# EFFICIENTLY COMPUTING NASH EQUILIBRIA IN ADVERSARIAL TEAM MARKOV GAMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Computing Nash equilibrium policies is a central problem in multi-agent reinforcement learning that has received extensive attention both in theory and in practice. However, in light of computational intractability barriers in general-sum games, provable guarantees have been thus far either limited to fully competitive or cooperative scenarios, or impose strong assumptions that are difficult to meet in most practical applications. In this work, we depart from those prior results by investigating infinite-horizon *adversarial team Markov games*, a natural and well-motivated class of games in which a team of identically-interested players—in the absence of any explicit coordination or communication—is competing against an adversarial player. This setting allows for a unifying treatment of zero-sum Markov games and Markov potential games, and serves as a step to model more realistic strategic interactions that feature both competing and cooperative interests. Our main contribution is the first algorithm for computing stationary $\epsilon$-approximate Nash equilibria in adversarial team Markov games with computational complexity that is polynomial in all the natural parameters of the game, as well as $1/\epsilon$. The proposed algorithm is based on performing independent policy gradient steps for each player in the team, in tandem with best responses from the side of the adversary; in turn, the policy for the adversary is then obtained by solving a carefully constructed linear program. Our analysis leverages non-standard techniques to establish the KKT optimality conditions for a nonlinear program with nonconvex constraints, thereby leading to a natural interpretation of the induced Lagrange multipliers.

## 1 INTRODUCTION

Multi-agent reinforcement learning (MARL) offers a principled framework for analyzing competitive interactions in dynamic and stateful environments in which agents' actions affect both the state of the world and the rewards of the other players. Strategic reasoning in such complex multi-agent settings has been guided by game-theoretic principles, leading to many recent landmark results in benchmark domains in AI (Bowling et al., 2015; Silver et al., 2017; Vinyals et al., 2019; Moravčík et al., 2017; Brown & Sandholm, 2019; 2018; Brown et al., 2020; Perolat et al., 2022). Most of these remarkable advances rely on scalable and often decentralized algorithms for computing *Nash equilibria* (Nash, 1951)—a standard game-theoretic notion of rationality—in two-player zero-sum games.

Nevertheless, while single-agent RL has enjoyed rapid theoretical progress over the last few years (*e.g.*, see (Jin et al., 2018; Agarwal et al., 2020; Li et al., 2021; Luo et al., 2019; Sidford et al., 2018), and references therein), a comprehensive understanding of the multi-agent landscape still remains elusive. Indeed, provable guarantees for efficiently computing Nash equilibria have been thus far limited to either fully competitive settings, such as two-player zero-sum games (Daskalakis et al., 2020; Wei et al., 2021; Sayin et al., 2021; Cen et al., 2021; Sayin et al., 2020; Condon, 1993), or environments in which agents are striving to coordinate towards a common global objective (Claus & Boutilier, 1998; Wang & Sandholm, 2002; Leonardos et al., 2021; Ding et al., 2022; Zhang et al., 2021b; Chen et al., 2022; Maheshwari et al., 2022; Fox et al., 2022).

However, many real-world applications feature both shared and competing interests between the agents. Efficient algorithms for computing Nash equilibria in such settings are much more scarce,

and typically impose restrictive assumptions that are difficult to meet in most applications (Hu & Wellman, 2003; Bowling, 2000). In fact, even in *stateless* two-player (normal-form) games, computing approximate Nash equilibria is computationally intractable (Daskalakis et al., 2009; Rubinstein, 2017; Chen et al., 2009; Etessami & Yannakakis, 2010)—subject to well-believed complexity-theoretic assumptions. As a result, it is common to investigate equilibrium concepts that are more permissive than Nash equilibria, such as *coarse correlated equilibria (CCE)* (Aumann, 1974; Moulin & Vial, 1978). Unfortunately, recent work has established strong lower bounds for computing even approximate (stationary) CCEs in turn-based stochastic two-player games (Daskalakis et al., 2022; Jin et al., 2022). Those negative results raise a central question:

*Are there natural multi-agent environments incorporating both competing and shared interests for which we can establish efficient algorithms for computing (stationary) Nash equilibria?* (★)

Our work makes concrete progress in this fundamental direction. Specifically, we establish the first efficient algorithm leading to Nash equilibria in *adversarial team Markov games*, a well-motivated and natural multi-agent setting in which a team of agents with a common objective is facing a competing adversary.

