG, which converges to an ϵ-NE for any ATMG, Γ, with an iteration and sample complexity that scales polynomially with $1/\epsilon$ and the parameters of Γ. To simply describe the algorithm, the team players initialize their policies and then the following two steps are repeated for T iterations:

1. the adversary approximately maximizes a *regularized version* of their value function, $V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x}) - \frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})\|^2$, using Algorithm 2, and then
2. every agent independently performs a gradient descent step on the value function.

During this process, all agents use only bandit feedback information in order to estimate the gradients of the value function. We remark that the learning dynamics remain uncoupled. The only instance of communication between agents is the fact that the team and the adversary take turns when updating their policies. During their turn, the adversary approximately best-responds.

Of particular interest is the sub-routine of Algorithm 2, VIS-REG-PG. It is effectively a directly parameterized policy gradient method for an objective function that is concave in the state-action visitation measure $\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{B}|}$. The objective function is merely the original value function plus a quadratic term, $-\frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})\|^2$. We remind the reader that due to the existence of this introduced regularizer, the utility of the adversary $\boldsymbol{u} = \nabla_{\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})} F_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})$. In order to estimate a gradient, the adversary needs to collect a number of trajectories, $\tau = (s_0, b_0, s_1, \ldots, s_{H-1}, b_{H-1}, s_H)$, each of length H. Notably, the adversary only uses the empirical state-action visitation measure for the purpose of gradient estimation of the regularized function.

https://doi.org/10.52202/079017-2948

Algorithm 1 Independent Stochastic Policy-Nested-Gradient (ISPNG)

Input: Accuracy $\epsilon > 0$

Based on ϵ, set stepsize η_x, T_x iterations, batch size M, truncation parameter ζ_x, and inner-loop accuracy $\epsilon_y > 0$. ▷ see Theorem C.3

$x_i^{(0)}(s,a) = \frac{1}{|\mathcal{A}_i|},\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}_i$. ▷ for all agents $i \in [n]$

for $t \leftarrow 1, 2, \ldots, T_x$ **do**

  $\boldsymbol{y}^{(t)} \leftarrow$ VIS-REG-PG$(\boldsymbol{x}^{(t-1)}, \epsilon_y)$ ▷ see Algorithm 2

  $\hat{\boldsymbol{g}}_i^{(t)} \leftarrow \text{REINFORCE}\left(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t)}; M\right)$ ▷ for all agents $i \in [n]$

  $\boldsymbol{x}_i^{(t)} \leftarrow \text{Proj}_{\mathcal{X}_i^{\zeta_x}}\left(\boldsymbol{x}_i^{(t-1)} - \eta_x \hat{\boldsymbol{g}}_i^{(t)}\right)$ ▷ for all agents $i \in [n]$

end for

$\boldsymbol{y}^{(T_x+1)} \leftarrow$ VIS-REG-PG$(\boldsymbol{x}^{T_x}, \epsilon_y)$

$\boldsymbol{x}^* \leftarrow \boldsymbol{x}^{(t^\star)}$ ▷ pick the best iterate

$\boldsymbol{y}^* \leftarrow \boldsymbol{y}^{(t^\star+1)}$

Algorithm 2 Visitation-Regularized Policy Gradient Algorithm (VIS-REG-PG)

Input: An MDP, a joint strategy of the team $\boldsymbol{x}$, and a desired accuracy $\epsilon > 0$.

Based on ϵ, set batch size K, sample traj. length H, stepsize η_y, truncation parameter ζ_y and regularization coeff. ν. ▷ see Theorem C.3

$y^{(0)}(s,b) \leftarrow \frac{1}{|\mathcal{B}|},\ \forall (s,b) \in \mathcal{S} \times \mathcal{B}$.

for Epoch $t \leftarrow 0, 1, \ldots, T_y$ **do**

  Independently sample K trajectories, $\mathcal{K}^{(t)}$, of length H under policy $\boldsymbol{y}^{(t)}$.

  $\hat{\boldsymbol{\lambda}}^{(t)} \leftarrow \dfrac{1}{K} \displaystyle\sum_{\tau \in \mathcal{K}^{(t)}} \tilde{\boldsymbol{\lambda}}(\tau | \boldsymbol{y}^{(t)})$,

  $\boldsymbol{u} \leftarrow \boldsymbol{r}(\boldsymbol{x}) - \nu \hat{\boldsymbol{\lambda}}^{(t)}$.

  $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)} \leftarrow \dfrac{1}{K} \displaystyle\sum_{\tau \in \mathcal{K}^{(t)}} \tilde{\boldsymbol{g}}(\tau | \boldsymbol{y}^{(t)}; \boldsymbol{u})$. ▷ $\tilde{\boldsymbol{g}}$ as in Definition 2.7.

  $\boldsymbol{y}^{(t+1)} \leftarrow \text{Proj}_{\mathcal{Y}^{\zeta_y}}(\boldsymbol{y}^{(t)} + \eta_y \hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})$.

end for

3.3 Analyzing Independent Stochastic Policy-Nested-Gradient

Algorithm 1, or ISPNG, is an instance of a nested-loop algorithm. As we have already informally stated, ISPNG runs gradient descent on the regularized max function $\Phi^\nu(\boldsymbol{x}) = \max_{\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x})} \left\{ \boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \frac{\nu}{2} \|\boldsymbol{\lambda}\|^2 \right\}$ for some parameter ν. This function has Hölder-continuous gradient and, as such, the convergence proof is underpinned by Theorem 3.1. Formally we state that:

Theorem 3.2 (Grad. Contunuity of Reg-Max Function)**.** *Let function $\Phi^\nu(\boldsymbol{x})$ be the maximum function of the regularized value function of an ATMG, with regularization coefficient $\nu > 0$. It is the case that, (i) Φ^ν is differentiable, (ii) $\nabla_{\boldsymbol{x}} \Phi^\nu$ is $(1/2, \ell_{1/2})$-Hölder continuous, i.e,*

$$\|\nabla_{\boldsymbol{x}} \Phi^\nu(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \Phi^\nu(\overline{\boldsymbol{x}})\| \leq \ell_{1/2} \|\boldsymbol{x} - \boldsymbol{x}'\|^{\frac{1}{2}}$$

with $\ell_{1/2} \coloneqq \dfrac{30 n^{\frac{1}{4}} |\mathcal{S}|^{\frac{5}{4}} \left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)^2}{\nu \min_s \rho(s) (1-\gamma)^{\frac{13}{2}}}$.

ISPNG manages to run gradient descent on function Φ^ν though the agents can never observe the exact gradient of Φ^ν. This is not only due to the randomness of gradient estimators but mainly because they cannot observe the adversary's actions and thus do not know the gradient w.r.t the regularizing term. Fortunately, the regularization coefficient plays a second role in bounding the inexactness error of the gradient estimates. For that reason, parameter ν admits a careful tuning.

Finally, the differentiability of Φ^ν and the per-player gradient domination property of the $V_{\boldsymbol{\rho}}$ implies that an ϵ-FOSP $\boldsymbol{x}^*$ and the corresponding best-response for the regularized value function, $\boldsymbol{y}^*$, constitute an ϵ-NE, leading to the main Theorem of this subsection:

Theorem 3.3 (Main Result; Formally in Theorem C.3)**.** *Given a desired accuracy $\epsilon > 0$, Algorithm 1 outputs a joint policy $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ for which it holds that,*

$$\mathbb{E}\left[V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*) - \min_{\boldsymbol{x}_i' \in \mathcal{X}_i} V_{\boldsymbol{\rho}}(\boldsymbol{x}_i', \boldsymbol{x}_{-i}^*, \boldsymbol{y}^*)\right] \leq \epsilon, \quad \forall i \in [n];$$

and

$$\mathbb{E}\left[\max_{\boldsymbol{y}' \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}') - V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] \leq \epsilon,$$

with a number of iterations and a number of samples that are $\mathsf{poly}\left(\frac{1}{\epsilon}, n, |\mathcal{S}|, \sum_i |\mathcal{A}_i| + |\mathcal{B}|, D_{\mathrm{m}}, \frac{1}{1-\gamma}, \frac{1}{\min_s \rho(s)}\right)$*. By D_{m} we denote the mismatch coefficient $D_{\mathrm{m}} \coloneqq \left\| \frac{\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}}{\boldsymbol{\rho}} \right\|_\infty$ (Definition C.2).*

4 Minimax in Nonconvex–Hidden-Strongly-Concave Functions

Finally, we would state a more general result compared to that of Theorem 3.3. We consider the general min-max nonconvex–nonconcave optimization problem, $\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$, when an additional structural assumption holds, *i.e.,* when f is nonconvex–hidden-strongly-concave. In particular, function f admits a reformulation of the form,

$$H(\boldsymbol{x}, \boldsymbol{u}) \coloneqq f\left(\boldsymbol{x}, c^{-1}(\boldsymbol{u}; \boldsymbol{x})\right),$$

where function H is a nonconvex–strongly-concave function defined on $\mathcal{X} \times \mathcal{U}$. The sets $\mathcal{X}$ and $\mathcal{U}$ are closed and convex, while $c(\cdot; \boldsymbol{x}) : \mathcal{Y} \to \mathcal{U}$ is an invertible mapping parametrized by $\boldsymbol{x}$. Moreover, we will denote $\mathcal{U}(\boldsymbol{x}) \coloneqq \{\boldsymbol{u} | \boldsymbol{u} = c(\boldsymbol{y}; \boldsymbol{x}),\ \forall \boldsymbol{y} \in \mathcal{Y}\}$. We further assume that the mapping c and its inverse are Lipschitz-continuous. Specifically,

Assumption 4.1. For the mapping c and its inverse, c^{-1}, it holds that

$$\begin{aligned} \|c(\boldsymbol{y}; \boldsymbol{x}) - c(\boldsymbol{y}'; \boldsymbol{x}')\| &\leq L_c(\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{y} - \boldsymbol{y}'\|), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}; \boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y} \\ \left\|c^{-1}(\boldsymbol{u}; \boldsymbol{x}) - c^{-1}(\boldsymbol{u}'; \boldsymbol{x}')\right\| &\leq L_{c^{-1}}(\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{u} - \boldsymbol{u}'\|), \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}; \boldsymbol{u}, \boldsymbol{u}' \in \mathcal{U}. \end{aligned}$$

If this is the case, the maximizer $\boldsymbol{u}^\star(\boldsymbol{x}) \coloneqq \mathrm{argmax}_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{x})} H(\boldsymbol{x}, \boldsymbol{u})$, is Hölder continuous w.r.t. $\boldsymbol{x}$ as stated by the following Theorem.

Theorem 4.1 (Formally in Theorem D.2)**.** *Let function $f(\boldsymbol{x}, \boldsymbol{y})$ be nonconvex–hidden-strongly-concave with a modulus of $\nu > 0$. Let also function H be a L_H-Lipschitz continuous and ℓ_H-smooth nonconvex–strongly-concave reformulation of f with an invertible mapping c for which Assumption 4.1 holds. Then,*

$$\|\boldsymbol{u}^\star(\boldsymbol{x}) - \boldsymbol{u}^\star(\boldsymbol{x}')\| \leq L_\star \|\boldsymbol{x} - \boldsymbol{x}'\|^{\frac{1}{2}}, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}.$$

where, $L_\star = O\left(\frac{1}{\nu}\right)$.

Theorem 4.2 (Convergence to an ϵ-SP; Formally in Theorem D.3)**.** *Let f be a nonconvex–hidden-strongly-concave function obeying to the same assumptions as f in Theorem 4.1 and $\epsilon > 0$. Further assume a maximization oracle with $O(\nu\epsilon^2)$-accuracy. There exists an algorithm that computes an ϵ-approximate saddle-point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ by making $T = O\left(\frac{1}{\nu^2\epsilon^3}\right)$ calls to the maximization oracle. Also, the maximization oracle can be implemented by stochastic gradient ascent with iteration complexity $T' = \tilde{O}\left(\frac{1}{\nu^3\epsilon^2}\right)$, and stepsize $\eta_y = O(\nu^2\epsilon^2)$.*

5 Conclusion, Future Work, and Limitations

Conclusions We expanded stochastic gradient techniques to be able to compute a stationary point in constrained optimization of nonconvex with weakly-smooth functions. We applied that result to design the first learning algorithm that computes an ϵ-approximate Nash equilibrium in adversarial team Markov games using a finite number of samples and iterations that scale polynomially with $1/\epsilon$ and the natural parameters of the game.

https://doi.org/10.52202/079017-2948

Future Work We believe that some questions that require further investigation are the following: (i) Is it possible to extend the techniques of [34] to establish convergence guarantees of stochastic gradient descent on nonconvex functions with Hölder-continuous gradient without batch-sampling of the gradient? (ii) Can we design a two-timescale gradient descent-ascent scheme for ATMGs that converges to a Nash equilibrium with best-iterate guarantees? (iii) Can we utilize some variance-reduction techniques to achieve a better sample complexity for learning an ϵ-NE in ATMGs?

Limitations The main limitations of our work are (i) the notion of independent learning as presented is weaker than the one presented in [27] – *i.e.*, our algorithm has an "inner loop", (ii) the fact that we did not present an example for which the function Φ^{ν} fails to be smooth; hence, it is unclear if we can prove the smoothness of this function and achieve tighter analysis. The first item can be addressed in future work by developing a two-timescale algorithmic approach. As for the second item, we remark that even if it is the case that Φ^{ν} is smooth for ATMGs, our provided convergence rates would be straightforwardly improved without any qualitative modification of the algorithm. Also, we would like to highlight that this discussion is related to Remark 2.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. FK carried over part of the research during an Archimedes Research Internship. IP would like to acknowledge an ICS research award and a startup grant from UCI.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[3] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[4] E. Altman. *Constrained Markov decision processes*. Routledge, 2021.

[5] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

[6] A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[7] N. Basilico, A. Celli, G. D. Nittis, and N. Gatti. Team-maxmin equilibrium: Efficiency bounds and algorithms. In S. Singh and S. Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017*, pages 356–362. AAAI Press, 2017.

[8] P. Bernhard and A. Rapaport. On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24(8):1163–1181, 1995.

[9] L. Bisi, L. Sabbioni, E. Vittori, M. Papini, and M. Restelli. Risk-averse trust region optimization for reward-volatility reduction. *arXiv preprint arXiv:1912.03193*, 2019.

[10] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter. From error bounds to the complexity of first-order descent methods for convex functions, 2016.

[11] C. Borgs, J. T. Chayes, N. Immorlica, A. T. Kalai, V. S. Mirrokni, and C. H. Papadimitriou. The myth of the folk theorem. *Games Econ. Behav.*, 70(1):34–43, 2010.

[12] M. Bowling, N. Burch, M. Johanson, and O. Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

[13] K. Brantley, M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

[14] N. Brown, A. Bakhtin, A. Lerer, and Q. Gong. Combining deep reinforcement learning and search for imperfect-information games. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[15] N. Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[16] N. Brown and T. Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[17] Y. Cai, H. Luo, C.-Y. Wei, and W. Zheng. Uncoupled and convergent learning in two-player zero-sum markov games. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

[18] Y. Cai, A. Oikonomou, and W. Zheng. Finite-time last-iterate convergence for learning in multi-player games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

 https://doi.org/10.52202/079017-2948

[19] V. Campos, A. Trott, C. Xiong, R. Socher, X. Giró-i Nieto, and J. Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, pages 1317–1327. PMLR, 2020.

[20] L. Carminati, F. Cacciamani, M. Ciccone, and N. Gatti. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 2638–2657. PMLR, 2022.

[21] A. Celli and N. Gatti. Computational results for extensive-form adversarial team games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 27952–27964, 2021.

[23] D. Chen, Q. Zhang, and T. T. Doan. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022.

[24] X. Chen, X. Deng, and S. Teng. Settling the complexity of computing two-player nash equilibria. *J. ACM*, 56(3):14:1–14:57, 2009.

[25] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[26] F. C. Chu and J. Y. Halpern. On the np-completeness of finding an optimal strategy in games with common payoffs. *Int. J. Game Theory*, 30(1):99–106, 2001.

[27] C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

[28] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM J. Comput.*, 39(1):195–259, 2009.

[29] C. Daskalakis, N. Golowich, and K. Zhang. The complexity of markov equilibrium in stochastic games. *CoRR*, abs/2204.03991, 2022.

[30] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[31] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9256–9266, 2018.

[32] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In A. Blum, editor, *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, volume 124 of *LIPIcs*, pages 27:1–27:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[33] C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In S. Khuller and V. V. Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1466–1478. ACM, 2021.

[34] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[35] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

[36] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization, 2021.

[37] D. Ding, C.-Y. Wei, K. Zhang, and M. R. Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. *arXiv preprint arXiv:2202.04129*, 2022.

[38] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[39] S. Emmons, C. Oesterheld, A. Critch, V. Conitzer, and S. Russell. For learning in symmetric teams, local optima are global nash equilibria. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 5924–5943. PMLR, 2022.

[40] L. Erez, T. Lancewicki, U. Sherman, T. Koren, and Y. Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*, pages 9343–9373. PMLR, 2023.

[41] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[42] G. Farina, A. Celli, N. Gatti, and T. Sandholm. Ex ante coordination and collusion in zero-sum multi-player extensive-form games. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 9661–9671, 2018.

[43] I. Fatkhullin, N. He, and Y. Hu. Stochastic optimization under hidden convexity, 2023.

[44] T. Fiez, L. Ratliff, E. Mazumdar, E. Faulkner, and A. Narang. Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29049–29063. Curran Associates, Inc., 2021.

[45] L. Flokas, E.-V. Vlatakis-Gkaragkounis, and G. Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent, 2021.

[46] R. Fox, S. M. McAleer, W. Overman, and I. Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022*, volume 151 of *Proceedings of Machine Learning Research*, pages 4414–4425. PMLR, 2022.

[47] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[48] U. Ghai, Z. Lu, and E. Hazan. Non-convex online learning via algorithmic equivalence. *Advances in Neural Information Processing Systems*, 35:22161–22172, 2022.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[50] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[51] E. Gorbunov, A. B. Taylor, and G. Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

 https://doi.org/10.52202/079017-2948

[52] K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

[53] T. Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973.

[54] K. A. Hansen, T. D. Hansen, P. B. Miltersen, and T. B. Sørensen. Approximability and parameterized complexity of minmax values. In *International Workshop on Internet and Network Economics*, pages 684–695. Springer, 2008.

[55] S. Hansen, W. Dabney, A. Barreto, T. Van de Wiele, D. Warde-Farley, and V. Mnih. Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*, 2019.

[56] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

[57] E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[58] S. He, Y. Jiang, H. Zhang, J. Shao, and X. Ji. Wasserstein unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6884–6892, 2022.

[59] Y.-C. Ho and K.-C. Chu. Team decision theory and information structures in optimal control problems–part i. *IEEE Transactions on Automatic Control*, 17(1):15–22, 1972.

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

[61] J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4:1039–1069, 2003.

[62] C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

[63] C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

[64] Y. Jin, V. Muthukumar, and A. Sidford. The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*, 2022.

[65] F. Kalogiannis, I. Anagnostides, I. Panageas, E.-V. Vlatakis-Gkaragkounis, V. Chatziafratis, and S. A. Stavroulakis. Efficiently computing nash equilibria in adversarial team markov games. In *The Eleventh International Conference on Learning Representations*, 2023.

[66] F. Kalogiannis, E.-V. Vlatakis-Gkaragkounis, and I. Panageas. Teamwork makes von neumann work: Min-max optimization in two-team zero-sum games. *ICLR*, 2023.

[67] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

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

[69] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. *Advances in Neural Information Processing Systems*, 27, 2014.

[70] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

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

[72] H. Liu and P. Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR, 2021.

[73] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[74] C. Maheshwari, M. Wu, D. Pai, and S. Sastry. Independent and decentralized learning in markov potential games, 2022.

[75] J. Marschak. Elements for a theory of teams. *Management Science*, 1(2):127–137, 1955.

[76] P. Mertikopoulos, B. Lecouat, H. Zenati, C. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[77] A. Mladenovic, I. Sakos, G. Gidel, and G. Piliouras. Generalized natural gradient flows in hidden convex-concave games and gans. In *International Conference on Learning Representations*, 2021.

[78] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 1497–1507. PMLR, 2020.

[79] M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[80] M. Mutti, R. De Santi, P. De Bartolomeis, and M. Restelli. Challenging common assumptions in convex reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4489–4502, 2022.

[81] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[82] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[83] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[84] K. A. Oikonomidis, E. Laude, P. Latafat, A. Themelis, and P. Patrinos. Adaptive proximal gradient methods are universal without approximation. *arXiv preprint arXiv:2402.06271*, 2024.

[85] F. Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

[86] J. Perolat, B. de Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S. H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, and K. Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning, 2022.

[87] M. Piccione and A. Rubinstein. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, 20(1):3–24, 1997.

 https://doi.org/10.52202/079017-2948

[88] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[89] R. Radner. Team Decision Problems. *The Annals of Mathematical Statistics*, 33(3):857 – 881, 1962.

[90] I. Sakos, E.-V. Vlatakis-Gkaragkounis, P. Mertikopoulos, and G. Piliouras. Exploiting hidden structures in non-convex games for convergence to nash equilibrium. *Advances in Neural Information Processing Systems*, 36, 2024.

[91] M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. Decentralized q-learning in zero-sum markov games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18320–18334. Curran Associates, Inc., 2021.

[92] M. O. Sayin, F. Parise, and A. Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv preprint arXiv:2010.04223*, 2020.

[93] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

[94] L. Schulman and U. V. Vazirani. The duality gap for two-team zero-sum games. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

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

[96] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

[97] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359, 2017.

[98] B. V. Stengel and D. Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21(1-2):309–321, 1997.

[99] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

[100] A. Tamar and S. Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

[101] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, Ç. Gülçehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. P. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nat.*, 575(7782):350–354, 2019.

[102] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 2007.

[103] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4259–4299. PMLR, 15–19 Aug 2021.

[104] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[105] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[106] M. Yashtini. On the global convergence rate of the gradient descent method for functions with hölder continuous gradients. *Optimization letters*, 10:1361–1370, 2016.

[107] T. Yu, Y. Tian, J. Zhang, and S. Sra. Provably efficient algorithms for multi-objective competitive rl. In *International Conference on Machine Learning*, pages 12167–12176. PMLR, 2021.

[108] T. Zahavy, B. O'Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

[109] T. Zahavy, Y. Schroecker, F. Behbahani, K. Baumli, S. Flennerhag, S. Hou, and S. Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. *arXiv preprint arXiv:2205.13521*, 2022.

[110] B. H. Zhang, G. Farina, and T. Sandholm. Team belief DAG form: A concise representation for team-correlated game-theoretic decision making. *CoRR*, abs/2202.00789, 2022.

[111] B. H. Zhang and T. Sandholm. Team correlated equilibria in zero-sum extensive-form games via tree decompositions. *CoRR*, abs/2109.05284, 2021.

[112] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

[113] J. Zhang, C. Ni, C. Szepesvari, M. Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

[114] R. Zhang, Z. Ren, and N. Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

[115] S. Zhang, B. Liu, and S. Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10905–10913, 2021.

[116] Y. Zhang and B. An. Computing team-maxmin equilibria in zero-sum multiplayer extensive-form games. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 2318–2325. AAAI Press, 2020.

[117] Y. Zhang and B. An. Converging to team-maxmin equilibria in zero-sum multiplayer games. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 11033–11043. PMLR, 2020.

 https://doi.org/10.52202/079017-2948

Appendix

Table of Contents

A Further Related Work

We accommodate this section to mention a brief collection of related literature in the fields of team games, reinforcement learning, and optimization. The literature is vast and we can only manage to mention some representative works.

A.1 Team Games

Research on team games has been a major focus in economic and group decision theory for decades [75, 53, 89, 59]. A key modern reference is [98], which introduced the *team-maxmin equilibrium (TME)* for normal-form games, where the team's strategy maximizes their minimal expected payoff against any adversary response. Despite their optimality, TMEs are computationally intractable even for 3-player team games [54, 11]. Recently, practical algorithms have been developed for multiplayer games [117, 116, 7]. Team equilibria are also relevant to two-player zero-sum games with imperfect recall [87].

Due to TME's intractability, *TMECor*, a relaxed equilibrium concept involving a *correlation device*, has been studied [42, 21, 7, 117, 111, 110, 20]. TMECor permits correlated strategies but can be impractical in certain scenarios [98]. TMECor is also NP-hard for *imperfect-information* extensive-form games (EFGs) [26], although fixed-parameter-tractable (FPT) algorithms have been developed for specific EFG classes [111, 110].

The computational aspects of standard Nash equilibrium (NE) in adversarial team games are not well-understood, even in normal-form games. Von Neumann's *minimax theorem* [102] does not apply to team games, rendering traditional methods ineffective. [94] characterized the *duality gap* between teams, while in [66] it was shown that standard no-regret learning dynamics, such as gradient descent and optimistic Hedge, may fail to converge to mixed NE in binary-action adversarial team games.

A.2 Reinforcement Learning

Multiagent RL Nash equilibrium computation has been central in multiagent RL. Notable algorithms, such as Nash-Q [60, 61], guarantee convergence to Nash equilibria only under strict game conditions. The behavior of independent policy gradient methods [93] remains poorly understood. The impossibility result by the authors of [56] precludes universal convergence to Nash equilibria even in normal-form games, aligning with the computational intractability (PPAD-completeness) of Nash equilibria in two-player general-sum games [28, 24]. Surprisingly, recent work shows similar hardness in turn-based stochastic games, making (stationary) CCEs intractable [29, 64].

Thus, research has focused on specific game classes, like Markov potential games [68, 37, 114, 23, 74, 46] or two-player zero-sum Markov games [27, 103, 91, 22, 92]. As noted, adversarial Markov team games can unify and extend these settings. Identifying multi-agent settings where Nash equilibria are efficiently computable is a key open problem (see, *e.g.*, [27]). Recent guarantees for convergence to Nash equilibria have been found in symmetric games, including symmetric team games [39]. Additionally, weaker solution concepts, relaxing either Markovian or stationarity properties, have gained attention [29, 62].

Convex RL Maximizing a value function regularized by a term that is strongly-concave with respect to the state-action visitation measure is an instance of a convex RL problem. In that sense, our work is also related to that strain of literature. Convex RL [108, 112] is a framework that generalizes standard MDP problems by considering the optimization of an objective function that is convex (or concave) in the state (or state-action) visitation measures that the agent's policies induce. The value function of standard RL has an objective function linear to that measure. Common well-known problems that are unified below the lens of convex are (i) "pure-exploration" RL [57], where the agent maximizes the entropy of the state visitation measure, (ii) imitation learning [1], where an agent minimizes the distance of the state visitation measure their policy induces and the one induced by an expert, (iii) risk-averse RL [47] where the agent optimizes an objective function that is sensitive to the tail behavior of the agent and not merely their expected behavior [100, 99, 25, 9, 115, 80], (iv) constrained RL [4], where an agent optimizes their value function while making sure to satisfy a number of constraints that are dependent on their state-action visitation measure [5, 107, 13, 2], (v) diverse skills discovery, where the goal is to drive learning agents to acquire a diverse set of emergent skills [19, 41, 52, 55, 58, 72, 96, 109].

A.3 Optimization

Min-max Optimization Min-max optimization studies problems of the form $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. When the objective function f is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$, the corresponding variational inequality (VI) is monotone, and a wide range of algorithms have been proposed for computing an approximate saddle-point — see, *e.g.*, [81, 69].

It is also known that standard Gradient Descent/Ascent (GDA) exhibits time-averaged convergence while the actual trajectory of iterates might cycle [30, 31]. Methods like Extra Gradient or techniques such as optimism are used to ensure convergence [30, 31, 32, 76, 78, 18, 51].

For more general objectives, we know how to compute approximate saddle-points when the weak Minty Property is satisfied [36] and for functions where one (or both) side satisfies the PŁcondition [83, 44, 105]. On the negative side, we know that the problem in its full generality (nonconvex–nonconcave landscape with coupled linear constraints) is computationally intractable [33].

Hidden-Convex Optimization This nascent field of optimization [43, 48] considers nonconvex objectives that can be reformulated, through a change of variables, into a convex objective. Further, in the context of game theory, the notion of hidden-monotonicity has made its appearance in [45] and the subsequent works [77, 90].

Weakly-Smooth Optimization The majority of references that we encounter for weakly-smooth minimization assume convexity and concern the unconstrained setting. We mention the important references of [35, 82] while also more recent works [106, 85, 84].

https://doi.org/10.52202/079017-2948

B Nonconvex Weakly-Smooth Constrained Optimization

In this section we prove that stochastic projected gradient descent with an stochastic inexact oracle converges to an ϵ-FOSP in functions with Hölder continuous gradient. We complement this section with the proof of folklore lemmas of constrained optimization that show that the "gradient mapping" (Definition B.1) is an appropriate surrogate of stationarity also for the family of functions we consider.

Definition B.1 (Gradient Mapping)**.** *We define the gradient mapping and stochastic gradient mapping, $\boldsymbol{r}_\eta$ and $\hat{\boldsymbol{r}}_\eta$, to be:*

- $\boldsymbol{r}_\eta(\boldsymbol{z}) \coloneqq \frac{1}{\eta}\left(\boldsymbol{z} - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z} - \eta \boldsymbol{g}\left(\boldsymbol{z}\right)\right)\right)$*, with a shorthand notation,* $\boldsymbol{r}_\eta^t \coloneqq \boldsymbol{r}_\eta(\boldsymbol{z}^)$*;*
- $\hat{\boldsymbol{r}}_\eta(\boldsymbol{z}) \coloneqq \frac{1}{\eta}\left(\boldsymbol{z} - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z} - \eta \hat{\boldsymbol{g}}(\boldsymbol{z})\right)\right)$*, similarly,* $\hat{\boldsymbol{r}}_\eta^t \coloneqq \hat{\boldsymbol{r}}_\eta(\boldsymbol{z}^)$ *.*

B.1 Auxiliary Lemmas

In general, demonstrating that the gradient mapping is an adequate surrogate of stationarity in differentiable constraint optimization relies on the Lipschitz continuity of the function. We make sure that this is the case when the gradient is only Hölder continuous with $p < 1$.

Lemma B.1 (Inexact-Gradient Mapping as a Stationarity Surrogate)**.** If $\left\|\boldsymbol{r}_\eta(\boldsymbol{z})\right\| \leq \epsilon$ for some $\boldsymbol{z} \in \mathcal{Z}$, it holds that:

$$\max_{\boldsymbol{z}' \in \mathcal{Z}, \|\boldsymbol{z}' - \boldsymbol{z}^+\| \leq 1} \left\langle -\nabla\phi(\boldsymbol{z}^+), \boldsymbol{z}' - \boldsymbol{z}^+ \right\rangle \leq \vartheta + \eta^2\epsilon + \ell_p \eta^p \epsilon^p,$$

where $\boldsymbol{z}^+ \coloneqq \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z} - \eta \boldsymbol{g}(\boldsymbol{z})\right)$.

Proof. In (Inexact Stoch-PGD), $\|\boldsymbol{g}(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z})\| \leq \vartheta$, $\forall \boldsymbol{z} \in \mathcal{Z}$. Since $\boldsymbol{z}^+ \coloneqq \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z} - \eta \boldsymbol{g}(\boldsymbol{z})\right)$, it holds that

$$\boldsymbol{z}^+ = \operatorname*{argmin}_{\boldsymbol{z}' \in \mathcal{Z}} \left\{ \left\|\boldsymbol{z}' - \left(\boldsymbol{z} - \eta \boldsymbol{g}(\boldsymbol{z})\right)\right\|^2 \right\}.$$

Due to the optimality condition, we have

$$-\left(\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\boldsymbol{g}(\boldsymbol{z})\right) \in N_{\mathcal{Z}}(\boldsymbol{z}^+),$$

where $N_{\mathcal{Z}}(\boldsymbol{z})$ is the normal cone of $\mathcal{Z}$ at $\boldsymbol{z}$, $N_{\mathcal{Z}}(\boldsymbol{z}) \coloneqq \{\boldsymbol{v} | \left\langle \boldsymbol{v}, \boldsymbol{z}' - \boldsymbol{z} \right\rangle \leq 0, \forall \boldsymbol{z}' \in \mathcal{Z}\}$. From the latter, we can conclude that

$$\begin{aligned}
&-\left(\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\nabla\phi(\boldsymbol{z})\right) - \frac{1}{\eta}\left(-\nabla\phi(\boldsymbol{z}) + \boldsymbol{g}(\boldsymbol{z})\right) \in N_{\mathcal{Z}}(\boldsymbol{z}^+) \\
&-\left(\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\nabla\phi(\boldsymbol{z})\right) \in N_{\mathcal{Z}}(\boldsymbol{z}^+) + B\left(\frac{\vartheta}{\eta}\right) \\
&-\frac{1}{\eta}\nabla\phi(\boldsymbol{z}^+) - \left(\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\nabla\phi(\boldsymbol{z}) - \frac{1}{\eta}\nabla\phi(\boldsymbol{z}^+)\right) \in N_{\mathcal{Z}}(\boldsymbol{z}^+) + B\left(\frac{\vartheta}{\eta}\right).
\end{aligned}$$

Now, we bound $\left\|\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\nabla\phi(\boldsymbol{z}) - \frac{1}{\eta}\nabla\phi(\boldsymbol{z}^+)\right\|$,

$$\begin{aligned}
\left\|\boldsymbol{z}^+ - \boldsymbol{z} + \frac{1}{\eta}\nabla\phi(\boldsymbol{z}) - \frac{1}{\eta}\nabla\phi(\boldsymbol{z}^+)\right\| &\leq \left\|\boldsymbol{z}^+ - \boldsymbol{z}\right\| + \frac{1}{\eta}\left\|\nabla\phi(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z}^+)\right\| \\
&\leq \left\|\boldsymbol{z}^+ - \boldsymbol{z}\right\| + \frac{\ell_p}{\eta}\left\|\boldsymbol{z}^+ - \boldsymbol{z}\right\|^p \\
&\leq \eta\epsilon + \frac{\ell_p}{\eta^{1-p}}\epsilon^p.
\end{aligned}$$

Therefore we have

$$-\nabla\phi(\boldsymbol{z}^+) \in N_{\mathcal{Z}}(\boldsymbol{z}^+) + B\left(\vartheta + \eta^2\epsilon + \ell_p\eta^p\epsilon^p\right).$$

The latter display implies the statement of the lemma. □

We immediately have the following corollary,

Corollary B.1. For any $\boldsymbol{z} \in \mathcal{Z}$, denote $\boldsymbol{z}^+ \coloneqq \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z} - \eta \boldsymbol{g}(\boldsymbol{z})\right)$. $\mathbb{E}\left[\left\|\boldsymbol{r}_\eta(\boldsymbol{z})\right\|\right] \leq \epsilon$ implies that

$$\mathbb{E}\left[\max_{\boldsymbol{z}' \in \mathcal{Z}, \|\boldsymbol{z}' - \boldsymbol{z}^+\| \leq 1} \left\langle -\nabla\phi(\boldsymbol{z}^+), \boldsymbol{z}' - \boldsymbol{z}^+ \right\rangle\right] \leq \vartheta + \eta^2\epsilon + \ell_p \eta^p \epsilon^p.$$

B.2 Stochastic PGD with Inexact Gradients

The folklore proof of gradient descent for nonconvex functions relies on the Lipschitz continuity of the gradient to prove convergence to a first-order stationary point. When the gradient are not Lipschitz continuous but continuous in the weaker notion of Hölder continuity implies the following fact that we will eventually use to prove a "descent lemma".

Fact B.1. Let a function $\phi : \mathcal{Z} \to \mathbb{R}$ with (p, ℓ_p)-Hölder continuous gradient. Then, it is the case that for all $\boldsymbol{z}, \boldsymbol{z}'$,

$$\left|\phi(\boldsymbol{z}') - \phi(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{z}' - \boldsymbol{z}\rangle\right| \leq \frac{\ell_p}{1+p}\left\|\boldsymbol{z}' - \boldsymbol{z}\right\|^{1+p}.$$

Following [35], we discuss functions with Hölder-continuous gradient (see Definition 2.1) through the framework of inexact oracle. We show that the answer $(\phi(\boldsymbol{z}), \nabla\phi(\boldsymbol{z}))$ of an exact oracle for a nonconvex function satisfying Hölder gradient continuity can be translated into some "inexact" information for a smooth function. Parameters δ, ℓ' in Proposition B.1 can be treated as "inexactness" parameters and will be chosen as appropriate parameters of the exponent p of Hölder continuity.

Proposition B.1. For given δ, ℓ_p, and a tuning of $\ell' \coloneqq \frac{\ell_p^{\frac{2}{1+p}}}{\delta^{\frac{1-p}{1+p}}}$, it holds that for any $\boldsymbol{x}, \boldsymbol{x}'$,

$$\frac{\ell_p}{1+p}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^{1+p} \leq \frac{\ell'}{2}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2 + \delta.$$

Proof. We let $\chi \coloneqq \|\boldsymbol{x} - \boldsymbol{x}'\|$. By choosing the optimal ℓ' we can verify that

$$2\max_{\chi \geq 0}\left\{\frac{\ell_p}{1+p}\chi^{-1+p} - \delta\chi^{-2}\right\} = \ell_p\left(\frac{\ell_p}{2\delta}\cdot\frac{1-p}{1+p}\right)^{\frac{1-p}{1+p}} \leq \frac{\ell_p^{\frac{2}{1+p}}}{\delta^{\frac{1-p}{1+p}}}.$$

Where in the inequality we use the fact that $\left(\frac{1-p}{2(1+p)}\right)^{\frac{1-p}{1+p}} \leq 1$ for $0 \leq p \leq 1$. Setting $\ell' = \frac{\ell_p^{\frac{2}{1+p}}}{\delta^{\frac{1-p}{1+p}}}$ yields the desired inequality. □

Now, we can use Proposition B.1 as in place of the "descent-lemma" to Theorem C.3 to prove convergence to an ϵ-FOSP.

Theorem B.1. *Let $\phi : \mathcal{Z} \to \mathbb{R}$ be a (p, ℓ_p)-weakly smooth nonconvex function. Further, assume a stochastic inexact gradient oracle $\hat{\boldsymbol{g}}$. I.e., it holds that $\mathbb{E}\left[\hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\right] = 0$ and $\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\right\|^2\right] \leq \frac{\sigma^2}{M}$ for some $\boldsymbol{g} : \mathcal{Z} \to \mathcal{Z}^*$ where $\|\boldsymbol{g}(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z})\| \leq \vartheta$, $\forall \boldsymbol{z} \in \mathcal{Z}$. Implementing T updates of the form* (Inexact Stoch-PGD) *using $\hat{\boldsymbol{g}}$ and a stepsize $\eta = \frac{1}{2\ell'}$ guarantees that:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2\right] \leq \frac{8\ell_p^{\frac{2}{1+p}}\left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right] - \phi^*\right)}{\delta^{\frac{1-p}{1+p}}T} + \frac{8\sigma^2}{M} + 8\ell_p^{\frac{2}{1+p}}\delta^{\frac{2p}{1+p}} + 4\vartheta^2.$$

We postpone the proof to state a corollary that might help the reader gain some intuition on how the iteration complexity scales with p.

https://doi.org/10.52202/079017-2948

Corollary B.2. Let ϕ, $\hat{\boldsymbol{g}}$, the update rule of (Inexact Stoch-PGD) as in Theorem B.1, and stepsize $\eta = (\frac{\epsilon^{1-p}}{2^{3-2p}\cdot\ell_p})^{\frac{1}{p}}$. For t^* drawn uniformly at random from $[1,\ldots,T]$, it holds that:

$$\mathbb{E}\left[\left\|\boldsymbol{r}_\eta^*\right\|^2\right] \leq \frac{8\ell_p^{\frac{2}{1+p}}\left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right]-\phi^*\right)}{\delta^{\frac{1-p}{1+p}}T} + 16\ell_p^{\frac{2}{1+p}}\delta^{\frac{2p}{1+p}} + 8\vartheta^2 + \frac{18\sigma^2}{M},$$

where $\boldsymbol{r}_\eta^* \coloneqq \boldsymbol{r}_\eta^{t^*}$. Furthermore, by setting the parameters as $T \geq \frac{8^{\frac{1+p}{p}}\ell_p^{\frac{1}{p}}\left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right]-\phi^*\right)}{\epsilon^{\frac{1+p}{p}}}$, $\delta \leq \frac{\left(\frac{\epsilon}{8}\right)^{\frac{1+p}{p}}}{\ell_p^{\frac{1}{p}}}$, $\vartheta \leq \frac{\epsilon}{8}$, and $M \geq \frac{9\sigma^2}{2\epsilon^2}$, it is guaranteed that there will exist a $t^\star \in \{0,\ldots,T-1\}$ such that $\mathbb{E}[\boldsymbol{r}_\eta(\boldsymbol{z}^{t^\star})] \leq \epsilon$.

Proof. For the first claim,

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{r}_\eta^*\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{\eta}\left(\boldsymbol{z}^{t^*} - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^{t^*} - \eta\boldsymbol{g}(\boldsymbol{z}^{t^*})\right)\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{\eta}\left(\boldsymbol{z}^{t^*} - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^{t^*} - \eta\hat{\boldsymbol{g}}(\boldsymbol{z}^{t^*})\right)\right)\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\frac{1}{\eta}\left(\mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^{t^*} - \eta\hat{\boldsymbol{g}}(\boldsymbol{z}^{t^*})\right) - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^{t^*} - \eta\boldsymbol{g}(\boldsymbol{z}^{t^*})\right)\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\hat{\boldsymbol{r}}_\eta^*\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1}{\eta}\left(\boldsymbol{z}^{t^*} - \eta\hat{\boldsymbol{g}}(\boldsymbol{z}^{t^*}) - \boldsymbol{z}^{t^*} - \eta\boldsymbol{g}(\boldsymbol{z}^{t^*})\right)\right\|^2\right] \\
&= 2\mathbb{E}\left[\left\|\hat{\boldsymbol{r}}_\eta^*\right\|^2\right] + 2\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}(\boldsymbol{z}^{t^*}) - \boldsymbol{g}(\boldsymbol{z}^{t^*})\right\|^2\right] \\
&\leq \frac{8\ell_p^{\frac{2}{1+p}}\left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right]-\phi^*\right)}{\delta^{\frac{1-p}{1+p}}T} + 16\ell_p^{\frac{2}{1+p}}\delta^{\frac{2p}{1+p}} + 8\vartheta^2 + \frac{18\sigma^2}{M}.
\end{aligned}$$

Where the last inequality follows from Theorem B.1 and the fact that $\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\right\|^2\right] \leq \frac{\sigma^2}{M}$. By setting the parameters as in the corollary, we have $\mathbb{E}[\boldsymbol{r}_\eta(\boldsymbol{z}^{t^\star})] \leq \epsilon$. □

Remark 1. *With the same parameters we choose in Corollary B.2, Lemma B.1 guarantees that for any $p \in (0,1]$, $\boldsymbol{r}_\eta(\boldsymbol{z})$ is a sufficient surrogate for stationarity. In particular, $\left\|\boldsymbol{r}_\eta(\boldsymbol{z})\right\| \leq \epsilon$ implies that*

$$-\nabla\phi(\boldsymbol{z}^+) \in N_{\mathcal{Z}}(\boldsymbol{z}^+) + B\left(\left(\left(\frac{8^{1-p}}{\ell_p}\right)^{\frac{2}{p}} + 9\right)\epsilon\right).$$

Finally, we state one more auxiliary claim before proceeding to the proof of Theorem C.3.