### 1.1 Our Results

Before we state our main result, let us first briefly introduce the setting of adversarial team Markov games; a more precise description is deferred to Section 2.1. To address Question (★), we study an infinite-horizon Markov (stochastic) game with a finite state space $\mathcal{S}$ in which a team of agents $\mathcal{N}_A := [n]$ with a common objective function is competing against a *single* adversary with opposing interests. Every agent $k \in [n]$ has a (finite) set of available actions $\mathcal{A}_k$, while $\mathcal{B}$ represents the adversary's set of actions. We will also let $\gamma \in [0, 1)$ be the *discounting factor*. Our goal will be to compute an (approximate) Nash equilibrium; that is, a strategy profile so that no player can improve via a unilateral deviation (see Definition 2.1). In this context, our main contribution is the first polynomial time algorithm for computing Nash equilibria in adversarial team Markov games:

**Theorem 1.1** (Informal)**.** *There is an algorithm (*IPGMAX*) that, for any $\epsilon > 0$, computes an $\epsilon$-approximate stationary Nash equilibrium in adversarial team Markov games, and runs in time*

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

A few remarks are in order. First, our guarantee significantly extends and unifies prior results that only applied to either *two-player* zero-sum Markov games or to *Markov potential games*; both of those settings can be cast as special cases of adversarial team Markov games (see Section 2.3). Further, the complexity of our algorithm, specified in Theorem 1.1, scales only with $\sum_{k \in \mathcal{N}_A} |\mathcal{A}_k|$ instead of $\prod_{k \in \mathcal{N}_A} |\mathcal{A}_k|$, bypassing what is often referred to as the *curse of multi-agents* (Jin et al., 2021). Indeed, viewing the team as a single "meta-player" would induce an action space of size $\prod_{k \in \mathcal{N}_A} |\mathcal{A}_k|$, which is *exponential* even if each agent in the team has only two actions. In fact, our algorithm operates without requiring any (explicit) form of coordination or communication between the members of the team (beyond the structure of the game), a feature that has been motivated in practical applications (von Stengel & Koller, 1997). Namely, scenarios in which communication or coordination between the members of the team is either overly expensive, or even infeasible; for an in depth discussion regarding this point we refer to (Schulman & Vazirani, 2017).

### 1.2 Overview of Techniques

To establish Theorem 1.1, we propose a natural and decentraliezd algorithm we refer to as *Independent Policy GradientMax* (IPGMAX). IPGMAX works in turns. First, each player in the team performs one independent policy gradient step on their value function with an appropriately selected learning rate $\eta > 0$. In turn, the adversary best responds to the current policy of the team. This exchange is repeated for a sufficiently large number of iterations $T$. Finally, IPGMAX includes an auxiliary subroutine, namely `AdvNashPolicy()`, which computes the Nash policy of the adversary; this will be justified by Proposition 1.1 we describe below.

Our analysis builds on the techniques of Lin et al. (2020)—developed for the saddle-point problem $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$—for characterizing GDMAX. Specifically, GDMAX consists of performing gradient descent steps, specifically on the function $\phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. Lin et al. (2020) showed that GDMAX converges to a point $\big(\hat{\boldsymbol{x}}, \boldsymbol{y}^*(\hat{\boldsymbol{x}})\big)$ such that $\hat{\boldsymbol{x}}$ is an approximate first-order stationary point of the *Moreau envelope* (see Definition 3.1) of $\phi(\boldsymbol{x})$, while $\boldsymbol{y}^*(\hat{\boldsymbol{x}})$ is a best response to $\hat{\boldsymbol{x}}$. Now if $f(\boldsymbol{x},\cdot)$ is *strongly-concave*, one can show (by Dantzig's theorem) that $\big(\hat{\boldsymbol{x}}, \boldsymbol{y}^*(\boldsymbol{x})\big)$ is an approximate first-order stationary point of $f$. However, our setting introduces further challenges since the value function $V_{\boldsymbol{\rho}}(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})$ is nonconvex-nonconcave.