Claim B.1. Consider an iterate of (Inexact Stoch-PGD), $\boldsymbol{z}^t$. Also, define $\boldsymbol{z}^+ = \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^t - \eta\boldsymbol{g}(\boldsymbol{z})\right)$, where $\boldsymbol{g}$ is the inexact-gradient oracle. It is the case that,

$$\left\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^+\right\| \leq \eta^2\frac{\sigma^2}{M}.$$

Proof. The proof follows easily from arguments we have already used,

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^+\right\|^2\right] &= \mathbb{E}\left[\left\|\mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^t - \eta\hat{\boldsymbol{g}}^t\right) - \mathrm{Proj}_{\mathcal{Z}}\left(\boldsymbol{z}^t - \eta\boldsymbol{g}^t\right)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\eta\hat{\boldsymbol{g}}^t - \eta\boldsymbol{g}^t\right\|^2\right] \\
&= \eta^2\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}^t - \boldsymbol{g}^t\right\|^2\right] \\
&\leq \eta^2\frac{\sigma^2}{M}.
\end{aligned}$$

□

Proof of Theorem B.1

Proof. Since $\|\boldsymbol{g}(\boldsymbol{z}) - \nabla\phi(\boldsymbol{z})\| \leq \vartheta$, from the weakly-smooth condition, we have

$$
\begin{aligned}
\phi\left(\boldsymbol{z}^{t+1}\right) &\leq \phi(\boldsymbol{z}^t) + \left\langle \nabla\phi\left(\boldsymbol{z}^t\right), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \right\rangle + \frac{\ell_p}{1+p}\left\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\right\|^{1+p} \\
&\leq \phi(\boldsymbol{z}^t) + \left\langle \nabla\phi\left(\boldsymbol{z}^t\right), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \right\rangle + \frac{\ell'}{2}\left\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\right\|^2 + \delta \qquad (1) \\
&= \phi(\boldsymbol{z}^t) + \left\langle \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \right\rangle + \left\langle \nabla\phi(\boldsymbol{z}^t) - \boldsymbol{g}(\boldsymbol{z}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \right\rangle + \frac{\ell'\eta^2}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \delta \\
&= \phi(\boldsymbol{z}^t) - \eta\left\langle \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \eta\left\langle \nabla\phi(\boldsymbol{z}^t) - \boldsymbol{g}(\boldsymbol{z}^t), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \frac{\ell'\eta^2}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \delta \qquad (2) \\
&= \phi(\boldsymbol{z}^t) - \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \eta\left\langle \nabla\phi(\boldsymbol{z}^t) - \boldsymbol{g}(\boldsymbol{z}^t), \hat{\boldsymbol{r}}_\eta^t \right\rangle \\
&\quad + \frac{\ell'\eta^2}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \delta \\
&\leq \phi(\boldsymbol{z}^t) - \eta\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \eta\left\langle \nabla\phi(\boldsymbol{z}^t) - \boldsymbol{g}(\boldsymbol{z}^t), \hat{\boldsymbol{r}}_\eta^t \right\rangle \\
&\quad + \frac{\ell'\eta^2}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \delta \qquad (3) \\
&\leq \phi(\boldsymbol{z}^t) - \eta\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \frac{\eta}{2}\left\|\nabla\phi(\boldsymbol{z}^t) - \boldsymbol{g}(\boldsymbol{z}^t)\right\|^2 \\
&\quad + \frac{\eta}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \frac{\ell'\eta^2}{2}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \delta \qquad (4) \\
&\leq \phi(\boldsymbol{z}^t) - \left(\frac{\eta}{2} - \frac{\ell'\eta^2}{2}\right)\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t \right\rangle + \frac{\eta}{2}\vartheta^2 + \delta \qquad (5) \\
&\leq \phi(\boldsymbol{z}^t) - \left(\frac{\eta}{2} - \frac{\ell'\eta^2}{2}\right)\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 \\
&\quad + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t - \boldsymbol{r}_\eta^t \right\rangle + \frac{\eta}{2}\vartheta^2 + \delta.
\end{aligned}
$$

Where

- (1) is because of Proposition B.1;
- in (2), we plug-in the definition of $\hat{\boldsymbol{r}}_\eta^t$;
- (3) uses the fact that $-\left\langle \hat{\boldsymbol{g}}(\boldsymbol{z}^t), \hat{\boldsymbol{r}}_\eta^t \right\rangle \leq -\frac{1}{\eta}\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2$;
- (4) is due to Young's inequality;
- in (5), we plug-in the error bound on the inexact-gradient oracle $\|\nabla\phi(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\|^2 \leq \vartheta$.

Continuing we have

$$
\begin{aligned}
&\phi\left(\boldsymbol{z}^{t+1}\right) \\
\leq\ &\phi(\boldsymbol{z}^t) - \left(\frac{\eta}{2} - \frac{\ell'\eta^2}{2}\right)\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \hat{\boldsymbol{r}}_\eta^t - \boldsymbol{r}_\eta^t \right\rangle \\
&+ \frac{\eta}{2}\vartheta^2 + \delta \\
\leq\ &\phi(\boldsymbol{z}^t) - \left(\frac{\eta}{2} - \frac{\ell'\eta^2}{2}\right)\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 + \eta\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta\left\|\hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right)\right\|^2 + \frac{\eta}{2}\vartheta^2 + \delta.
\end{aligned}
$$

https://doi.org/10.52202/079017-2948

Summing for $t = 0, \ldots, T-1$,

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \phi\left(\boldsymbol{z}^{t+1}\right) \\
\leq\ &\sum_{t=0}^{T-1} \left(\phi\left(\boldsymbol{z}^{t}\right) - \left(\frac{\eta}{2} - \frac{\ell' \eta^2}{2}\right) \|\hat{\boldsymbol{r}}_\eta^t\|^2 + \eta \left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta \left\| \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right) \right\|^2 \right) \\
&+ \frac{\eta}{2} \vartheta^2 T + \delta T.
\end{aligned}
$$

This is equivalent to

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \left(\frac{\eta}{2} - \frac{\ell' \eta^2}{2}\right) \left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2 \\
\leq\ &\phi\left(\boldsymbol{z}^0\right) - \phi\left(\boldsymbol{z}^T\right) + \sum_{t=0}^{T-1} \left(\eta \left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta \left\| \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right) \right\|^2 \right) \\
&+ \frac{\eta}{2} \vartheta^2 T + \delta T \\
\leq\ &\phi\left(\boldsymbol{z}^0\right) - \phi^* + \sum_{t=0}^{T-1} \left(\eta \left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle + \eta \left\| \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right) \right\|^2 \right) + \frac{\eta}{2} \vartheta^2 T + \delta T.
\end{aligned}
$$

Taking expectations, we have

$$
\begin{aligned}
&\left(\frac{\eta}{2} - \frac{\ell' \eta^2}{2}\right) \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2\right] \\
\leq\ &\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right] - \phi^* + \sum_{t=0}^{T-1} \left(\eta \mathbb{E}\left[\left\langle \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right), \boldsymbol{r}_\eta^t \right\rangle\right] + \eta \mathbb{E}\left[\left\| \hat{\boldsymbol{g}}\left(\boldsymbol{z}^t\right) - \boldsymbol{g}\left(\boldsymbol{z}^t\right) \right\|^2\right] \right) \\
&+ \delta T + \frac{\eta}{2} \vartheta^2 T \\
\leq\ &\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right] - \phi^* + \eta \frac{\sigma^2}{M} T + \delta T + \frac{\eta}{2} \vartheta^2 T.
\end{aligned}
$$

By setting $\eta \leftarrow \frac{1}{2\ell'}$, it holds that

$$
\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\hat{\boldsymbol{r}}_\eta^t\right\|^2\right] &\leq \frac{8\ell' \left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right] - \phi^*\right)}{T} + \frac{8\sigma^2}{M} + 8\ell' \delta + 4\vartheta^2 \\
&= \frac{8\ell_p^{\frac{2}{1+p}} \left(\mathbb{E}\left[\phi\left(\boldsymbol{z}^0\right)\right] - \phi^*\right)}{\delta^{\frac{1-p}{1+p}} T} + \frac{8\sigma^2}{M} + 8\ell_p^{\frac{2}{1+p}} \delta^{\frac{2p}{1+p}} + 4\vartheta^2.
\end{aligned}
$$

This completes the proof.

□

C Adversarial Team Markov Games

In this section we the formal proofs of our claims regarding ATMGs. Before proceeding, let us provide a roadmap of the current section:

- Beginning, in Table 1 we offer a concise summary of our ATMG-related notation.
- We proceed to present a number of crucial facts regarding MDPs in Appendix C.1. In particular, facts regarding the state-action visitation measure.
- In Appendix C.3, we demonstrate that the regularized value function has a unique maximizer that changes in Hölder-continuous way w.r.t. to team policies $\boldsymbol{x}$. This leads to the Hölder-continuity of the gradient of the regularized maximum function Φ^{ν} (see Theorem C.2).
- Having established the latter, in Section 3.2 we invoke the results on gradient descent for nonconvex functions with Hölder continuous gradient to get our main theorem regarding ϵ-NE learning in ATMGs (Theorem C.3).
- The tuning of the parameters of Theorem C.3 is supported by (i) Appendix C.5, where we get precise guarantees for maximizing the regularizing value function w.r.t. the adversary's policy $\boldsymbol{y}$ (Theorem C.4) (ii) Appendix C.6, where we define and analyze the gradient estimators used by the agents of the MG.

https://doi.org/10.52202/079017-2948

Table 1: Notation

Parameters of the model:	
$\mathcal{S}$	State space
$\mathcal{N}$	Set of players
r	Reward function of the adversary
n	Number of players in the team
$\mathcal{A}_i$	Action space of player i of the team
$\mathcal{A}$	Team's joint action space
$\mathcal{B}$	Action space of the adversary
A_i	Number of actions available to player i of the team
B	Number of actions available to the adversary
$\mathcal{X}_i$	The set of feasible directly parameterized policies of player i: $\mathcal{X}_i \coloneqq \Delta(\mathcal{A}_i)^S$
$\mathcal{X}$	The set of feasible directly parameterized policies of the team: $\mathcal{X} \coloneqq \bigtimes_{i=1}^n \mathcal{X}_i$
$\mathcal{Y}$	The set of feasible directly parameterized policies of the adversary player: $\mathcal{Y} \coloneqq (\mathcal{B})^S$
$\mathbb{P}(s'\|s, \boldsymbol{a}, b)$	Probability of transitioning from state s to s' under the action profile $(\boldsymbol{a}, b)$
$\mathbb{P}(\boldsymbol{x}, \boldsymbol{y})$	The (row-stochastic) transition matrix of the Markov chain induced by $(\boldsymbol{x}, \boldsymbol{y})$
γ	Discount factor
$d_{s_0}^{\boldsymbol{x},\boldsymbol{y}}(s)$	The (un-normalized) state visitation measure for policy $(\boldsymbol{x}, \boldsymbol{y})$
$\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})$	The state-action visitation measure of the adversary when the team is playing policy $\boldsymbol{x}$
$\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y}), V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$	The value vector per-state, the expected value under initial distribution $\boldsymbol{\rho}$
$\boldsymbol{V}^\nu(\boldsymbol{x}, \boldsymbol{y}), V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \boldsymbol{y})$	The *regularized* value vector per-state, the expected value under initial distribution $\boldsymbol{\rho}$
τ	A trajectory of states and joint actions of the Markov game
Estimators:	
$\tilde{\boldsymbol{\lambda}}$	A single estimate of the state-action visitation measure
$\hat{\boldsymbol{\lambda}}$	The estimator of the state-action visitation measure
$\tilde{\boldsymbol{g}}$	A single estimate of the gradient
$\hat{\boldsymbol{g}}_{\boldsymbol{y}}$	The estimator of the gradient
Parameters:	
L_V	Lipschitz constant of the value function $V_{\boldsymbol{\rho}}(\cdot, \cdot)$
ℓ_V	Smoothness constant of the value function $V_{\boldsymbol{\rho}}(\cdot, \cdot)$
D_{m}	Distribution mismatch coefficient
Additional notation:	
$\Phi(\boldsymbol{x})$	Maximum of the value function given $\boldsymbol{x}$: $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$
$\Phi^\nu(\boldsymbol{x})$	Maximum of the *regularized* value function given $\boldsymbol{x}$: $\Phi^\nu(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \boldsymbol{y})$

C.1 Further Background on Markov Decision Processes

We need additional preliminaries on Markov decision processes (MDPs). Specifically, we will discuss the properties of the *(discounted) state and state-action visitation measure*. These measures represent the "discounted" expected amount of time the Markov chain—induced by the players' fixed policies—spends at state s (respectively, at a state action pair (s, b)) starting from initial state s'. Each visit is weighted by a discount factor γ^h, where h is the visit time. Notably, in [3] it is defined as a probability measure, meaning that for an initial state distribution $\boldsymbol{\rho}$, the discounted state visitation distribution sums to 1. For convenience, we will use the unnormalized definition from [88, Chapter 6.10], which sums to $\frac{1}{1-\gamma}$. This is why we refer to it as a *measure* instead of a *distribution*.

Definition C.1 (State Visit. Measure)**.** *Given an initial state distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ and a stationary joint policy $\boldsymbol{\pi} \in \Pi$, we define the state visitation frequency $d_{\bar{s}}^{\boldsymbol{\pi}}$ as follows:*

$$d_{\bar{s}}^{\boldsymbol{\pi}}(s) = \sum_{h=0}^{\infty} \gamma^h \, \mathbb{P}(s_h = s | \boldsymbol{\pi}, s_0 = \bar{s}).$$

Additionally, expanding the definition, we define $d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\bar{s} \sim \boldsymbol{\rho}} \left[d_{\bar{s}}^{\boldsymbol{\pi}}(s) \right]$.

For convenience, the expression $d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s)$ is utilized to represent the state visitation measure resulting from the policies $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$.

Fact C.1. Let MDP, $\mathcal{M}(\mathcal{S}, \mathcal{B}, \mathbb{P}, r, \gamma)$. Let a policy $\boldsymbol{\pi} \in \Delta(\mathcal{B})^{|\mathcal{S}|}$. For the corresponding state-action visitation measure $\boldsymbol{\lambda} \in \mathbb{R}^{S \times B}$ and the state visitation measure $\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}} \in \mathbb{R}^{S}$, it holds that,

$$\lambda_{s,b}(\boldsymbol{\pi}) = d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}(s)\pi(s, b), \quad \forall s \in \mathcal{S}, \forall b \in \mathcal{B}.$$

A quantity that is important in contemporary RL literature is that of the mismatch coefficient which we formally define here.

Definition C.2 (Distribution Mismatch Coefficient)**.** *Let $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ be a full-support distribution over states, and Π be the joint set of policies. We define the distribution mismatch coefficient D as*

$$D_{\mathrm{m}} \coloneqq \sup_{\boldsymbol{\pi} \in \Pi} \left\| \frac{\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}}{\boldsymbol{\rho}} \right\|_{\infty},$$

where $\frac{\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}}{\boldsymbol{\rho}}$ denotes element-wise division.

The following theorem that relates policies and visitation measures is essential to our analysis.

Theorem C.1 ([88, Theorem 6.9.1])**.** *Consider an adversarial Markov game Γ and a fixed team policy $\boldsymbol{x}$,*

(i) Any $\boldsymbol{y} \in \mathcal{Y}$ defines a feasible state-action visitation measure $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{B}|}$; namely,

$$\lambda_{s,b}(\boldsymbol{y}; \boldsymbol{x}) \coloneqq \sum_{\bar{s} \in \mathcal{S}} \rho(\bar{s}) \cdot \mathbb{E}_{\boldsymbol{y}} \left[\gamma^t \, \mathbb{P}(s^{(t)} = s, b^{(t)} = b \mid \boldsymbol{x}, s^{(0)} = \bar{s}) \right].$$

(ii) Any feasible state-action visitation measure $\boldsymbol{\lambda}$ defines a feasible $\boldsymbol{y} \in \mathcal{Y}$; namely,

$$y_{s,b} \coloneqq \frac{\lambda(s, b)}{\sum_{b' \in \mathcal{B}} \lambda(s, b')}, \ \forall (s, b) \in \mathcal{S} \times \mathcal{B}.$$

Further, for any such $\boldsymbol{y} \in \mathcal{Y}$ it holds that $\lambda_{s,b}(\boldsymbol{y}; \boldsymbol{x}) = \lambda(s, b), \ \forall (s, b) \in \mathcal{S} \times \mathcal{B}$, where $\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})$ is the induced discounted state-action measure.

An implication of the latter theorem is the fact that $\boldsymbol{\lambda}(\cdot; \boldsymbol{x})$ is a "1–1" mapping between policies and visitation measures. Following, we see that this mapping is also Lipschitz-continuous and smooth (see Lemmas C.1 to C.3).

Lemma C.1. For any initial distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$, function $V_{\boldsymbol{\rho}}$ is L-Lipschitz and ℓ-smooth with $L \coloneqq \frac{\sqrt{\sum_{i=1}^{n} |\mathcal{A}_i| + |\mathcal{B}|}}{(1-\gamma)^2}$ and $\ell \coloneqq \frac{2\gamma\left(\sum_{i=1}^{n} |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}$, in other words

$$|V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) - V_{\boldsymbol{\rho}}(\boldsymbol{x}', \boldsymbol{y}')| \leq \frac{\sqrt{\sum_{i=1}^{n} |\mathcal{A}_i| + |\mathcal{B}|}}{(1-\gamma)^2} \left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}') \right\|;$$

$$\left\| \nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) - \nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}', \boldsymbol{y}') \right\| \leq \frac{2\gamma \left(\sum_{i=1}^{n} |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3} \left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}') \right\|.$$

https://doi.org/10.52202/079017-2948

Proof. The proof follows from Lemma 4.4 in [68]. □

Lemma C.2. Let $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{S}||\mathcal{B}|}$ be the state-action visitation measure for the adversary, then $\boldsymbol{\lambda}$ is L_λ-Lipschitz continuous and ℓ_λ-smooth w.r.t to policy $(\boldsymbol{x}, \boldsymbol{y})$. Specifically, we have

$$\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\| \leq \frac{|\mathcal{S}|^{\frac{1}{2}}\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^2}\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right),$$

and

$$\|\nabla\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \nabla\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\| \leq \frac{2|\mathcal{S}|^{\frac{1}{2}}\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)^{\frac{3}{2}}}{(1-\gamma)^3}\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).$$

Proof. Each $\lambda_{s,b}$ can be considered as a value function for the given state s and the reward function is $r(a', b') = \mathbb{1}(b = b')$. Then by applying Lemma C.1, we have

$$|\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x}) - \lambda_{s,b}(\boldsymbol{y}';\boldsymbol{x}')| \leq \frac{\sqrt{\sum_i |\mathcal{A}_i| + |\mathcal{B}|}}{(1-\gamma)^2}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right), \tag{6}$$

$$\|\nabla\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x}) - \nabla\lambda_{s,b}(\boldsymbol{y}';\boldsymbol{x}')\| \leq \frac{2\gamma\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right). \tag{7}$$

From Equation (6) we get

$$\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|_\infty \leq \frac{\sqrt{\sum_i |\mathcal{A}_i| + |\mathcal{B}|}}{(1-\gamma)^2}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).$$

This implies

$$\begin{aligned}
\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\| &\leq \sqrt{|\mathcal{S}||\mathcal{B}|}\,\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|_\infty \\
&\leq \frac{\sqrt{|\mathcal{S}||\mathcal{B}|}\sqrt{\sum_i |\mathcal{A}_i| + |\mathcal{B}|}}{(1-\gamma)^2}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right) \\
&\leq \frac{\sqrt{|\mathcal{S}|}\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^2}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).
\end{aligned}$$

Similarly, from Equation (7), we have

$$\|\nabla\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x}) - \nabla\lambda_{s,b}(\boldsymbol{y}';\boldsymbol{x}')\| \leq \frac{2\gamma\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).$$

Thus

$$\begin{aligned}
\|\nabla\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \nabla\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|_F &\leq \sqrt{|\mathcal{S}||\mathcal{B}|}\cdot\max_{s\in\mathcal{S},b\in\mathcal{B}}\|\nabla\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x}) - \nabla\lambda_{s,b}(\boldsymbol{y}';\boldsymbol{x}')\| \\
&\leq \frac{2\gamma\sqrt{|\mathcal{S}||\mathcal{B}|}\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).
\end{aligned}$$

Where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix. Finally, we have

$$\begin{aligned}
\|\nabla\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \nabla\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\| &\leq \|\nabla\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \nabla\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|_F \\
&\leq \frac{2\gamma\sqrt{|\mathcal{S}||\mathcal{B}|}\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right) \\
&\leq \frac{2\gamma\sqrt{|S|\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)^3}}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right) \\
&\leq \frac{2\sqrt{|S|\left(\sum_i |\mathcal{A}_i| + |\mathcal{B}|\right)^3}}{(1-\gamma)^3}\cdot\left(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|\right).
\end{aligned}$$

□

Lemma C.3. Consider $\lambda_{\text{inv}}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) := \frac{\lambda_{s,b}(\boldsymbol{y};\boldsymbol{x})}{\sum_{b'}\lambda_{s,b'}(\boldsymbol{y};\boldsymbol{x})}$, which is a function that maps the adversary's state-action visitation measure $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) \in \Lambda(\boldsymbol{x})$ to the adversary's policy $\boldsymbol{y} \in \mathcal{Y}$. For any fixed team policy $\boldsymbol{x}$, λ_{inv} is $L_{\lambda_{\text{inv}}}$-Lipschitz continuous with respect $\boldsymbol{\lambda}$ where $L_{\lambda_{\text{inv}}} = \max_{s\in\mathcal{S}} \frac{2}{\boldsymbol{\rho}(s)(1-\gamma)}$. Specifically, it holds that

$$\|\boldsymbol{y} - \boldsymbol{y}'\| \leq L_{\lambda_{\text{inv}}} \|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x})\|.$$

Proof. Take the partial derivative of $\lambda_{\text{inv}}(\boldsymbol{\lambda})$, we have

$$\begin{aligned}
\left|\frac{\partial}{\partial\lambda_{s,b}}\frac{\lambda_{s,b}}{\sum_{b'}\lambda_{s,b'}}\right| &= \left|\frac{1}{\sum_{b'}\lambda_{s,b'}} - \frac{\lambda_{s,b}}{(\sum_{b'}\lambda_{s,b'})}\right| \\
&\leq \left|\frac{1}{\sum_{b'}\lambda_{s,b'}}\right| + \left|\frac{\lambda_{s,b}}{(\sum_{b'}\lambda_{s,b'})}\right| \\
&\leq \max_{s\in\mathcal{S}}\left\{\left|\frac{1}{\rho(s)}\right| + \frac{1}{\rho(s)(1-\gamma)}\right\} \\
&\leq \max_{s\in\mathcal{S}} \frac{2}{\rho(s)(1-\gamma)}.
\end{aligned}$$

This implies the Lipschitz continuity. □

The Regularized Value Function.

Lemma C.4. Function $V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) \coloneqq V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - \frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2$ is L_ν-Lipschitz continuous and ℓ_ν-smooth, where $L_\nu \coloneqq L + \frac{\nu L_\lambda}{2(1-\gamma)}$ and $\ell_\nu \coloneqq \ell + \frac{\nu\ell_\lambda}{2(1-\gamma)} + \frac{\nu L_\lambda^2}{2}$.

Proof. For Lipschitz continuity, we have

$$\begin{aligned}
&\left|V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x}',\boldsymbol{y}')\right| \\
\leq\ & |V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - V_{\boldsymbol{\rho}}(\boldsymbol{x}',\boldsymbol{y}')| + \frac{\nu}{2}\left|\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2 - \|\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|^2\right| \\
\leq\ & |V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - V_{\boldsymbol{\rho}}(\boldsymbol{x}',\boldsymbol{y}')| + \frac{\nu}{2}\max_{\boldsymbol{x},\boldsymbol{y}}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\| \cdot \left|\,\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\| - \|\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|\,\right| \\
\leq\ & |V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - V_{\boldsymbol{\rho}}(\boldsymbol{x}',\boldsymbol{y}')| + \frac{\nu}{2}\cdot\frac{1}{1-\gamma}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\| \\
\leq\ & \left(L + \frac{\nu L_\lambda}{2(1-\gamma)}\right)\cdot(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|).
\end{aligned}$$

For smoothness, denote the Jacobian matrix of $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$ w.r.t to $(\boldsymbol{x},\boldsymbol{y})$ by $\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{y})$, it holds that

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{x}}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2 - \nabla_{\boldsymbol{x}}\|\boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|^2\right\| \\
=\ & \left\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\right\| \\
\leq\ & \left\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\right\| \\
&+ \left\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}') - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')^\top\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\right\| \\
\leq\ & \|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|\,\|\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{y}) - \mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\| + \|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|\,\|\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\| \\
\leq\ & \frac{1}{1-\gamma}\|\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{y}) - \mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\| + \|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}';\boldsymbol{x}')\|\,\|\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}',\boldsymbol{y}')\| \qquad (8) \\
\leq\ & \left(\frac{\ell_\lambda}{1-\gamma} + L_\lambda\cdot L_\lambda\right)\cdot(\|\boldsymbol{x}-\boldsymbol{x}'\| + \|\boldsymbol{y}-\boldsymbol{y}'\|).
\end{aligned}$$

https://doi.org/10.52202/079017-2948

Where in (8) we used the fact that $\|\mathbf{J}_{\boldsymbol{\lambda}}(\boldsymbol{x}', \boldsymbol{y}')\| \leq L_\lambda$. We conclude that

$$\begin{aligned}
&\left\|\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y}) - \nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}', \boldsymbol{y}')\right\| \\
= &\left\|\nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) - \nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}', \boldsymbol{y}') + \frac{\nu}{2}\left(\nabla_{\boldsymbol{x}} \|\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})\|^2 - \nabla_{\boldsymbol{x}} \|\boldsymbol{\lambda}(\boldsymbol{y}'; \boldsymbol{x}')\|^2\right)\right\| \\
\leq &\|\nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) - \nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}', \boldsymbol{y}')\| + \frac{\nu}{2}\left\|\nabla_{\boldsymbol{x}} \|\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{x})\|^2 - \nabla_{\boldsymbol{x}} \|\boldsymbol{\lambda}(\boldsymbol{y}'; \boldsymbol{x}')\|^2\right\| \\
\leq &\left(\ell + \frac{\nu \ell_\lambda}{2(1-\gamma)} + \frac{\nu L_\lambda^2}{2}\right) \cdot (\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{y} - \boldsymbol{y}'\|).
\end{aligned}$$

□

Finally, we compute the Lipschitz continuity parameter of the reward vector that we already used in our previous claims.

Lemma C.5. Let $\boldsymbol{r}(\boldsymbol{x})$ be the reward function for the adversary when the team is playing policy $\boldsymbol{x}$. Then it holds that

$$\|\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x}')\| \leq L_r \|\boldsymbol{x} - \boldsymbol{x}'\|,$$

where $L_r = \sqrt{S}\left(\sum_{i=1}^n |\mathcal{A}_i| + B\right)$.

Proof.

$$\begin{aligned}
\|\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x}')\| &\leq \sqrt{|\mathcal{S}||\mathcal{B}|}\|r(\boldsymbol{x}) - r(\boldsymbol{x}')\|_\infty \\
&= \sqrt{|\mathcal{S}||\mathcal{B}|} \max_{s,b} |r(s, \boldsymbol{x}, b) - r(s, \boldsymbol{x}', b)| \\
&= \sqrt{|\mathcal{S}||\mathcal{B}| \sum_{i=1} |\mathcal{A}_i|}\|\boldsymbol{x} - \boldsymbol{x}'\| \qquad (9) \\
&\leq \sqrt{|\mathcal{S}|}\left(\sum_{i=1}^n |\mathcal{A}_i| + B\right)\|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned}$$

Where (9) follows from Claim D.9. in [65]. □

C.2 Auxiliary Lemmas

Bounding the stationarity error on the truncated simplex. The ζ-truncated simplex, $\Delta^{m,\zeta}$ is defined a the set of all probability vectors with no entry smaller than $\zeta > 0$. More formally, for a given dimension m and a $0 < \zeta \leq \frac{1}{m}$, the ζ-truncated simplex is defined to be

$$\Delta^{m,\zeta} = \left\{\boldsymbol{x} \;\middle|\; x_i \geq \zeta, \sum_{i=1}^m x_i = 1\right\}.$$

Lemma C.6. [40, Lemma 15] Let $\Delta^{m,\zeta}$ be the ζ-truncated m-simplex. If $0 \leq \zeta \leq \frac{1}{2m}$, then for all $\boldsymbol{x} \in \Delta^m$, there exists a $\boldsymbol{x}_\zeta \in \Delta^{m,\zeta}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\zeta\| \leq 2\zeta m$.

Proposition C.1 (Stationarity on the trunc. simplex)**.** Let an L_f-Lipschitz continuous differentiable function $f : \Delta^m \to \mathbb{R}$. Also, let an ϵ-approximate stationary point $\boldsymbol{x}_\zeta$ when the feasibility set is the truncated simplex $\Delta^{m,\zeta}$ such that

$$\left\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x}'_\zeta - \boldsymbol{x}_\zeta \right\rangle \leq \epsilon, \quad \forall \boldsymbol{x}'_\zeta \in \Delta^{m,\zeta}.$$

Then, $\boldsymbol{x}_\zeta$ is an $(\epsilon + 2\zeta m L_f)$-stationary point when the entire simplex is considered, *i.e,*

$$\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x} - \boldsymbol{x}_\zeta \rangle \leq \epsilon + 2\zeta m L_f, \quad \forall \boldsymbol{x} \in \Delta^m.$$

Proof. Consider $\boldsymbol{x}'_\zeta \in \Delta^{m,\zeta}$ such that $\left\|\boldsymbol{x} - \boldsymbol{x}'_\zeta\right\| \leq 2\zeta m$, such point exists due to Lemma C.6, we have

$$\begin{aligned}
\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x} - \boldsymbol{x}_\zeta \rangle &= \left\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x}'_\zeta - \boldsymbol{x}_\zeta \right\rangle + \left\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x} - \boldsymbol{x}'_\zeta \right\rangle \\
&\leq \epsilon + 2\zeta m L_f.
\end{aligned}$$

Where in the last inequality we use the fact that for all $\boldsymbol{x}'_\zeta \in \Delta^{m,\zeta}$, we have $\left\langle -\nabla f(\boldsymbol{x}_\zeta), \boldsymbol{x}'_\zeta - \boldsymbol{x}_\zeta \right\rangle \leq \epsilon$ and $\|\nabla f(\boldsymbol{x}_\zeta)\| \leq L_f$. □

From stationarity to optimality.

Lemma C.7 (Gradient Domination)**.** Let a single-agent MDP with action-space $\mathcal{A}$ and directly-parametrized policy $\boldsymbol{x} \in \Delta^\zeta(\mathcal{A})^{|\mathcal{S}|}$. Then it holds that

$$\max_{\boldsymbol{x}^\star \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^\star) - V_{\boldsymbol{\rho}}(\boldsymbol{x}) \leq \frac{1}{1-\gamma} D_{\mathrm{m}} \max_{\boldsymbol{x}' \in \Delta^\zeta(\mathcal{A})^{|\mathcal{S}|}} (\boldsymbol{x}' - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}) + \frac{2 D_{\mathrm{m}} \zeta |\mathcal{S}||\mathcal{A}| L}{1-\gamma}.$$

Proof. The proof follows easily from Proposition C.1 and the gradient domination property [3, Lemma 4.1]. □

C.3 Continuity of the maximizers

We begin this section by firstly introducing a proposition which we will leverage in Lemma C.8.

Proposition C.2. Consider the inequality of the form $\alpha\lambda^2 \leq \beta\lambda\chi + \gamma\chi$ where λ and χ are variables and α, β, γ are coefficients. Under the constraints that $\alpha, \beta, \gamma, \lambda, \chi \geq 0$, there is no solution of the form $\lambda \leq c\chi$ for any finite constant $c \geq 0$.

Proof. Solving the quadratic inequality for λ gives:

$$0 \leq \lambda \leq \frac{\beta\chi + \sqrt{\chi(4\alpha\gamma + \beta^2\chi)}}{2\alpha}.$$

We search for a positive constant c such that $\lambda \leq \frac{\beta\chi + \sqrt{\chi(4\alpha\gamma+\beta^2\chi)}}{2\alpha} \leq c\chi$, or equivalently,

$$\frac{1}{2\alpha}\left(\beta + \sqrt{\frac{4\alpha\gamma}{\chi} + \beta^2}\right) \leq c.$$

By observing that when α, β, γ are all positive constants and as $\chi \to 0$, $c \to \infty$, we conclude that no such finite constant c exists. □

We first define the maximizer for the regularized function $\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \frac{\nu}{2}\|\boldsymbol{\lambda}\|^2$. Since the function is strongly-concave w.r.t. $\boldsymbol{\lambda}$, the maximizing $\boldsymbol{\lambda}$ is unique. Specifically we denote

$$\boldsymbol{\lambda}^\star(\boldsymbol{x}) \coloneqq \operatorname*{argmax}_{\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x})} \left\{ \boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \frac{\nu}{2}\|\boldsymbol{\lambda}\|^2 \right\}. \tag{10}$$

Now we are ready to show an important lemma regarding to $\boldsymbol{\lambda}^\star(\boldsymbol{x})$.

Lemma C.8 (Continuity of the max. of reg. functions)**.** For any adversarial Markov game Γ, $\boldsymbol{\lambda}^\star(\boldsymbol{x})$ defined in (10) is $(1/2, L_\star)$-Hölder continuous, specifically

$$\|\boldsymbol{\lambda}^*(\boldsymbol{x}_1) - \boldsymbol{\lambda}^*(\boldsymbol{x}_2)\| \leq L_\star \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^{1/2},$$

where $L_\star \coloneqq \frac{2(n)^{1/4}}{\nu(1-\gamma)^{3/2}} |\mathcal{S}|^{1/2} \left(\sum_{k=1}^n |\mathcal{A}_k| + |\mathcal{B}|\right)^{\frac{3}{4}}$.

Proof. Consider team policies, $\boldsymbol{x}_1, \boldsymbol{x}_2$. It holds true for the unique maximizers $\boldsymbol{\lambda}^\star(\boldsymbol{x}_1), \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)$ of the adversary's regularized value function that,

$$\begin{aligned} \Big(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\Big)^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)) \leq 0, \quad \forall \boldsymbol{\lambda}_1 \in \Lambda(\boldsymbol{x}_1); \\ \Big(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\Big)^\top (\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)) \leq 0, \quad \forall \boldsymbol{\lambda}_2 \in \Lambda(\boldsymbol{x}_2), \end{aligned} \tag{11}$$

where $\Lambda(\boldsymbol{x})$ is the set of feasible state-action visitation measures of the adversary given team's policy $\boldsymbol{x}$. To bound the distance between the two vectors $\boldsymbol{\lambda}^\star(\boldsymbol{x}_1), \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)$, we observe that for any measure $\overline{\boldsymbol{\lambda}} \in \Lambda(\boldsymbol{x}_1) \cup \Lambda(\boldsymbol{x}_2)$, there exist a measure $\overline{\boldsymbol{\lambda}}_1 \in \Lambda(\boldsymbol{x}_1)$ that shares the same adversary's policy $\boldsymbol{y}$ as

https://doi.org/10.52202/079017-2948

in $\overline{\boldsymbol{\lambda}}$. It then follows from Lemma C.2 that $\left\|\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}\right\| \leq L_\lambda \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|$. Therefore we have for all $\overline{\boldsymbol{\lambda}} \in \Lambda(\boldsymbol{x}_1)$,

$$\begin{aligned}\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right) &= \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\right) \\ &\quad + \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\overline{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right) \\ &\leq L_\lambda \sqrt{|\mathcal{S}||\mathcal{B}|} \left(1 + \frac{\nu}{1-\gamma}\right) \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|. \end{aligned} \tag{12}$$

Where the last inequality follows from (11) and the fact that $\left\|\boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x})\right\| \leq \sqrt{|\mathcal{S}||\mathcal{B}|}\left(1 + \frac{\nu}{1-\gamma}\right)$ for any $\boldsymbol{x} \in \mathcal{X}$. Similarly, it also holds that for all $\overline{\boldsymbol{\lambda}} \in \Lambda(\boldsymbol{x}_1)$,

$$\left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)^\top \left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) \leq L_\lambda \sqrt{|\mathcal{S}||\mathcal{B}|} \left(1 + \frac{\nu}{1-\gamma}\right) \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|. \tag{13}$$

Plugging in $\overline{\boldsymbol{\lambda}} \leftarrow \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)$ and $\overline{\boldsymbol{\lambda}} \leftarrow \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)$ into (12) and (13) respectively

$$\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right) \leq L_\lambda \sqrt{|\mathcal{S}||\mathcal{B}|} \left(1 + \frac{\nu}{1-\gamma}\right) \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|,$$

$$\left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)^\top \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_1) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) \leq L_\lambda \sqrt{|\mathcal{S}||\mathcal{B}|} \left(1 + \frac{\nu}{1-\gamma}\right) \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|.$$

Adding the two inequalities results in

$$\begin{aligned}&\Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)\Big)^\top \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_1) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) \\ &\quad \leq 2L_\lambda \sqrt{|\mathcal{S}||\mathcal{B}|} \left(1 + \frac{\nu}{1-\gamma}\right) \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|. \end{aligned} \tag{14}$$

On the other hand, since $\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\lambda} - \frac{\nu}{2}\|\boldsymbol{\lambda}\|^2$ is ν-strongly concave in $\boldsymbol{\lambda}$, we have for all $\boldsymbol{\lambda}_1 \in \Lambda(\boldsymbol{x}_1)$.

$$\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}_1\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big) + \nu\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \leq 0.$$

We again use the fact that for every $\overline{\boldsymbol{\lambda}} \in \overline{\Lambda}$, it holds that there exists $\overline{\boldsymbol{\lambda}}_1 \in \Lambda(\boldsymbol{x}_1)$ s.t. $\|\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\| \leq L_\lambda\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. Therefore it holds that for any $\overline{\boldsymbol{\lambda}} \in \overline{\Lambda}$,

$$\begin{aligned}0 &\geq \left(\overline{\boldsymbol{\lambda}}_1 + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}_1 + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}})\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big) \\ &\quad + \nu\|\overline{\boldsymbol{\lambda}}_1 + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \\ &= \left((\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)) + (\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}})\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}_1} + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}})\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big) \\ &\quad + \nu\|\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}\|^2 + \nu\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 + 2\nu\left\langle\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\rangle \\ &= \left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}_1 + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}})\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big) \\ &\quad + \nu\|\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}\|^2 + \nu\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 + 2\nu\left\langle\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\rangle \\ &\quad + \left(\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}_1 + (\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}})\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big).\end{aligned}$$

Rearranging, we have

$$\begin{aligned}&\left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big) + \nu\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \\ &\leq \underbrace{-\nu\left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\right)}_{\Omega_1} \\ &\quad \underbrace{-\nu\|\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}\|^2 - 2\nu\left\langle\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\rangle}_{\Omega_2} \\ &\quad \underbrace{-(\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}})^\top\Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}_1 + \nu(\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}})\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\Big)}_{\Omega_3}.\end{aligned}$$

We bound Ω_1, Ω_2, and Ω_3 separately.

- For Ω_1, since $\|\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\| \leq L_\lambda \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, we have

$$\begin{aligned}\Omega_1 \leq \left|\nu \left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\right)\right| &\leq \nu \left\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\| \left\|\overline{\boldsymbol{\lambda}} - \overline{\boldsymbol{\lambda}}_1\right\| \\ &\leq \frac{2\nu}{1-\gamma}\sqrt{|\mathcal{S}||\mathcal{B}|} \cdot L_\lambda \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.\end{aligned}$$

  Where we use the fact that $\left\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\| \leq \left\|\overline{\boldsymbol{\lambda}}\right\| + \|\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|$ and $\|\boldsymbol{\lambda}\| \leq \|\boldsymbol{\lambda}\|_1 = \frac{1}{1-\gamma}$.

- For Ω_2, only the second term is possibly non-negative, it holds that

$$\left|\left\langle \overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right\rangle\right| \leq L_\lambda \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \frac{2}{1-\gamma}\sqrt{|\mathcal{S}||\mathcal{B}|}.$$

  Resulting in

$$\Omega_2 \leq \frac{4\nu}{1-\gamma}\sqrt{|\mathcal{S}||\mathcal{B}|} L_\lambda \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

- For Ω_3:

$$\Omega_3 \leq \left|(\overline{\boldsymbol{\lambda}}_1 - \overline{\boldsymbol{\lambda}})^\top \Big(- \nu\overline{\boldsymbol{\lambda}}_1 + \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\Big)\right| \leq \frac{2\nu}{1-\gamma}\sqrt{|\mathcal{S}||\mathcal{B}|} L_\lambda \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

Finally, by putting the bounds of $\Omega_1, \Omega_2, \Omega_3$ and setting $L' := \frac{8\nu}{1-\gamma}\sqrt{|\mathcal{S}||\mathcal{B}|} L_\lambda$, we have for all $\overline{\boldsymbol{\lambda}} \in \overline{\Lambda}$,

$$\left(\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\overline{\boldsymbol{\lambda}}\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\right) + \nu\|\overline{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \leq L'\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{15}$$

Concluding, we plug $\overline{\boldsymbol{\lambda}} \leftarrow \boldsymbol{\lambda}^\star(\boldsymbol{x}_2)$ in (15), resulting

$$\begin{aligned}\left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\right) + \nu\|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \\ \leq L'\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.\end{aligned} \tag{16}$$

Adding (14) and (16), we have

$$\begin{aligned}&2L_\lambda\sqrt{|\mathcal{S}||\mathcal{B}|}\left(1 + \frac{\nu}{1-\gamma}\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + L'\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \\ &\geq \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)\right) \\ &\quad + \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \left(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)\right) + \nu\|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 \\ &= \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)\Big) + \nu\|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2.\end{aligned}$$

Rearranging we get

$$\begin{aligned}\nu\|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|^2 &\leq \left(\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\right)^\top \Big(\left(\boldsymbol{r}(\boldsymbol{x}_2) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right) - \left(\boldsymbol{r}(\boldsymbol{x}_1) - \nu\boldsymbol{\lambda}^\star(\boldsymbol{x}_2)\right)\Big) \\ &\quad + L''\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \\ &\leq L_r\|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + L''\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,\end{aligned}$$

where $L'' := 2\left(1 + \frac{\nu}{1-\gamma}\right)\sqrt{|\mathcal{S}||\mathcal{B}|} L_\lambda + L' = \left(2 + \frac{10\nu}{1-\gamma}\right)\sqrt{|\mathcal{S}||\mathcal{B}|} L_\lambda$.