For this reason, we take a more refined approach. We first show in Proposition 3.1 that IPGMAX is guaranteed to converge to a policy profile $\big(\hat{\boldsymbol{\pi}}_{\text{team}}, \cdot\big)$ such that $\hat{\boldsymbol{\pi}}_{\text{team}}$ is an $\epsilon$-nearly stationary point of $\max_{\boldsymbol{\pi}_{\text{adv}}} V_{\rho}(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})$. Then, the next key step and the crux in the analysis is to show that $\hat{\boldsymbol{\pi}}_{\text{team}}$ *can be extended to an $O(\epsilon)$-approximate Nash equilibrium policy:*

**Proposition 1.1** (Informal)**.** If $\hat{\boldsymbol{\pi}}_{\text{team}}$ is an $\epsilon$-nearly stationary point of $\max_{\boldsymbol{\pi}_{\text{adv}}} V_{\rho}(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})$, there exists a policy for the adversary $\hat{\boldsymbol{\pi}}_{\text{adv}}$ so that $(\hat{\boldsymbol{\pi}}_{\text{team}}, \hat{\boldsymbol{\pi}}_{\text{adv}})$ is an $O(\epsilon)$-approximate Nash equilibrium.

In the special case of normal-form games, a similar extension theorem was recently obtained by Anonymous (2022). In particular, that result was derived by employing fairly standard linear programming techniques. In contrast, our more general setting introduces several new challenges, not least due to the nonconvexity-nonconcavity of the objective function.

Indeed, our analysis leverages more refined techniques stemming from nonlinear programming. More precisely, while we make use of standard policy gradient properties, similar to the single-agent MDP setting (Agarwal et al., 2021; Xiao, 2022), our analysis does not rely on the so-called *gradient-dominance* property (Bhandari & Russo, 2019), as that property does not hold in a team-wise sense. Instead, inspired by an alternative proof of Shapley's theorem (Shapley, 1953) for two-person zero-sum Markov games (Filar & Vrieze, 2012, Chapter 3), we employ mathematical programming. One of the central challenges is that the induced nonlinear program has a set of nonconvex constraints. As such, even the existence of (nonnegative) Lagrange multipliers satisfying the KKT conditions is not guaranteed, thereby necessitating more refined analysis techniques.

To this end, we employ the *Arrow-Hurwiz-Uzawa constraint qualification* (Theorem A.1) in order to establish that the local optima are contained in the set of KKT points (Corollary B.1). Then, we leverage the structure of adversarial team Markov games to characterize the induced Lagrange multipliers, showing that a subset of these can be used to establish Proposition 1.1; incidentally, this also leads to an efficient algorithm for computing a (near-)optimal policy of the adversary. Finally, we also remark that controlling the approximation error—an inherent barrier under policy gradient methods—in Proposition 1.1 turns out to be challenging. We bypass this issue by constructing "relaxed" programs that incorporate some imprecision in the constraints. A more detailed overview of our algorithm and the analysis is given in Section 3.

## 2 PRELIMINARIES

In this section, we introduce the relevant background and our notation. Section 2.1 describes adversarial team Markov games. Section 2.2 then defines some key concepts from multi-agent MDPs, while Section 2.3 describes a generalization of adversarial team Markov games, beyond identically-interested team players, allowing for a richer structure in the utilities of the team—namely, adversarial Markov potential games.

**Notation.** We let $[n] := \{1, \ldots, n\}$. 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 denote by $\|\cdot\| := \|\cdot\|_2$ the Euclidean norm. 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 game; precise statements are given in the Appendix. For the convenience of the reader, a comprehensive overview of our notation is given in A.3.

### 2.1 ADVERSARIAL TEAM MARKOV GAMES

An *adversarial team Markov game* (or an adversarial team *stochastic* game) is the Markov game extension of static, normal-form adversarial team games (Von Stengel & Koller, 1997). The game

is assumed to take place in an infinite-horizon discounted setting in which a team of identically-interested agents gain what the adversary loses. Formally, the game $\mathcal{G}$ is represented by a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \mathcal{A}, \mathcal{B}, r, \mathbb{P}, \gamma, \rho)$ whose components are defined as follows.

- $\mathcal{S}$ is a finite and nonempty set of *states*, with cardinality $S := |\mathcal{S}|$;
- $\mathcal{N}$ is the set of players, partitioned into a set of $n$ team agents $\mathcal{N}_A := [n]$ and a single *adversary*
- $\mathcal{A}_k$ is the action space of each player in the team $k \in [n]$, so that $\mathcal{A} := \bigtimes_{k\in[n]} \mathcal{A}_k$, while $\mathcal{B}$ is the action space of the adversary. We also let $A_k := |\mathcal{A}_k|$ and $B := |\mathcal{B}|$;[1]
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to (0, 1)$ is the (deterministic) instantaneous *reward function*[2] representing the (normalized) payoff of the adversary, so that for any $(s, \boldsymbol{a}, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$r(s, \boldsymbol{a}, b) + \sum_{k=1}^{n} r_k(s, \boldsymbol{a}, b) = 0, \tag{1}$$