By setting $\lambda = \|\boldsymbol{\lambda}^\star(\boldsymbol{x}_2) - \boldsymbol{\lambda}^\star(\boldsymbol{x}_1)\|$ and $\chi = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, we consider the inequality $\nu\lambda^2 \leq L_r\lambda\chi + L''\chi$ with coefficients $\nu, L_r, L'' \geq 0$ and variables λ and χ. Choosing $\alpha \leftarrow \nu$, $\beta \leftarrow L_r$, and $\gamma \leftarrow L''$, then Proposition C.2 implies that $\boldsymbol{\lambda}^\star(\boldsymbol{x})$ is not Lipschitz-continuous w.r.t the team policy $\boldsymbol{x}$. Hence, we consider a solution of the form $0 \leq \lambda \leq \frac{\beta\chi + \sqrt{\chi(4\alpha\gamma + \beta^2\chi)}}{2\alpha} \leq c\chi^p$, where $\frac{1}{2} - p \geq 0$. We choose $p = \frac{1}{2}$ since it yields the best convergence rate from Theorem B.1. Solving the above inequality with $p = \frac{1}{2}$ gives that

https://doi.org/10.52202/079017-2948

- if $\chi = 0$, the inequality holds trivially;
- if $\chi > 0$, we have

$$c \geq \frac{\beta\chi + \sqrt{\chi(4\alpha\gamma + \beta^2\chi)}}{2\alpha\sqrt{\chi}} = \frac{\beta\sqrt{\chi}}{2\alpha} + \frac{\sqrt{4\alpha\gamma + \beta^2\chi}}{2\alpha}.$$

Since $\chi = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \leq \sqrt{n|\mathcal{S}|}\mathrm{Diam}_{\mathcal{X}_i} = \sqrt{2n|\mathcal{S}|}$, by plugging in the coefficients, we have

$$c = \frac{1}{\nu}\sqrt{2L_r^2 X + 4\nu L''} \leq \frac{2(n)^{1/4}}{\nu(1-\gamma)^{3/2}}|\mathcal{S}|^{1/2}\left(\sum_{k=1}^{n}|\mathcal{A}_k| + |\mathcal{B}|\right)^{\frac{3}{4}}.$$

By setting $L_\star = c$, we conclude that

$$\|\boldsymbol{\lambda}^*(\boldsymbol{x}_1) - \boldsymbol{\lambda}^*(\boldsymbol{x}_2)\| \leq L_\star \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^{1/2}.$$

□

We are now ready to show that $\Phi^\nu(\boldsymbol{x})$ is weakly-smooth.

Theorem C.2 (Hölder Continuous Max Value Func.)**.** *Let function $\Phi^\nu(\boldsymbol{x})$ be the maximum function of the regularized value function, $\Phi^\nu(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \boldsymbol{y})$. It is the case that,*

- *Φ^ν is differentiable,*
- *$\nabla_{\boldsymbol{x}}\Phi^\nu$ is $(1/2, \ell_{1/2})$-Hölder continuous, i.e,*

$$\|\nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}')\| \leq \ell_{1/2}\|\boldsymbol{x} - \boldsymbol{x}'\|^{1/2},$$

with $\ell_{1/2} \coloneqq \frac{30n^{\frac{1}{4}}|\mathcal{S}|^{\frac{5}{4}}\left(\sum_i|\mathcal{A}_i|+|\mathcal{B}|\right)^2}{\nu\min_s\rho(s)(1-\gamma)^{\frac{13}{2}}}$.

Proof. Since $\Phi^\nu(\boldsymbol{x})$ has a unique maximizer $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{x})$, by applying Danskin's Theorem [8] and the "1–1" correspondence between $\boldsymbol{\lambda}$ and $\boldsymbol{y}$ (Theorem C.1), we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}')\| &= \left\|\nabla_{\boldsymbol{x}}V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{\lambda}^\star(\boldsymbol{x}))) - \nabla_{\boldsymbol{x}}V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}', \boldsymbol{y}(\boldsymbol{\lambda}^\star(\boldsymbol{x}')))\right\| \\
&\leq \ell_\nu\left(\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{y}(\boldsymbol{\lambda}^\star(\boldsymbol{x})) - \boldsymbol{y}(\boldsymbol{\lambda}^\star(\boldsymbol{x}'))\|\right) \\
&\leq \ell_\nu(\|\boldsymbol{x} - \boldsymbol{x}'\| + L_{\lambda_{\mathrm{inv}}}\|\boldsymbol{\lambda}^\star(\boldsymbol{x}) - \boldsymbol{\lambda}^\star(\boldsymbol{x}')\|) \\
&\leq \ell_\nu\left((2n|\mathcal{S}|)^{\frac{1}{4}} + L_{\lambda_{\mathrm{inv}}}L_\star\right)\cdot\|\boldsymbol{x} - \boldsymbol{x}'\|^{\frac{1}{2}}.
\end{aligned}$$

Where in the last inequality we used Lemma C.8 and the fact that $\|\boldsymbol{x} - \boldsymbol{x}'\| \leq \sqrt{2n|\mathcal{S}|}$. Plugging in the coefficients in Lemma C.8, Lemma C.3, and Lemma C.4, it yields that

$$\|\nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}\Phi^\nu(\boldsymbol{x}')\| \leq \frac{30n^{\frac{1}{4}}|\mathcal{S}|^{\frac{5}{4}}\left(\sum_i|\mathcal{A}_i| + |\mathcal{B}|\right)^2}{\nu\min_s\boldsymbol{\rho}(s)(1-\gamma)^{\frac{13}{2}}}\cdot\|\boldsymbol{x} - \boldsymbol{x}'\|^{\frac{1}{2}}.$$

□

C.4 Analysis of ISPNG: Proof of Theorem 3.3

In this part we show that Algorithm 1 converges to an ϵ-NE. Essentially, Algorithm 1 implements projected gradient descent on the regularized maximum function $\Phi^\nu : \mathcal{X} \to \mathbb{R}$ with a stochastic ϑ-inexact gradient oracle. Function Φ^ν is Hölder-continuous (see Theorem C.2) and as such we can invoke Theorem C.3 to prove convergence to an ϵ-FOSP.

The inexactness of the gradient oracle, ϑ, is the sum of two error sources:

1. the fact that the adversary can only approximately maximize the regularized value function $V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \boldsymbol{y})$ — the iteration and sample complexity of maximizing $V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x}, \cdot)$ is provided in Theorem C.4;

2. the exact estimation of $\nabla\Phi^\nu$ requires estimation of the adversary's policy $\boldsymbol{y}$ — as we assume that the agents do not observe each other's actions this is impossible. Nevertheless, in Lemma C.9 it is proven that the inexactness error is bounded and controlled through the regularizer's coefficient ν.

After quantifying ϑ as the function of the latter two terms, the optimality gap of Algorithm 2 and a term that scales with $O(\nu)$, we can tune the rest of the parameters accordingly.

The resulting ϵ-FOSP, thanks to the gradient domination property (Lemma C.7), corresponds to an ϵ-NE.

Bounding the error of the inexact gradient. Following, we prove that the inexactness error of the gradient oracle is bounded by a function of the controllable parameter ν.

Lemma C.9 (Inexact gradient)**.** Let $V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x},\boldsymbol{y}) \coloneqq V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) - \frac{\nu}{2}\sum_s \|d^{\boldsymbol{x},\boldsymbol{y}}(s)\boldsymbol{y}(s)\|^2$, $\boldsymbol{y}^\nu(\boldsymbol{x}) \coloneqq \operatorname{argmax}_{\boldsymbol{y}}\{V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x},\boldsymbol{y})\}$, and $\boldsymbol{g}(\boldsymbol{x}) \coloneqq \nabla_{\boldsymbol{x}} V(\boldsymbol{x},\boldsymbol{y}_\nu(\boldsymbol{x}))$. Then, it holds that

$$\|\boldsymbol{g}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x},\boldsymbol{y})\|_2 \leq \frac{\nu|\mathcal{S}|^{\frac{1}{2}}\left(\sum_{i=1}^n |\mathcal{A}_i| + |\mathcal{B}|\right)}{(1-\gamma)^3}.$$

Proof. We observe that

$$\begin{aligned}\|\boldsymbol{g}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}^\nu(\boldsymbol{x},\boldsymbol{y})\| &= \left\|\nabla_{\boldsymbol{x}}\left(-\frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2\right)\right\| \\ &= \nu\,\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|\,\|\nabla_{\boldsymbol{x}}\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\| \\ &\leq \frac{\nu}{1-\gamma}L_\lambda.\end{aligned}$$

Where in the last inequality we use the fact that $\|\boldsymbol{\lambda}\| \leq \frac{1}{1-\gamma}$ and $\nabla_{\boldsymbol{x}}\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) \leq L_\lambda$.

□

Learning an ϵ-NE. We can now compile the intermediate statements to guarantee that ISPNG computes an ϵ-NE for any desired accuracy $\epsilon > 0$ within a finite number of iterations and samples.

Theorem C.3. *Consider an adversarial team Markov game Γ and Algorithm 1,* ISPNG*, with an outer-loop parameter tuning of:*

- $T = \frac{1061683200 D_{\mathrm{m}}^5 n^{\frac{1}{2}}|\mathcal{S}|^{\frac{9}{2}}\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)^6}{(1-\gamma)^{24}(\min_s \rho(s))^2\epsilon^5}$;
- $\eta_x = \frac{(\min_s \rho(s))^2(1-\gamma)^{22}\epsilon^3}{33177600 D_{\mathrm{m}}^3 n^{\frac{1}{2}}|\mathcal{S}|^{\frac{9}{2}}\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)^6}$;
- $\zeta_x = \frac{(1-\gamma)^3\epsilon}{6D_{\mathrm{m}}|\mathcal{S}|\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)^{\frac{3}{2}}}$;
- $M = \frac{2034 D_{\mathrm{m}}^3|\mathcal{S}|\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)^{\frac{7}{2}}}{(1-\gamma)^{10}(\min_s \rho(s))^4\epsilon^3}\max\left\{\frac{(1-\gamma)^4(\min_s \rho(s))^4\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)}{|\mathcal{S}|}, \frac{9}{2}\right\}$.

*Also, let the tuning of the inner-loop subroutine Algorithm 2 (*VIS-REG-PG*) be:*

- $\nu = \frac{(1-\gamma)^4\epsilon}{48 D_{\mathrm{m}}|\mathcal{S}|\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)}$;
- $T_y = \tilde{O}\left(\frac{D_{\mathrm{m}}^5|\mathcal{S}|^6(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|)^9}{(1-\gamma)^{21}(\min_s \rho(s))^4\epsilon^5}\right)$;
- $\eta_y = \frac{(1-\gamma)^{28}(\min_s \rho(s))^4\epsilon^4}{978447237120 D_{\mathrm{m}}^4|\mathcal{S}|^5(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|)^8}$;
- $\zeta_y = \frac{(1-\gamma)^{15}(\min_s \rho(s))^2\epsilon^3}{18432 D_{\mathrm{m}}^2|\mathcal{S}|^{\frac{7}{2}}(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|)^6}$;
- $K = \frac{19365101568 D_{\mathrm{m}}^4|\mathcal{S}|^7\left(\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|\right)^{12}}{(1-\gamma)^{36}(\min_s \rho(s))^4\epsilon^6}$;

https://doi.org/10.52202/079017-2948

- $H = \frac{2}{1-\gamma}\log\left(\frac{2293235712 D_{\mathrm{m}}^4 |\mathcal{S}|^4 (\sum_{i=1}^n |\mathcal{A}_i|+|\mathcal{B}|)^6}{(1-\gamma)^{22}(\min_s \rho(s))^2 \epsilon^4}\right)$.

It is the case that the output of the algorithm, $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, will be an ϵ-NE in expectation. Specifically, we have

$$\mathbb{E}\left[V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*) - \min_{\boldsymbol{x}_i' \in \mathcal{X}_i} V_{\boldsymbol{\rho}}(\boldsymbol{x}_i', \boldsymbol{x}_{-i}^*, \boldsymbol{y}^*)\right] \leq \epsilon, \quad \forall i \in [n]$$

and

$$\mathbb{E}\left[\max_{\boldsymbol{y}' \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}') - V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] \leq \epsilon.$$

Proof. Let $\boldsymbol{x}^*, \boldsymbol{y}^*$ be the final output of the algorithm, from Lemma C.7, we have

$$\begin{aligned}
&\mathbb{E}\left[V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*) - \min_{\boldsymbol{x}_i' \in \mathcal{X}_i} V_{\boldsymbol{\rho}}(\boldsymbol{x}_i', \boldsymbol{x}_{-i}^*, \boldsymbol{y}^*)\right] \\
&\leq \frac{1}{1-\gamma} D_{\mathrm{m}} \mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left(-\nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^*)\right] + \frac{2 D_{\mathrm{m}} \zeta |\mathcal{S}| \left(\sum_{i=1}^n |\mathcal{A}_i| + |\mathcal{B}|\right) L}{1-\gamma} \\
&\leq \frac{1}{1-\gamma} D_{\mathrm{m}} \mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^*)\right] + \frac{\nu}{1-\gamma} L_\lambda \mathrm{Diam}_{\mathcal{X}_i} \\
&\quad + \frac{2 D_{\mathrm{m}} \zeta |\mathcal{S}| \left(\sum_{i=1}^n |\mathcal{A}_i| + |\mathcal{B}|\right) L}{1-\gamma}. \qquad (17)
\end{aligned}$$

Where (17) follows from Lemma C.9. As for the first term of (17), let us define $\boldsymbol{y}^\star(\boldsymbol{x}) = \mathrm{argmax}_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{x}^+ = \mathrm{Proj}_{\mathcal{X}}\left(\boldsymbol{x}^{t^*-1} - \eta_x \nabla V_{\boldsymbol{\rho}}(\boldsymbol{x}^{t^*-1}, \boldsymbol{y}^\star(\boldsymbol{x}^{t^*-1}))\right)$. Then it holds that,

$$\begin{aligned}
&\mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left\{\left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^*)\right\}\right] \\
&= \mathbb{E}\Big[\max_{\boldsymbol{x}_i'} \Big\{ \left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+, \boldsymbol{x}_{-i}^*, \boldsymbol{y}^\star(\boldsymbol{x}^+))\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^*) \\
&\qquad + \left(\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+, \boldsymbol{x}_{-i}^* \boldsymbol{y}^\star(\boldsymbol{x}^+)) - \nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^*)\Big\}\Big] \\
&\leq \mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left\{\left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+ \boldsymbol{x}_{-i}^*, \boldsymbol{y}^\star(\boldsymbol{x}^+))\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^+)\right\}\right] + L_\nu \mathbb{E}\left[\left\|\boldsymbol{x}_i^+ - \boldsymbol{x}_i^*\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+, \boldsymbol{x}_{-i}^*, \boldsymbol{y}^\star(\boldsymbol{x}^+)) - \nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right\|\right] \cdot \mathrm{Diam}_{\mathcal{X}_i} \qquad (18) \\
&\leq \mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left\{\left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+, \boldsymbol{x}_{-i}^*, \boldsymbol{y}^\star(\boldsymbol{x}^+))\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^+)\right\}\right] + L_\nu \mathbb{E}\left[\left\|\boldsymbol{x}_i^+ - \boldsymbol{x}_i^*\right\|\right] \\
&\quad + \ell_\nu \left(\mathbb{E}\left[\left\|\boldsymbol{x}^+ - \boldsymbol{x}^*\right\|\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^*\right\|\right]\right) \cdot \mathrm{Diam}_{\mathcal{X}_i} \qquad (19) \\
&\leq \mathbb{E}\left[\max_{\boldsymbol{x}_i'} \left\{\left(-\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}_i^+, \boldsymbol{x}_{-i}^*, \boldsymbol{y}^\star(\boldsymbol{x}^+))\right)^\top (\boldsymbol{x}_i' - \boldsymbol{x}_i^+)\right\}\right] + L_\nu \mathbb{E}\left[\left\|\boldsymbol{x}_i^+ - \boldsymbol{x}_i^*\right\|\right] \\
&\quad + \ell_\nu \left(\mathbb{E}\left[\left\|\boldsymbol{x}^+ - \boldsymbol{x}^*\right\|\right] + \mathbb{E}\left[\left\|\boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^\star(\boldsymbol{x}^*)\right\|\right] + \mathbb{E}\left[\|\boldsymbol{y}^\star(\boldsymbol{x}^*) - \boldsymbol{y}^*\|\right]\right) \cdot \mathrm{Diam}_{\mathcal{X}_i}.
\end{aligned}$$

Where

- Equation (18) is due to $\left\|\nabla V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y})\right\| \leq L_\nu$;
- Equation (19) follows from the fact that $V_{\boldsymbol{\rho}}^{\nu}(\boldsymbol{x}, \boldsymbol{y})$ is ℓ_ν-smooth.

By choosing parameters specified above and combining Corollaries B.1 and B.2, Theorem C.4, Lemma C.12, and Claim B.1, we have the desired result. On the other hand, since

$$\begin{aligned}\mathbb{E}\left[\max_{\boldsymbol{y}'\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}') - V_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^*)\right] &\le \mathbb{E}\left[\max_{\boldsymbol{y}'\in\mathcal{Y}} V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}') - V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^*)\right] + \frac{\nu}{(1-\gamma)^2} \\ &= \mathbb{E}\left[V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^\star(\boldsymbol{x}^*)) - V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x}^*,\boldsymbol{y}^*)\right] + \frac{\nu}{(1-\gamma)^2} \quad (20) \\ &\le L_\nu E\left[\|\boldsymbol{y}^\star(\boldsymbol{x}^*) - \boldsymbol{y}^*\|\right] + \frac{\nu}{(1-\gamma)^2}.\end{aligned}$$

Where in (20) we use the fact that $\|\boldsymbol{\lambda}\|^2 \le \frac{1}{(1-\gamma)^2}$. Combining Theorem C.4, Lemma C.12, and choosing parameters specified above gives the desired result. □

C.5 Visitation-Regularized Policy Gradient Analysis

In this section, we consider the direct parameterization for the policy of the adversary. For any policy $\boldsymbol{y} \in \mathcal{Y}$, for any state $s \in \mathcal{S}$ and any action $b \in \mathcal{B}$, we have

$$y(b|s) = y_{s,b}.$$

Where $y_{s,b}$ denotes $(s,b)^{th}$ entry of the policy vector $\boldsymbol{y}$. In this section, we mainly focus on solving the following policy optimization problem:

$$\max_{\boldsymbol{y}\in\boldsymbol{y}^\zeta} V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) := \max_{\boldsymbol{y}\in\mathcal{Y}^\zeta}\left\{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2\right\}. \quad (21)$$

Where $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$ is the state-action visitation measure under policy $\boldsymbol{y}$ as in Definition 2.4. $\boldsymbol{r}(\boldsymbol{x})$ is the induced pay-off vector for the adversary when the team is playing according to strategy $\boldsymbol{x}$ and ν is the regularization coefficient. Then by policy gradient theorem [113], denote $F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) = V^{\nu}_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y})$ we have

$$\begin{aligned}\nabla_{\boldsymbol{y}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) &= \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right]^\top\left(\boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right) \\ &= \mathbb{E}_{\boldsymbol{\rho},\boldsymbol{y}}\left[\sum_{h=0}^{\infty}\gamma^h\cdot\left(\boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right)_{s_h,b_h}\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y(b'_h|s_{h'})\right)\right].\end{aligned}$$

Given the direct parameterization, we can show the following lemmas:

Lemma C.10. For any adversarial policy $\boldsymbol{y}$ and state-action pair (s,b), we have $||\nabla_{\boldsymbol{y}}\log y(b|s)|| \le \frac{1}{\zeta}$, $||\nabla^2_{\boldsymbol{y}}\log y(b|s)|| \le \frac{1}{\zeta^2}$, and $||\nabla_{\boldsymbol{y}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))|| \le \frac{1}{(1-\gamma)^2\zeta} + \frac{\nu}{(1-\gamma)^3\zeta}$ for any fixed $\boldsymbol{x}$.

Proof. By direct parameterization $y(b|s) = y_{s,b}$, we have

$$\|\nabla_{\boldsymbol{y}}\log y(b|s)\| = \|\nabla_{\boldsymbol{y}}\log y_{s,b}\| = \left\|\frac{1}{y_{s,b}}\boldsymbol{e}_{s,b}\right\| \le \frac{1}{\zeta}. \quad (22)$$

Where $\boldsymbol{e}_{s,b}$ denotes the standard basis for the $(s,b)^{th}$ entry. Similarly, we have

$$\|\nabla^2_{\boldsymbol{y}}\log y(b|s)\| = \left\|\mathrm{diag}\left(\frac{1}{y^2_{s,b}}\right)\right\| \le \frac{1}{\zeta^2}. \quad (23)$$

Where $\mathrm{diag}(\cdot)$ denotes the standard diagonal matrix. For the policy gradient, we show that

$$\begin{aligned}\|\nabla_{\boldsymbol{y}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))\| &= \left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right]^\top\left(\boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right)\right\| \\ &= \left\|\mathbb{E}\Big[\sum_{h=0}^{\infty}\gamma^h\cdot(r(\boldsymbol{x})_{s_hb_h} - \nu\lambda(\boldsymbol{y};\boldsymbol{x})_{s_hb_h})\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y(b_{h'}|s_{h'})\right)\Big]\right\| \\ &\le \sum_{h=0}^{\infty}\gamma^h\cdot(1+\frac{\nu}{1-\gamma})\cdot(h+1)\cdot\frac{1}{\zeta} \quad (24) \\ &\le \frac{1}{(1-\gamma)^2\zeta} + \frac{\nu}{(1-\gamma)^3\zeta}.\end{aligned}$$

Where (24) is due to (22). □

https://doi.org/10.52202/079017-2948

Before we proceed to show the convergence towards global optimality for (21), we first define the notion of Moreau envelope and the proximal point.

Definition C.3 (Moreau Envelope and Proximal Point)**.** *For any $\boldsymbol{y} \in \mathcal{Y}^{\zeta}$, we use $F^{\nu}_{1/\beta}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))$to denote the Moreau envelope of function $F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))$ such that*

$$F^{\nu}_{1/\beta}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) := \max_{\boldsymbol{z}\in\mathcal{Y}^{\zeta}} \left\{ F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x})) - \frac{\beta}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\|^2 \right\}.$$

Moreover, we define the proximal point $\hat{\boldsymbol{y}}_{1/\beta}$ of Moreau envelope as following:

$$\hat{\boldsymbol{y}} := \operatorname*{argmax}_{\boldsymbol{z}\in\mathcal{Y}^{\zeta}} \left\{ F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) - \frac{\beta}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\|^2 \right\}.$$

Now we proceed to show the following lemma:

Lemma C.11. When running Algorithm 2, for any $t \geq 0$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{y}^{(t+1)} - \hat{\boldsymbol{y}}^{(t)}\right\|^2 \middle| \boldsymbol{y}^{(t)}\right] \leq & (1-\eta_y\beta)\left\|\boldsymbol{y}^{(t)} - \hat{\boldsymbol{y}}^{(t)}\right\|^2 + 2(1-\eta_y\beta)\eta_y(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H \\ & + \eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}})\right\|^2 \middle| \boldsymbol{y}^{(t)}\right].\end{aligned}$$

Where $\mathcal{C}_3 = \sqrt{|\mathcal{S}||\mathcal{B}|}\frac{6H}{(1-\gamma)^3\zeta}$.

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}^{t+1}-\hat{\boldsymbol{y}}^{(t)}\right\|^2\middle|\boldsymbol{y}^{(t)}\right] \\
=&\mathbb{E}\left[\left\|\mathrm{Proj}_{\mathcal{Y}}(\boldsymbol{y}^{(t)}+\eta_y\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}})-\mathrm{Proj}_{\mathcal{Y}}\left((1-\eta_y\beta)\hat{\boldsymbol{y}}^{(t)}+\eta_y\beta\boldsymbol{y}^{(t)}+\eta_y\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))\right)\right\|^2\middle|\boldsymbol{y}^{(t)}\right] \qquad (25)\\
\leq&\mathbb{E}\left[\left\|\boldsymbol{y}^{(t)}+\eta_y\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}}-\left((1-\eta_y\beta)\hat{\boldsymbol{y}}^{(t)}+\eta_y\beta\boldsymbol{y}^{(t)}+\eta_y\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))\right)\right\|^2\middle|\boldsymbol{y}^{(t)}\right] \\
=&\mathbb{E}\left[\left\|(1-\eta_y\beta)(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})+\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}}\right)\right\|^2\middle|\boldsymbol{y}^{(t)}\right] \\
=&\mathbb{E}\left[\left\|(1-\eta_y\beta)(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})+\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)+\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}}\right)\right\|^2\middle|\boldsymbol{y}^{(t)}\right] \\
=&\left\|(1-\eta_y\beta)(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})+\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\|^2 \\
&+2(1-\eta_y\beta)\eta_y\mathbb{E}\left[\left\langle(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})-\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right),\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}})\right\rangle\middle|\boldsymbol{y}^{(t)}\right] \\
&+\eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}^{(t)}_{\boldsymbol{y}})\right\|^2\middle|\boldsymbol{y}^{(t)}\right]. \qquad (26)
\end{aligned}$$

Where (25) follows from Lemma 3.2 in [34]. For the first part of (26), we have

$$\begin{aligned}
&\left\|(1-\eta_y\beta)(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})+\eta_y\left(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\|^2 \\
=&(1-\eta_y\beta)^2\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2+\eta_y^2\left\|\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2 \\
&+2(1-\eta_y\beta)\eta_y\left\langle\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)},\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\rangle \\
\leq&(1-\eta_y\beta)^2\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2+\eta_y^2\ell_\nu^2\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2+2(1-\eta_y\beta)\eta_y\ell_\nu\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2 \qquad (27)\\
=&(1-\eta_y\beta)\left(1-\eta_y\beta+2\eta_y\ell_\nu+\frac{\eta_y^2\ell_\nu^2}{1-\eta_y\beta}\right)\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2.
\end{aligned}$$

Where (27) follows from Lemma C.4. By setting η_y, β such that $2\eta_y\ell_\nu \leq \frac{\eta_y\beta}{2}$, and $\frac{\eta_y^2\ell_\nu^2}{1-\eta_y\beta} \leq \frac{\eta_y\beta}{2}$, we have

$$\begin{aligned}
&\left\|(1-\eta_y\beta)(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})+\eta_y(\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})))\right\|^2 \\
\leq&(1-\eta_y\beta)\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2. \qquad (28)
\end{aligned}$$

For the third part in (26), we have

$$
\begin{aligned}
&2(1-\eta_y\beta)\eta_y\mathbb{E}\Big[\Big\langle(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})-\eta_y\big(\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\big),\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\Big\rangle\Big|\boldsymbol{y}^{(t)}\Big]\\
\leq&2(1-\eta_y\beta)\eta_y\Big\|(\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)})-\eta_y\big(\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\big)\Big\|\cdot\mathbb{E}\Big[\Big\|\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\Big\|\Big|\boldsymbol{y}^{(t)}\Big]\\
\leq&2(1-\eta_y\beta)\eta_y(1+\eta_y\ell_\nu)\cdot\big\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\big\|\cdot\mathbb{E}\big[\big\|\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\big\|\big] \qquad (29)\\
\leq&2(1-\eta_y\beta)\eta_y(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H. \qquad (30)
\end{aligned}
$$

Where

- $\mathcal{C}_3=\sqrt{|\mathcal{S}||\mathcal{B}|}\frac{6H}{(1-\gamma)^3\zeta}$;
- (29) follows from Lemma C.4;
- (30) is due to Lemma C.14.

Combine (26), (28), and (30), we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{y}^{t+1}-\hat{\boldsymbol{y}}^{(t)}\right\|^2\Big|\boldsymbol{y}^{(t)}\right]\leq&(1-\eta_y\beta)\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2+2(1-\eta_y\beta)\eta_y(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H\\
&+\eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\right\|^2\Big|\boldsymbol{y}^{(t)}\right].
\end{aligned}
$$

□

We then show the result for convergence to optimality for (21).

Theorem C.4. *By setting* $\eta_y=\frac{\nu\epsilon}{10\ell_\nu\sigma^2L_{\lambda_{\text{inv}}}^2}$ *and* $H=\frac{2\log(1/\nu\epsilon)}{1-\gamma}$. *After running Algorithm 2 for* $T=\mathcal{O}\left(\frac{\ell_\nu L_{\lambda_{\text{inv}}}^2}{\nu}\log\left(\frac{1}{\epsilon}\right)+\frac{\ell_\nu\sigma^2L_{\lambda_{\text{inv}}}^4}{\nu^2\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ *iterations, we have*

$$
\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}))-F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x}))\right]\leq\epsilon.
$$

Where $\boldsymbol{y}_\zeta^\star$ *is the unique maximizer for the optimization problem* (21).

Proof. From Theorem 1 in [43], by setting $\beta=4\ell_\nu$, $\alpha\leq2\eta_y\ell_\nu$, and $\eta_y\leq\frac{2}{9\ell_\nu}$. Then for any $\boldsymbol{z}\in\mathcal{Y}^\zeta$, we have

$$
\begin{aligned}
&\mathbb{E}\left[F_{1/\beta}^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t+1)};\boldsymbol{x}))\right]\\
\geq&\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x}))-(1+s)\frac{\beta}{2}\left\|\hat{\boldsymbol{y}}^{(t)}-\boldsymbol{y}^{(t+1)}\right\|^2-\left(1+\frac{1}{s}\right)\frac{\beta}{2}\left\|\hat{\boldsymbol{y}}^{(t)}-\boldsymbol{z}\right\|^2\right]\\
\geq&\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x}))-(1+s)(1-\eta_y\beta)\frac{\beta}{2}\left\|\boldsymbol{y}^{(t)}-\hat{\boldsymbol{y}}^{(t)}\right\|^2\right]\\
&-\left(1+\frac{1}{s}\right)\frac{\beta}{2}\mathbb{E}[\left\|\hat{\boldsymbol{y}}^{(t)}-\boldsymbol{z}\right\|^2]-2\eta_y(1+s)(1-\eta_y\beta)(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H\\
&-(1+s)\frac{\beta}{2}\eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\right\|^2\Big|\boldsymbol{y}^{(t)}\right]. \qquad (31)
\end{aligned}
$$

Where (31) is due to Lemma C.11. By setting $s=\frac{\eta_y\beta}{2}$, we have $(1+s)(1-\eta_y\beta)\leq1-\frac{\eta_y\beta}{2}$, $1+s\leq2$, and $1+\frac{1}{s}\leq\frac{3}{\eta_y\beta}$. From Theorem 1 in [43], we get

$$
\begin{aligned}
&\mathbb{E}\left[F_{1/\beta}^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t+1)};\boldsymbol{x}))\right]\\
\geq&(1-\alpha)\mathbb{E}\left[F_{1/\beta}^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right]+\alpha F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}))-\beta\eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\right\|^2\Big|\boldsymbol{y}^{(t)}\right]\\
&-\left(\frac{3L_{\lambda_{\text{inv}}}^2\alpha^2}{2\eta_y}-\frac{(1-\alpha)\alpha\nu}{2}\right)\mathbb{E}\left[\left\|\boldsymbol{\lambda}(\hat{\boldsymbol{y}}^{(t)};\boldsymbol{x})-\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})\right\|^2\right]-2\eta_y(1-\alpha)(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H.
\end{aligned}
$$

https://doi.org/10.52202/079017-2948

Define $\Lambda_t := \mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu_{1/\beta}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right]$, by setting $\left(\frac{3L^2_{\lambda_{\text{inv}}}\alpha^2}{2\eta_y} - \frac{(1-\alpha)\alpha\nu}{2}\right) \leq 0$, we have

$$\Lambda_{t+1} \leq (1-\alpha)\Lambda_t + \beta\eta_y^2\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\right\|^2 \left|\boldsymbol{y}^{(t)}\right.\right] + 2\eta_y(1-\alpha)(1+\eta_y\ell_\nu)\cdot\mathcal{C}_3\gamma^H.$$

Summing over T iterations, and denote $\mathbb{E}\left[\left\|\nabla_{\boldsymbol{y}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)})\right\|^2 \left|\boldsymbol{y}^{(t)}\right.\right] = \sigma^2$, we get

$$\Lambda_T \leq (1-\alpha)^T\Lambda_0 + \frac{4\ell_\nu\eta_y^2}{\alpha}\sigma^2 + \frac{2\eta_y(1-\alpha)(1+\eta_y\ell_\nu)}{\alpha}\cdot\mathcal{C}_3\gamma^H.$$

By setting $H = \frac{2\log(1/\nu\epsilon)}{1-\gamma}$, $\alpha \leq \min\left\{2\eta_y\ell_\nu, \frac{\nu\eta_y}{2L^2_{\lambda_{\text{inv}}}}\right\}$, and $\eta_y = \frac{2}{9\ell_\nu}, \frac{\nu\epsilon}{10\ell_\nu\sigma^2L^2_{\lambda_{\text{inv}}}}$, after

$$T = \mathcal{O}\left(\frac{\ell_\nu L^2_{\lambda_{\text{inv}}}}{\nu}\log\left(\frac{1}{\epsilon}\right) + \frac{\ell_\nu\sigma^2L^4_{\lambda_{\text{inv}}}}{\nu^2\epsilon}\log\left(\frac{1}{\epsilon}\right)\right).$$

iterations, we get $\Lambda_T \leq \epsilon$. Where $\sigma^2 = \frac{\mathcal{C}_1}{K} + \mathcal{C}_2\cdot\gamma^{2H}$, $\mathcal{C}_1 = \frac{57}{(1-\gamma)^6\zeta^2}$, $\mathcal{C}_2 = \frac{126H^2}{(1-\gamma)^6\zeta^2}$, $\mathcal{C}_3 = \sqrt{|\mathcal{S}||\mathcal{B}|}\frac{6H}{(1-\gamma)^3\zeta}$. Since $F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))$ is smooth with respect to the state-action visitation measure $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$. We have

$$\begin{aligned}
\Lambda_T &= \mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu_{1/\beta}(\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x}))\right] \\
&= \mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - \max_{\boldsymbol{z}\in\mathcal{Y}}\left\{F^\nu(\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x})) - \frac{\beta}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{z};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x})\right\|^2\right\}\right] \\
&\geq \mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x})) + \frac{\beta}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x})\right\|^2\right] \\
&= \mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x}))\right].
\end{aligned}$$

Therefore

$$\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(T)};\boldsymbol{x}))\right] \leq \Lambda_T \leq \epsilon.$$

□

Define $\boldsymbol{y}^\star \in \mathcal{Y}$ such that $\boldsymbol{y}^\star = \operatorname{argmax}_{\boldsymbol{y}\in\mathcal{Y}}\left\{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) - \frac{\nu}{2}\|\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|^2\right\}$. We bound the distance between the the optimal $\boldsymbol{y}^\star$ and $\boldsymbol{y}^{(T)}$ from Algorithm 2.

Lemma C.12. For any $\boldsymbol{y} \in \mathcal{Y}^\zeta$, if $\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))\right] \leq \epsilon$, then we have

$$\mathbb{E}\left[\|\boldsymbol{y}^\star - \boldsymbol{y}\|\right] \leq L_{\lambda_{\text{inv}}}\left(\sqrt{\frac{8L_\lambda|\mathcal{B}|\zeta}{(1-\gamma)\nu}} + \sqrt{\frac{2\epsilon}{\nu}}\right).$$

Proof. Since $F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))$ is ν-strongly concave with respect to $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$, we have

$$\begin{aligned}
F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) &\geq F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) + \left\langle\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\rangle \\
&\quad + \frac{\nu}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\|^2 \qquad (32) \\
&= F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})) + \frac{\nu}{2}\left\|\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\|^2.
\end{aligned}$$

Where (32) holds because $\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})$ is the optimal solution for $F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))$ for any $\boldsymbol{y} \in \mathcal{Y}$. Therefore

$$\mathbb{E}\left[\left\|\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\right\|\right] \leq \sqrt{\frac{2}{\nu}\cdot\mathbb{E}\left[F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}))\right]} \leq \sqrt{\frac{2\epsilon}{\nu}}.$$

From the definition of $\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})$, it holds that for all $\boldsymbol{y}_\zeta \in \mathcal{Y}^\zeta$, we have $\left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) \right\rangle \leq 0$. Combine with Lemma C.6 and consider $\boldsymbol{y}^\star \in \mathcal{Y}$, we have

$$
\begin{aligned}
&\left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\rangle \\
&= \left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\rangle \qquad (33) \\
&= \left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) \right\rangle + \left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\rangle \\
&\leq \left\langle -\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\rangle \\
&\leq \left\| \nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) \right\| \cdot \left\| \boldsymbol{\lambda}(\boldsymbol{y}_\zeta;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\| \\
&\leq \frac{4L_\lambda |\mathcal{B}| \zeta}{1-\gamma}. \qquad (34)
\end{aligned}
$$

Where

- in (33) $\boldsymbol{y}_\zeta \in \mathcal{Y}^\zeta$ is chosen such that $\|\boldsymbol{y}^\star - \boldsymbol{y}_\zeta\| \leq 2\zeta|\mathcal{B}|$ according to C.6;
- (34) holds because $\|\nabla_{\boldsymbol{\lambda}} F^\nu(\boldsymbol{\lambda})\| \leq \frac{2}{1-\gamma}$ and $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})$ is L_λ-continuous.

Since $F^\nu(\boldsymbol{\lambda})$ is ν-strongly concave w.r.t $\boldsymbol{\lambda}$, we have

$$
\begin{aligned}
&\frac{\nu}{2} \left\| \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) \right\|^2 \\
\leq & F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})) - F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) + \left\langle \nabla_{\boldsymbol{\lambda}} F^\nu((\boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x})), \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) \right\rangle \\
\leq & \frac{4L_\lambda |\mathcal{B}| \zeta}{1-\gamma}.
\end{aligned}
$$

Thus we conclude that

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{y}^\star - \boldsymbol{y}\|\right] &\leq L_{\lambda_{\text{inv}}} \mathbb{E}\left[\|\boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x})\|\right] \\
&\leq L_{\lambda_{\text{inv}}} \left(\left\| \boldsymbol{\lambda}(\boldsymbol{y}^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) \right\| + \mathbb{E}\left[\left\| \boldsymbol{\lambda}(\boldsymbol{y}_\zeta^\star;\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{x}) \right\| \right] \right) \\
&\leq L_{\lambda_{\text{inv}}} \left(\sqrt{\frac{8L_\lambda |\mathcal{B}| \zeta}{(1-\gamma)\nu}} + \sqrt{\frac{2\epsilon}{\nu}} \right).
\end{aligned}
$$

□

https://doi.org/10.52202/079017-2948

C.6 Regarding the Gradient and Visitation Estimators

In this subsection we will quantify the bias and variance of the gradient and state-action visitation estimators used in Algorithms 1 and 2. In particular, REINFORCE:

- the gradient estimator for team agents is implemented by sampling a trajectory with horizon length, H, that is drawn from a geometric distribution for the team, and
- while, the state-action visitation estimators that the adversary uses come from sampled trajectories of a fixed horizon length H.

In the former case, the estimator is unbiased while in the second case the bias decays exponentially in H.

C.6.1 REINFORCE for Vanilla Policy Gradient

In the present work, the team agents only need to implement a batch version of REINFORCE [104]. That is, they get estimates $\hat{\boldsymbol{g}}_k^{(t)} = \frac{1}{M}\sum_{j=1}^{M} \tilde{\boldsymbol{g}}_k^{(t)}$, where:

$$\tilde{\boldsymbol{g}}_{i,j}^{(t)} = \sum_{h_j=1}^{H_j} r_i^{(h_j)} \sum_{h=1}^{H_j} \nabla \log x_i \left(a^{(h_j)}|s^{(h_j)}\right), \tag{REINFORCE}$$

with each H_j is a random variable following a geometric distribution with parameter $(1-\gamma)$.

Although the authors of [27] use ζ-greedy parametrization in order to bound the variance of the estimator, policies drawn from the ζ-truncated simplex imply the same inequality needed to bound the variance. Hence, we invoke the corresponding lemma.

Lemma C.13 ([27, Lemma 2])**.** When Equation (REINFORCE) is implemented with H following a geometric distribution with a parameter $1-\gamma$, and agent k selects policies from the ζ-truncated simplex on each state, it is the case that the gradient estimates satisfy:

$$\mathbb{E}\left[\hat{\boldsymbol{g}}_k^{(t)}\right] - \nabla_{\boldsymbol{x}_k} V_{\boldsymbol{\rho}}(\boldsymbol{x}^t, \boldsymbol{y}^t) = 0;$$

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_k^{(t)} - \nabla_{\boldsymbol{x}_k} V_{\boldsymbol{\rho}}(\boldsymbol{x}^t, \boldsymbol{y}^t)\right\|^2\right] \leq 24\frac{|\mathcal{A}_k^2|}{\zeta(1-\gamma)}.$$

C.6.2 Gradient Estimation for Visitation-Regularized Policy Gradient

In this subsection we will describe (i) a state-action visitation estimator with bounded bias and variance and (ii) a gradient estimator of the regularized value function whose bias and variance are also bounded.

Bounding the variance of the a gradient estimator with a deterministic choice of H was significantly less demanding than doing so with a randomized choice. This comes at the cost with a non-zero bias that nevertheless decays exponentially in H. For any policy of the adversary $\boldsymbol{y} \in \mathcal{Y}$, we introduce the H-horizon truncated state-action visitation measure

$$\lambda_{H,s,b}(\boldsymbol{y};\boldsymbol{x})_{s,b} := \sum_{h=0}^{H-1} \gamma^h \, \mathbb{P}(s_h = s, b_h = b|\boldsymbol{y}, s_0 \sim \boldsymbol{\rho}). \tag{35}$$

Where $\lambda_{H,s,b}(\boldsymbol{y};\boldsymbol{x})$ denotes the $(s,b)^{th}$ entry of $\lambda_H(\boldsymbol{y};\boldsymbol{x})$. For any reward vector $\boldsymbol{r}$, we have

$$\left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y};\boldsymbol{x})\right]^\top \boldsymbol{r} = \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h \cdot r(s_h, b_h) \cdot \left(\sum_{h'=0}^{h} \nabla_{\boldsymbol{y}} \log y(b_{h'}|s_{h'})\right) \middle| \boldsymbol{y}, s_0 \sim \boldsymbol{\rho}\right].$$

C.6.3 Controlling the Estimation Bias and Variance

In this subsection we will present a detailed analysis regarding estimators defined in Definition 2.6, Definition 2.7, and the ones used in Algorithm 2. Particularly, in Lemma C.14 we bound the bias of aforementioned estimators. This bias is inevitable for our analysis due to the fact we are

sampling trajectories of a finite length H over an infinite horizon. Proceeding to Lemma C.16 and Lemma C.17, we bound the variance of the state-action distribution measure estimator $\hat{\boldsymbol{\lambda}}^{(t)}$ and the gradient estimator $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ w.r.t their biased means. Finally, in Lemma C.18, we bound the distance between the gradient estimator $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ and the actual gradient $\nabla_{\boldsymbol{y}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))$.

Lemma C.14 (Bounded Bias of the Estimators)**.** For any adversary's policy $\boldsymbol{y} \in \mathcal{Y}$. We let $\tau = (s_0, b_0, s_1, b_1, \cdots, s_{H-1}, b_{H-1})$ be an H-length trajectory sampled from $\boldsymbol{y}$, then we have $\mathbb{E}_{\tau \sim \boldsymbol{y}}\left[\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y})\right] = \boldsymbol{\lambda}_H(\boldsymbol{y}; \boldsymbol{x})$ and $\mathbb{E}_{\tau \sim \boldsymbol{y}}\left[\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}; \boldsymbol{r})\right] = \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}; \boldsymbol{x})\right]^{\top} \boldsymbol{r}$. This implies that in Algorithm 2, $\mathbb{E}\left[\hat{\boldsymbol{\lambda}}^{(t)}\right] = \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})$ and $\mathbb{E}\left[\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}\right] = \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \boldsymbol{r}^{(t)}$. Moreover, we have:

- $\left\|\mathbb{E}\left[\hat{\boldsymbol{\lambda}}^{(t)}\right] - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right\| \leq \frac{\gamma^H}{1-\gamma}$, and
- $\left\|\mathbb{E}\left[\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}\right] - \nabla_{\boldsymbol{y}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right\| \leq \left(\frac{H+1}{(1-\gamma)\zeta} + \frac{\nu H + \nu + 1}{(1-\gamma)^2 \zeta} + \frac{\nu}{(1-\gamma)^3 \zeta}\right) \cdot \gamma^H$.

Proof. From the definition, we have

$$\mathbb{E}_{\tau \sim \boldsymbol{y}}\left[\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y})\right] = \boldsymbol{\lambda}_H(\boldsymbol{y}; \boldsymbol{x}), \quad \mathbb{E}_{\tau \sim \boldsymbol{y}}\left[\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}; \boldsymbol{r})\right] = \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}; \boldsymbol{x})\right]^{\top} \boldsymbol{r}.$$

Therefore,

$$\mathbb{E}\left[\hat{\boldsymbol{\lambda}}^{(t)}\right] = \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x}), \quad \mathbb{E}\left[\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}\right] = \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \boldsymbol{r}^{(t)}.$$

Then it holds that

$$\begin{aligned}
\left\|\mathbb{E}\left[\hat{\boldsymbol{\lambda}}^{(t)}\right] - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right\| &= \left\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right\| \\
&= \left\|\sum_{h=H}^{\infty} \gamma^h \cdot \mathbb{P}(s_h = s, b_h = b|\boldsymbol{y}^{(t)}, s_0 \sim \boldsymbol{\rho}) \cdot \boldsymbol{e}_{s_h, b_h}\right\| \\
&\leq \gamma^H \cdot \sum_{h=0}^{\infty} (\gamma^h \cdot 1) \\
&= \frac{\gamma^H}{1-\gamma}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\left\|\mathbb{E}\left[\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}\right] - \nabla_{\boldsymbol{y}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right\| \\
=& \left\|\left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \boldsymbol{r}^{(t)} - \nabla_{\boldsymbol{y}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right\| \\
=& \left\|\left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \nabla_{\boldsymbol{\lambda}} F^{\nu}(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})) - \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \nabla_{\boldsymbol{\lambda}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right\| \\
\leq& \left\|\left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} \left(\nabla_{\boldsymbol{\lambda}} F^{\nu}(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})) - \nabla_{\boldsymbol{\lambda}} F(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right)\right\| \\
&+ \left\|\left(\left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top} - \left[\nabla_{\boldsymbol{y}} \boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x})\right]^{\top}\right) \nabla_{\boldsymbol{\lambda}} F^{\nu}(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)}; \boldsymbol{x}))\right\|. \qquad (36)
\end{aligned}$$

https://doi.org/10.52202/079017-2948

For the first part in the above inequality, we have

$$
\begin{aligned}
&\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\left(\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\| \\
=&\left\|\sum_{h=0}^{H-1}\gamma^h\cdot\left(\frac{\partial F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}))}{\partial\lambda_{s_h,b_h}}-\frac{\partial F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))}{\partial\lambda_{s_h,b_h}}\right)\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right\| \\
\le&\sum_{h=0}^{\infty}\gamma^h\cdot\left\|\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}))-\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|_\infty\cdot\left\|\left(\sum_{h'=0}^{\infty}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right\| \\
\le&\sum_{h=0}^{\infty}\gamma^h\cdot\left\|\nu\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})-\nu\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|_\infty\cdot\left\|\left(\sum_{h'=0}^{\infty}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right\| \\
\le&\sum_{h=0}^{\infty}\gamma^h\cdot\nu\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})-\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\|_1\cdot(h+1)\cdot\frac{1}{\zeta} \qquad (37)\\
\le&\frac{\nu}{(1-\gamma)^2\zeta}\cdot\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})-\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\|_1 \\
\le&\frac{\nu}{(1-\gamma)^2\zeta}\cdot\left(\sum_{s,b}\sum_{h=H}^{\infty}\gamma^h\cdot\mathbb{P}(s_h=s,b_h=b|\boldsymbol{y},s_0\sim\boldsymbol{\rho})\right) \\
\le&\frac{\nu}{(1-\gamma)^3\zeta}\cdot\gamma^H. \qquad (38)
\end{aligned}
$$

For the second part in (36), we have

$$
\begin{aligned}
&\left\|\left(\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\right)\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\| \\
=&\left\|\mathbb{E}\left[\sum_{h=H}^{\infty}\gamma^h\cdot\frac{\partial F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))}{\partial\lambda_{s_h,b_h}}\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right]\right\| \\
\le&\sum_{h=H}^{\infty}\gamma^h\cdot\left(1+\frac{\nu}{1-\gamma}\right)\cdot(h+1)\cdot\frac{1}{\zeta} \\
\le&\left(1+\frac{\nu}{1-\gamma}\right)\cdot\left(\frac{H+1}{1-\gamma}+\frac{1}{(1-\gamma^2)}\right)\cdot\frac{1}{\zeta}\cdot\gamma^H \\
=&\left(\frac{H+1}{(1-\gamma)\zeta}+\frac{\nu H+\nu+1}{(1-\gamma)^2\zeta}+\frac{\nu}{(1-\gamma)^3\zeta}\right)\cdot\gamma^H. \qquad (39)
\end{aligned}
$$

Combining (36), (38), and (39), we get the result. □

Before we proceed to analyze the variance of the estimators, we first show the Lipschitz continuity of the gradient estimator.

Lemma C.15. Let $\tau=\{s_0,b_0,s_1,b_1,\cdots,s_{H-1},b_{H-1}\}$ be an arbitrary H-length trajectory. The gradient estimator satisfies

- For any policy $\boldsymbol{y}$, for any reward vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$,

$$\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y};\boldsymbol{r}_1)-\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y};\boldsymbol{r}_2)\|\le\frac{1}{(1-\gamma)^2\zeta}\cdot\|\boldsymbol{r}_1-\boldsymbol{r}_2\|_\infty.$$

- For any policies $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, for any reward vectors $\boldsymbol{r}$,

$$\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}_1;\boldsymbol{r})-\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}_2;\boldsymbol{r})\|\le\left(\frac{1}{(1-\gamma)^2\zeta^2}+\frac{\nu}{(1-\gamma)^3\,\zeta^2}\right)\cdot\|\boldsymbol{y}_1-\boldsymbol{y}_2\|.$$

Proof.

$$\begin{aligned}\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y};\boldsymbol{r}_1)-\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y};\boldsymbol{r}_2)\| &= \left\|\sum_{h=0}^{H-1}\gamma^h\cdot(r_1(s_h,b_h)-r_2(s_h,b_h))\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y(b_{h'}|s_{h'})\right)\right\| \\ &\le \sum_{h=0}^{H-1}\gamma^h\cdot\|\boldsymbol{r}_1-\boldsymbol{r}_2\|_\infty\cdot(h+1)\cdot\frac{1}{\zeta} \qquad (40)\\ &\le \frac{1}{(1-\gamma)^2\zeta}\cdot\|\boldsymbol{r}_1-\boldsymbol{r}_2\|_\infty.\end{aligned}$$

$$\begin{aligned}&\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}_1,\boldsymbol{r})-\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}_2,\boldsymbol{r})\| \\ \le\ & \left\|\sum_{h=0}^{H-1}\gamma^h\cdot r(s_h,b_h)\cdot\left(\sum_{h'=0}^{h}(\nabla_{\boldsymbol{y}}\log y_1(b_{h'}|s_{h'})-\nabla_{\boldsymbol{y}}\log y_2(b_{h'}|s_{h'}))\right)\right\| \\ \le\ & \sum_{h=0}^{H-1}\gamma^h\cdot r(s_h,b_h)\cdot(h+1)\cdot\frac{1}{\zeta^2}\cdot\|\boldsymbol{y}_1-\boldsymbol{y}_2\| \qquad (41)\\ \le\ & \frac{(1+\frac{\nu}{1-\gamma})}{(1-\gamma)^2\zeta^2}\cdot\|\boldsymbol{y}_1-\boldsymbol{y}_2\| \\ =\ & \left(\frac{1}{(1-\gamma)^2\zeta^2}+\frac{\nu}{(1-\gamma)^3\zeta^2}\right)\cdot\|\boldsymbol{y}_1-\boldsymbol{y}_2\|.\end{aligned}$$

Where

- (40) follows from (22);
- (41) is because of (23).

□

Now we analyze the variance of the estimators in the algorithm, we start with showing the following lemma.

Lemma C.16 (Bounded Var. of Visit. Estimator)**.** For $\hat{\boldsymbol{\lambda}}^{(t)}$ in Algorithm 2, the variance is bounded. It holds that

$$\mathbb{E}\left[\|\hat{\boldsymbol{\lambda}}^{(t)}-\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\|^2\right]\le\frac{1}{K(1-\gamma)^2}.$$

Where $\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})$ is the truncated state-action visitation measure for $\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})$ defined in (35)

Proof. It holds that

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{\lambda}}^{(t)}-\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}_i}\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y}^{(t)})-\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|^2\right] \\ &= \frac{1}{K}\cdot\mathbb{E}\left[\left\|\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y}^{(t)})-\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|^2\right] \qquad (42)\\ &\le \frac{1}{K}\cdot\mathbb{E}\left[\left\|\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y}^{(t)})\right\|^2\right] \qquad (43)\\ &\le \frac{1}{K(1-\gamma)^2}. \qquad (44)\end{aligned}$$

Where:

- (42) is due to $\mathbb{E}\left[\hat{\boldsymbol{\lambda}}^{(t)}\right]=\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})$ and the fact that trajectories $\tau\in\mathcal{K}^{(t)}$ are independently sampled;

https://doi.org/10.52202/079017-2948

- (43) is because the variance is bounded by the second moment;
- (44) is because $\|\tilde{\boldsymbol{\lambda}}(\tau|\boldsymbol{y}^{(t)})\| \leq \frac{1}{1-\gamma}$.

□

Now we analyze the variance of gradient estimator $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ by providing the following lemma:

Lemma C.17 (Bounded Var. of Grad. Estimator)**.** For $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ in Algorithm 2, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)} - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right] \leq \frac{3}{K(1-\gamma)^4\zeta^2} + \frac{6\nu}{K(1-\gamma)^5\zeta^2} + \frac{9\nu^2}{K(1-\gamma)^6\zeta^2}.$$

Proof. We denote $\boldsymbol{r}^\star = \nabla_{\boldsymbol{\lambda}} F^\nu\left(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right) = \boldsymbol{r}(\boldsymbol{x}) - \nu\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)} - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^{(t)}) - \frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star) + \frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)\right.\right. \\
&\left.\left. - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star + \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right] \\
\leq&3\mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\left(\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^{(t)}) - \tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)\right)\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star) - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right]. \qquad (45)
\end{aligned}$$

Where (45) is due to Cauchy-Schwarz inequality. For the first part in (45), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\left(\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^{(t)}) - \tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)\right)\right\|^2\right] \\
\leq&\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\mathbb{E}\left[\left\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^{(t)}) - \tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)\right\|^2\right] \qquad (46) \\
\leq&\frac{1}{(1-\gamma)^4\zeta^2}\cdot\mathbb{E}\left[\left\|\boldsymbol{r}^{(t)} - \boldsymbol{r}^\star\right\|_\infty^2\right] \qquad (47) \\
\leq&\frac{\nu^2}{(1-\gamma)^4\zeta^2}\cdot E\left[\left\|\hat{\boldsymbol{\lambda}}^{(t)} - \boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|\right] \qquad (48) \\
\leq&\frac{\nu^2}{K(1-\gamma)^6\zeta^2}. \qquad (49)
\end{aligned}$$

Where:

- (46) is due to Cauchy-Schwarz inequality;
- (47) follows from Lemma C.15;
- (48) follows the same proof as in (37);

- (49) is because of Lemma C.16.

For the second part in (45), it holds that,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{K}\sum_{\tau\in\mathcal{K}^{(t)}}\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\boldsymbol{r}^\star\right\|^2\right]\\
=&\frac{1}{K}\mathbb{E}\left[\left\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\boldsymbol{r}^\star\right\|^2\right] &(50)\\
\leq&\frac{1}{K}\mathbb{E}\left[\left\|\tilde{\boldsymbol{g}}(\tau|\boldsymbol{y}^{(t)};\boldsymbol{r}^\star)\right\|^2\right] &(51)\\
=&\frac{1}{K}\mathbb{E}\left[\left\|\sum_{h=0}^{H-1}\gamma^h\cdot r^\star(s_h,b_h)\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right\|^2\right]\\
\leq&\frac{1}{K}\left(\sum_{h=0}^{H-1}\gamma^h\cdot\left(1+\frac{\nu}{1-\gamma}\right)\cdot\frac{1}{\zeta}\cdot(h+1)\right)^2 &(52)\\
\leq&\frac{1}{K(1-\gamma)^4\zeta^2}+\frac{2\nu}{K(1-\gamma)^5\zeta^2}+\frac{\nu^2}{K(1-\gamma)^6\zeta^2}. &(53)
\end{aligned}$$

Where

- (50) is due to Lemma C.14 and the fact that trajectories τ are independently sampled;
- (51) is because variance is bounded by second moment;
- (52) follows from (22).

Finally for the last part in (45), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\boldsymbol{r}^\star-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\boldsymbol{r}^{(t)}\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top(\boldsymbol{r}^\star-\boldsymbol{r}^{(t)})\right\|^2\right]\\
\leq&\mathbb{E}\left[\left\|\sum_{h=0}^{H-1}\gamma^h\cdot\|\boldsymbol{r}^\star-\boldsymbol{r}^{(t)}\|_\infty\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right\|^2\right]\\
\leq&\left(\sum_{h=0}^{H-1}\gamma^h\cdot\nu\cdot(h+1)\cdot\frac{1}{\zeta}\right)^2\cdot\mathbb{E}\left[\left\|\hat{\boldsymbol{\lambda}}^{(t)}-\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|^2\right] &(54)\\
\leq&\frac{\nu^2}{(1-\gamma)^4\zeta^2}\cdot\frac{1}{K(1-\gamma)^2} &(55)\\
=&\frac{\nu^2}{K(1-\gamma)^6\zeta^2}. &(56)
\end{aligned}$$

Where

- (54) is due to (22) and (37);
- (55) is because of Lemma C.16.

https://doi.org/10.52202/079017-2948

Combine (45), (49), (53) and (56), we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right]\\
\leq&\frac{3\nu^2}{K(1-\gamma)^6\zeta^2}+3\left(\frac{1}{K(1-\gamma)^4\zeta^2}+\frac{2\nu}{K(1-\gamma)^5\zeta^2}+\frac{\nu^2}{K(1-\gamma)^6\zeta^2}\right)+\frac{3\nu^2}{K(1-\gamma)^6\zeta^2}\\
=&\frac{3}{K(1-\gamma)^4\zeta^2}+\frac{6\nu}{K(1-\gamma)^5\zeta^2}+\frac{9\nu^2}{K(1-\gamma)^6\zeta^2}.
\end{aligned}$$

□

After bounding the variance of $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ in Algorithm 2, we can prove the following lemma

Lemma C.18 (Bounded Dist. with Actual Grad.)**.** Consider $\boldsymbol{y}^{(t)}$ and $\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}$ in Algorithm 2, it holds that

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2\right]\leq\frac{\mathcal{C}_1}{K}+\mathcal{C}_2\cdot\gamma^{2H}.$$

Where

$$\mathcal{C}_1=\frac{57}{(1-\gamma)^6\zeta^2},\quad \mathcal{C}_2=\frac{126H^2}{(1-\gamma)^6\zeta^2}.$$

Proof. Let $\boldsymbol{r}^\star=\nabla_{\boldsymbol{\lambda}}F^\nu\left(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right)=\boldsymbol{r}(\boldsymbol{x})-\nu\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2\right]\\
=&\mathbb{E}\Bigg[\Bigg\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}+\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star\\
&+\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\Bigg\|^2\Bigg]\\
\leq&3\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}\right\|^2\right]\\
&+3\mathbb{E}\left[\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^{(t)}-\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star\right\|^2\right]\\
&+3\mathbb{E}\left[\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2\right].
\end{aligned}\tag{57}$$

Notice that the first part in (57) is bounded in Lemma C.17 and the second part is bounded in (56). For the last part, observe that

$$
\begin{aligned}
&\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star - \nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2 \\
=&\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2 \\
=&\Bigg\| \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \left(\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right) \\
&+ \left(\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\right) \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\Bigg\|^2 \\
\le& 2\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \left(\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\|^2 \\
&+ 2\left\|\left(\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\right) \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2. \qquad (58)
\end{aligned}
$$

Where (58) is follows from Cauchy-Schwarz inequality. For the first part, we have

$$
\begin{aligned}
&\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \left(\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\|^2 \\
\le& \left(\sum_{h=0}^{\infty}\gamma^h\cdot\nu\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\|_1\cdot(h+1)\cdot\frac{1}{\zeta}\right)^2 \qquad (59) \\
\le& \frac{\nu^2}{(1-\gamma)^4\zeta^2}\cdot\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\|_1^2.
\end{aligned}
$$

Where (59) is because of (37). Since

$$
\begin{aligned}
\left\|\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right\|_1^2 &= \left(\sum_{h=H}^{\infty}\sum_{s,b}\gamma^t\,\mathbb{P}(s_h=s,b_h=b|\boldsymbol{y}^{(t)},s_0\sim\boldsymbol{\rho})\right)^2 \\
&= \left(\gamma^H\sum_{h=0}^{\infty}\gamma^h\cdot 1\right)^2 \\
&\le \frac{\gamma^{2H}}{(1-\gamma)^2}.
\end{aligned}
$$

We have

$$
\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \left(\nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})) - \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right)\right\|^2 \le \frac{\nu^2}{(1-\gamma)^6\zeta^2}\cdot\gamma^{2H}. \qquad (60)
$$

For the second part in (58), it holds that

$$
\begin{aligned}
&\left\|\left(\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top - \left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top\right) \nabla_{\boldsymbol{\lambda}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2 \\
=&\left\|\mathbb{E}\left[\sum_{h=H}^{\infty}\gamma^h\cdot\nabla_\lambda F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))_{s_h,b_h}\cdot\left(\sum_{h'=0}^{h}\nabla_{\boldsymbol{y}}\log y^{(t)}(b_{h'}|s_{h'})\right)\right]\right\|^2 \\
\le&\left(\sum_{h=H}^{\infty}\gamma^h\cdot\left(1+\frac{\nu}{1-\gamma}\right)\cdot(h+1)\cdot\frac{1}{\zeta}\right)^2 \\
\le&\left(1+\frac{\nu}{1-\gamma}\right)^2\cdot\frac{1}{\zeta^2}\cdot\left(\frac{(H+1)^2}{(1-\gamma)^2}+\frac{1}{(1-\gamma)^4}\right)\cdot\gamma^{2H} \\
=&\left(\frac{(H+1)^2}{(1-\gamma)^2\zeta^2}+\frac{2\nu(H+1)^2}{(1-\gamma)^3\zeta^2}+\frac{\nu^2(H+1)^2+1}{(1-\gamma)^4\zeta^2}+\frac{2\nu}{(1-\gamma)^5\zeta^2}+\frac{\nu^2}{(1-\gamma)^6}\right)\cdot\gamma^{2H}. \qquad (61)
\end{aligned}
$$

https://doi.org/10.52202/079017-2948

Combine (58), (60), and (61) we get

$$\begin{aligned}&\left\|\left[\nabla_{\boldsymbol{y}}\boldsymbol{\lambda}_H(\boldsymbol{y}^{(t)};\boldsymbol{x})\right]^\top \boldsymbol{r}^\star - \nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2\\ \leq& 2\left(\frac{(H+1)^2}{(1-\gamma)^2\zeta^2}+\frac{2\nu(H+1)^2}{(1-\gamma)^3\zeta^2}+\frac{\nu^2(H+1)^2+1}{(1-\gamma)^4\zeta^2}+\frac{2\nu}{(1-\gamma)^5\zeta^2}+\frac{2\nu^2}{(1-\gamma)^6\zeta^2}\right)\cdot\gamma^{2H}. \quad (62)\end{aligned}$$

Now combine Lemma C.17, (56), (57), and (62), we get

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\boldsymbol{y}}^{(t)}-\nabla_{\boldsymbol{y}}F^\nu(\boldsymbol{\lambda}(\boldsymbol{y}^{(t)};\boldsymbol{x}))\right\|^2\right]\leq\frac{\mathcal{C}_1}{K}+\mathcal{C}_2\cdot\gamma^{2H}.$$

□

D Nonconvex–Hidden-Strongly-Concave Optimization

In this section we generalize our results to the more general setting of any constrained min-max optimization problem of the form $\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}}$ when f is nonconvex–hidden-strongly-concave. In particular:

- In Theorem D.2 we prove the differentiability and Hölder continuity of the max function $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$ by utilizing the Hölder continuity of the maximizers w.r.t. to $\boldsymbol{x}$ (Theorem D.1).
- Finally, in Theorem D.3 we prove that Algorithm 3 (SGDMAX) [70, Algorithm 4] with an appropriate tuning converges to an ϵ-SP for nonconvex–hidden-concave functions.

We begin by stating the assumptions we make.

Assumption D.1. Let f be a function defined on $\mathcal{X}\times\mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets. L-Lipschitz continuous and ℓ-smooth.

Assumption D.2. Let c be a "1–1" mapping between $\mathcal{Y}$ and a compact convex set $\mathcal{U}$ parameterized by $\boldsymbol{x}\in\mathcal{X}$. Further, we assume that c and its inverse c^{-1} are L_c- and $L_{c^{-1}}$-Lipschitz continuous.

Assumption D.3. Let H be a nonconvex–strongly-concave reformulation of f (as in Assumption D.1) for a mapping c (as in Assumption D.2). We assume H to be L_H-Lipschitz continuous and ℓ_H-smooth.

Moving on, we can show that the maximizers $\boldsymbol{u}^\star(\cdot)$ are Hölder continuous w.r.t. to $\boldsymbol{x}$.

Theorem D.1 (Continuity of the maximizers)**.** *Let a function nonconvex–nonconcave function f, c, H as in Assumptions D.1 to D.3. We define $\boldsymbol{u}^\star(\boldsymbol{x}) \coloneqq \operatorname{argmax}_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{x})} H(\boldsymbol{x},\boldsymbol{u})$, then it is the case that*

$$\|\boldsymbol{u}^\star(\boldsymbol{x}_1) - \boldsymbol{u}^\star(\boldsymbol{x}_2)\| \le L_\star \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^{1/2}.$$

Where $L_\star = \frac{1}{2\nu}\left(2\ell_H\sqrt{\mathrm{Diam}_{\mathcal{X}}} + 2\sqrt{\nu(1+2\ell_H)L_c\mathrm{Diam}_{\mathcal{U}} + 2\nu L_c L_H}\right)$.

Proof. Consider any $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$, since $\boldsymbol{u}^\star$ is the maximizer, it holds that

$$\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right)^\top\left(\boldsymbol{u}_1 - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right) \le 0, \quad \forall \boldsymbol{u}_1\in\mathcal{U}(\boldsymbol{x}_1);$$
$$\nabla H\left(\boldsymbol{x}_2,\boldsymbol{u}^\star(\boldsymbol{x}_2)\right)^\top\left(\boldsymbol{u}_2 - \boldsymbol{u}^\star(\boldsymbol{x}_2)\right) \le 0, \quad \forall \boldsymbol{u}_2\in\mathcal{U}(\boldsymbol{x}_2).$$

We now consider $\overline{\boldsymbol{u}}$ that belong to the set $\overline{\mathcal{U}} = \mathcal{U}(\boldsymbol{x}_1)\cup\mathcal{U}(\boldsymbol{x}_2)$. We observe that due to the Lipschitz mapping, for every $\overline{\boldsymbol{u}}\in\overline{\mathcal{U}}$, there exist a $\boldsymbol{u}_1\in\mathcal{U}(\boldsymbol{x}_1)$ such that $\|\overline{\boldsymbol{u}} - \boldsymbol{u}_1\| \le L_c\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. Similar argument holds for $\boldsymbol{u}_2\in\mathcal{U}(\boldsymbol{x}_2)$. Therefore, from previous two inequalities, we have

$$\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right)^\top\left(\overline{\boldsymbol{u}} - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right) \le L_c L_H\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \quad \forall\overline{\boldsymbol{u}}\in\overline{\mathcal{U}};$$
$$\nabla H\left(\boldsymbol{x}_2,\boldsymbol{u}^\star(\boldsymbol{x}_2)\right)^\top\left(\overline{\boldsymbol{u}} - \boldsymbol{u}^\star(\boldsymbol{x}_2)\right) \le L_c L_H\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \quad \forall\overline{\boldsymbol{u}}\in\overline{\mathcal{U}}.$$

Where in the above inequalities we used the fact that $\nabla H(\boldsymbol{x},\boldsymbol{u}) \le L_H$. We plug in $\overline{\boldsymbol{u}}\leftarrow\boldsymbol{u}^\star(\boldsymbol{x}_2)$ and $\overline{\boldsymbol{u}}\leftarrow\boldsymbol{u}^\star(\boldsymbol{x}_1)$ accordingly,

$$\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right)^\top\left(\boldsymbol{u}^\star(\boldsymbol{x}_1) - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right) \le L_c L_H\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,$$
$$\nabla H\left(\boldsymbol{x}_2,\boldsymbol{u}^\star(\boldsymbol{x}_2)\right)^\top\left(\boldsymbol{u}^\star(\boldsymbol{x}_1) - \boldsymbol{u}^\star(\boldsymbol{x}_2)\right) \le L_c L_H\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

Adding the two inequalities results in,

$$\Big(\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right) - \nabla H\left(\boldsymbol{x}_2,\boldsymbol{u}^\star(\boldsymbol{x}_2)\right)\Big)^\top\left(\boldsymbol{u}^\star(\boldsymbol{x}_1) - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right) \le 2L_c L_H\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{63}$$

Since $H(\boldsymbol{x},\cdot)$ is ν-strongly concave in $\boldsymbol{u}$ for all $\boldsymbol{x}$, it holds that

$$\left(\boldsymbol{u}_1 - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right)^\top\Big(\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}_1\right) - \nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right)\Big) + \nu\|\boldsymbol{u}_1 - \boldsymbol{u}^\star(\boldsymbol{x}_1)\|^2 \le 0 \quad \forall\boldsymbol{u}_1\in\mathcal{U}(\boldsymbol{x}_1).$$

We again consider feasibility set $\overline{\boldsymbol{u}}\in\overline{\mathcal{U}}$. Since for every $\overline{\boldsymbol{u}}\in\overline{\mathcal{U}}$, there exists $\boldsymbol{u}_1\in\mathcal{U}(\boldsymbol{x}_1)$ s.t. $\|\overline{\boldsymbol{u}} - \boldsymbol{u}_1\| \le L_c\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. We have

$$\left(\boldsymbol{u}_1 + (\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}) - \boldsymbol{u}^\star(\boldsymbol{x}_1)\right)^\top\Big(\nabla H(\boldsymbol{x}_1,\boldsymbol{u}_1) + (\nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}}) - \nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}})) - \nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^\star(\boldsymbol{x}_1)\right)\Big)$$
$$+ \nu\|\boldsymbol{u}_1 + (\overline{\boldsymbol{u}} - \overline{\boldsymbol{u}}) - \boldsymbol{u}^\star(\boldsymbol{x}_1)\|^2 \le 0, \quad \forall\boldsymbol{u}_1\in\mathcal{U}(\boldsymbol{x}_1).$$

https://doi.org/10.52202/079017-2948

We rearrange the latter display into

$$
\begin{aligned}
&\left(\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right)^{\top}\Big(\nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}})-\nabla H(\boldsymbol{x}_1,\boldsymbol{u}^{\star}(\boldsymbol{x}_1))\Big)+\nu\|\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\|^2\\
&\leq \underbrace{-\left(\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right)^{\top}\Big(\nabla H(\boldsymbol{x}_1,\boldsymbol{u}_1)-\nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}})\Big)}_{\Omega_1}\\
&\quad\underbrace{-\left(\boldsymbol{u}_1-\overline{\boldsymbol{u}}\right)^{\top}\Big(\nabla H(\boldsymbol{x}_1,\boldsymbol{u}_1)-\nabla H(\boldsymbol{x}_1,\boldsymbol{u}^{\star}(\boldsymbol{x}_1))\Big)}_{\Omega_2}\\
&\quad\underbrace{-\nu\left\|\boldsymbol{u}_1-\overline{\boldsymbol{u}}\right\|^2-2\nu\left\langle\boldsymbol{u}_1-\overline{\boldsymbol{u}},\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right\rangle}_{\Omega_3}.
\end{aligned}
$$

We bound Ω_1, Ω_2, and Ω_3 separately.

- For Ω_1, we have

$$
\begin{aligned}
-\left(\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right)^{\top}\Big(\nabla H(\boldsymbol{x}_1,\boldsymbol{u}_1)-\nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}})\Big) &\leq \mathrm{Diam}_{\mathcal{U}}\cdot\ell_H\|\boldsymbol{u}_1-\overline{\boldsymbol{u}}\|\\
&\leq \mathrm{Diam}_{\mathcal{U}}\ell_H L_c\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|.
\end{aligned}
$$

- For Ω_2, it holds that

$$
-(\boldsymbol{u}_1-\overline{\boldsymbol{u}})^{\top}\left(\nabla H(\boldsymbol{x}_1,\boldsymbol{u}_1)-\nabla H(\boldsymbol{x}_1,\boldsymbol{u}^{\star}(\boldsymbol{x}_1))\right)\leq L_c\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|\cdot\ell_H\mathrm{Diam}_{\mathcal{U}}.
$$

- For Ω_3, since the first term is always non-positive, we only need to bound the second term:

$$
\begin{aligned}
-\left\langle\boldsymbol{u}_1-\overline{\boldsymbol{u}},\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right\rangle &\leq \left\|\boldsymbol{u}_1-\overline{\boldsymbol{u}}\right\|\left\|\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right\|\\
&\leq L_c\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|\cdot\mathrm{Diam}_{\mathcal{U}}.
\end{aligned}
$$

Combining Ω_1, Ω_2, and Ω_3, we conclude that

$$
\begin{aligned}
&\left(\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right)^{\top}\Big(\nabla H(\boldsymbol{x}_1,\overline{\boldsymbol{u}})-\nabla H(\boldsymbol{x}_1,\boldsymbol{u}^{\star}(\boldsymbol{x}_1))\Big)+\nu\|\overline{\boldsymbol{u}}-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\|^2\\
&\leq (1+2\ell_H)L_c\mathrm{Diam}_{\mathcal{U}}\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|. \qquad (64)
\end{aligned}
$$

Plugging in $\overline{\boldsymbol{u}}\leftarrow\boldsymbol{u}^{\star}(\boldsymbol{x}_2)$ in (64) and combine it with (63), we get

$$
\begin{aligned}
\nu\|\boldsymbol{u}^{\star}(\boldsymbol{x}_2)-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\|^2 &\leq \left(\boldsymbol{u}^{\star}(\boldsymbol{x}_2)-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\right)^{\top}\Big(\nabla H\left(\boldsymbol{x}_2,\boldsymbol{u}^{\star}(\boldsymbol{x}_2)\right)-\nabla H\left(\boldsymbol{x}_1,\boldsymbol{u}^{\star}(\boldsymbol{x}_2)\right)\Big)\\
&\quad+L''\|\boldsymbol{x}_1-\boldsymbol{x}_2\|\\
&\leq \ell_H\|\boldsymbol{u}^{\star}(\boldsymbol{x}_2)-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\|\|\boldsymbol{x}_1-\boldsymbol{x}_2\|+L''\|\boldsymbol{x}_1-\boldsymbol{x}_2\|.
\end{aligned}
$$

Where $L''=(1+2\ell_H)L_c\mathrm{Diam}_{\mathcal{U}}+2L_cL_H$.

Similarly to Lemma C.8, we can set $\lambda=\|\boldsymbol{u}^{\star}(\boldsymbol{x}_2)-\boldsymbol{u}^{\star}(\boldsymbol{x}_1)\|$ and $\chi=\|\boldsymbol{x}_1-\boldsymbol{x}_2\|$ and consider the inequality $\nu\lambda^2\leq\ell_H\lambda\chi+L''\chi$. We aim to find the solution of the form $\frac{\ell_H\chi+\sqrt{\chi(4\nu L''+\ell_H^2\chi)}}{2\nu}\leq c\sqrt{\chi}$. By setting $L_\star=c$ and solve for c gives

$$
L_\star=\frac{1}{2\nu}\left(2\ell_H\sqrt{\mathrm{Diam}_{\mathcal{X}}}+2\sqrt{\nu(1+2\ell_H)L_c\mathrm{Diam}_{\mathcal{U}}+2\nu L_cL_H}\right).
$$

□

Finally, we show that Φ is differentiable and Hölder-continuous.

Theorem D.2. *Let function Φ be $\Phi(\boldsymbol{x})\coloneqq\max_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{x})}\{H(\boldsymbol{x},\boldsymbol{u})\}$. Its gradient $\nabla\Phi$ is $(1/2,\ell_{1/2})$-Hölder continuous,*

$$
\|\nabla\Phi(\boldsymbol{x})-\nabla\Phi(\boldsymbol{x}')\|\leq\ell_{1/2}\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|^{\frac{1}{2}},
$$

where $\ell_{1/2}\coloneqq\left((1+L_{c^{-1}})\sqrt{\mathrm{Diam}_{\mathcal{X}}}+L_{c^{-1}}L_\star\right)\ell$.

Proof.

$$\begin{aligned}
\|\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{x}')\| &= \left\|\nabla f\left(\boldsymbol{x}, c^{-1}(\boldsymbol{u}^\star(\boldsymbol{x}); \boldsymbol{x})\right) - \nabla f\left(\boldsymbol{x}', c^{-1}(\boldsymbol{u}^\star(\boldsymbol{x}'); \boldsymbol{x}')\right)\right\| \\
&\leq \ell\left\|\boldsymbol{x} - \boldsymbol{x}'\right\| + \ell\left\|c^{-1}(\boldsymbol{u}^\star(\boldsymbol{x}); \boldsymbol{x}) - c^{-1}(\boldsymbol{u}^\star(\boldsymbol{x}'); \boldsymbol{x}')\right\| \\
&\leq \ell\left\|\boldsymbol{x} - \boldsymbol{x}'\right\| + \ell L_{c^{-1}}\left(\left\|\boldsymbol{u}^\star(\boldsymbol{x}) - \boldsymbol{u}^\star(\boldsymbol{x}')\right\| + \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|\right) \qquad (65) \\
&\leq (1 + L_{c^{-1}})\ell\left\|\boldsymbol{x} - \boldsymbol{x}'\right\| + L_{c^{-1}}L_\star\ell\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^{\frac{1}{2}} \qquad (66) \\
&\leq \left((1 + L_{c^{-1}})\sqrt{\mathrm{Diam}_{\mathcal{X}}} + L_{c^{-1}}L_\star\right)\ell\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^{\frac{1}{2}}.
\end{aligned}$$

Where

- in (65) we invoke the Lipschitz continuity of function $c^{-1}(\cdot)$;
- (66) follows from Theorem D.1.

□

Following, SGDMAX is presented where we assume a stochastic gradient oracle $G = (G_x, Gy) : \mathcal{X} \times \mathcal{Y} \times \Xi \to \mathbb{R}^d$ that is unbiased and has a bounded variance:

Algorithm 3 SGDMAX

Input: Initialization $\boldsymbol{x}^{(0)}$, stepsize η_x, T_x iterations, batch size M, oracle accuracy ζ.

1: **for** $t \leftarrow 1, 2, \ldots, T$ **do**
2: $\quad \boldsymbol{y}^{(t)} \leftarrow \mathsf{max\text{-}oracle}\left(f(\boldsymbol{x}^{(t)}, \cdot); \zeta\right)$
3: $\quad \hat{\boldsymbol{g}}^{(t)} \leftarrow \frac{1}{M}\sum_{j=1}^{M} G_x\left(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t)}, \xi_j^{(t)}\right)$
4: $\quad \boldsymbol{x}^{(t)} \leftarrow \mathrm{Proj}_{\mathcal{X}}\left(\boldsymbol{x}_i^{(t-1)} - \eta_x\hat{\boldsymbol{g}}^{(t)}\right)$
5: **end for**
6: $\boldsymbol{y}^{(T+1)} \leftarrow \mathsf{max\text{-}oracle}\left(f(\boldsymbol{x}^{(T)}, \cdot); \zeta\right)$

Finally, we can state the theorem of convergence to an ϵ-approximate saddle-point.

Theorem D.3. *Let a function f as the one in Theorem D.1. For a desired accuracy $\epsilon > 0$, Algorithm 3, (*SGDMAX*) with a tuning of $T_x = O\left(\frac{\ell_{1/2}^2}{\epsilon^3}\right)$, η_x, a* max-oracle *accuracy $\zeta = O\left(\frac{\nu\epsilon^2}{\ell^2}\right)$, and a batch size of $M = \max\left\{1, \frac{9\sigma^2}{2\epsilon^2}\right\}$ guarantees that there exists a $t^* \in [T]$, such that,*

$$-\nabla_x f\left(\boldsymbol{x}^{(t^\star)}, \boldsymbol{y}^{(t^\star+1)}\right)^\top\left(\boldsymbol{x}' - \boldsymbol{x}^{(t^\star)}\right) \leq \epsilon,\ \forall \boldsymbol{x}' \in \mathcal{X};$$

$$\nabla_y f\left(\boldsymbol{x}^{(t^\star)}, \boldsymbol{y}^{(t^\star+1)}\right)^\top\left(\boldsymbol{y}' - \boldsymbol{y}^{(t^\star)}\right) \leq \epsilon,\ \forall \boldsymbol{y}' \in \mathcal{Y}.$$

Further, the max-oracle*, of accuracy ζ, can be implemented by $T_y = \tilde{O}\left(\frac{L}{L_c^2\nu} + \frac{L\sigma^2}{L_c^4+\nu^2}\frac{1}{\zeta}\right)$ iterations of stochastic projected gradient ascent with a step size $\eta_y = \min\left\{\frac{2}{9L}, \frac{L_c^2\nu\zeta}{10L\sigma^2}\right\}$.*

Proof. The proof follows easily from the proof of projected gradient ascent in hidden-strongly-concave function found [43, Theorem 6] and Theorems B.1 and D.2. □

Remark 2. *It has been shown that when a function f enjoys a hidden-strongly-concave reformulation, it satisfies global the Proximal-PŁ condition (or equivalently, global KŁ condition) [43, 67]. While the equivalence between global KŁ condition and quadratic growth condition has been proven [10, 38] when f is concave, to the authors' best knowledge, this equivalence still remains unclear when f is nonconcave. This means that we cannot use [83] to prove the smoothness of the maximum function when the feasibility set of the maximizing variable is constrained.*

https://doi.org/10.52202/079017-2948

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As claimed in the abstraction and introduction, our main contribution is the design and analysis of a learning algorithm for Adversarial Team Markov Games with polynomial iteration and sample complexity (in the parameters of the underlying Markov Game); this is captured by Theorem 3.3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are captured in conclusion and future work for investigation. Moreover, all necessary assumptions have been properly cited.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proof of the main result and auxiliary claims and lemmas can be found in the appendix and are properly referenced. All the assumptions have been properly defined in the Preliminaries section and section 3, 4, e.g., see Assumption 4.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-2948

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is of theoretical nature about learning Nash equilibria in Markov Games. The authors believe that the paper is aligned with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is of theoretical nature about learning Nash equilibria in Markov Games. The authors do not foresee any societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

https://doi.org/10.52202/079017-2948

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2948