  and for any $k \in [n]$,

$$r_k(s, \boldsymbol{a}, b) = r_{\text{team}}(s, \boldsymbol{a}, b). \tag{2}$$

- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is the *transition probability function*, so that $\mathbb{P}(s'|s, \boldsymbol{a}, b)$ denotes the probability of transitioning to state $s' \in \mathcal{S}$ when the current state is $s \in \mathcal{S}$ under the action profile $(\boldsymbol{a}, b) \in \mathcal{A} \times \mathcal{B}$;
- $\gamma \in [0, 1)$ is the *discount factor*; and
- $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ is the *initial state distribution* over the state space. We will assume that $\boldsymbol{\rho}$ is full-support, meaning that $\rho(s) > 0$ for all $s \in \mathcal{S}$.

In other words, an adversarial team Markov game is a subclass of general-sum infinite-horizon multi-agent discounted MDPs under the restriction that all but a single (adversarial) player have identical interests (see (2)), and the game is globally zero-sum—in the sense of (1). As we point out in Section 2.3, (2) can be relaxed in order to capture *(adversarial) Markov potential games* (Definition 2.2), without qualitatively altering our results.

### 2.2 Policies, Value Function, and Nash Equilibria

**Policies.** A *stationary*—that is, time-invariant—policy $\boldsymbol{\pi}_k$ for an agent $k$ is a function mapping a given state to a distribution over available actions, $\boldsymbol{\pi}_k : \mathcal{S} \ni s \mapsto \boldsymbol{\pi}_k(\cdot|s) \in \Delta(\mathcal{A}_k)$. We will say that $\boldsymbol{\pi}_k$ is *deterministic* if for every state there is some action that is selected with probability 1 under policy $\boldsymbol{\pi}_k$. For convenience, we will let $\Pi_{\text{team}} : \mathcal{S} \to \Delta(\mathcal{A})$ and $\Pi_{\text{adv}} : \mathcal{S} \to \Delta(\mathcal{B})$ denote the policy space for the team and the adversary respectively. We may also write $\Pi : \mathcal{S} \to \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ to denote the joint policy space of all agents.

**Direct Parametrization.** Throughout this paper we will assume that players employ *direct policy parametrization*. That is, for each player $k \in [n]$, we let $\mathcal{X}_k := \Delta(\mathcal{A}_k)^S$ and $\boldsymbol{\pi}_k = \boldsymbol{x}_k$ so that $x_{k,s,a} = \boldsymbol{\pi}_k(a|s)$. Similarly, for the adversary, we let $\mathcal{Y} := \Delta(\mathcal{B})^S$ and $\boldsymbol{\pi}_{\text{adv}} = \boldsymbol{y}$ so that $y_{s,a} = \boldsymbol{\pi}_{\text{adv}}(a|s)$. (Extending our results to other policy parameterizations, such as soft-max (Agarwal et al., 2021), is left for future work.)

**Value Function.** The *value function* $V_s : \Pi \ni (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n, \boldsymbol{\pi}_{\text{adv}}) \mapsto \mathbb{R}$ is defined as the expected cumulative discounted reward received by the adversary under the joint policy $(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}}) \in \Pi$ and the initial state $s \in \mathcal{S}$, where $\boldsymbol{\pi}_{\text{team}} := (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$. In symbols,

$$V_s(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}}) := \mathbb{E}_{(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})} \left[ \sum_{t=0}^{\infty} \gamma^t r(s^{(t)}, \boldsymbol{a}^{(t)}, b^{(t)}) \middle| s_0 = s \right], \tag{3}$$

where the expectation is taken over the trajectory distribution induced by $\boldsymbol{\pi}_{\text{team}}$ and $\boldsymbol{\pi}_{\text{adv}}$. When the initial state is drawn from a distribution $\boldsymbol{\rho}$, the value function takes the form $V_{\boldsymbol{\rho}}(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}}) := \mathbb{E}_{s\sim\boldsymbol{\rho}}\Big[V_s(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})\Big]$.

[1] To ease the notation, and without any essential loss of generality, we will assume throughout that the action space does not depend on the state.

[2] Assuming that the reward is positive is without any loss of generality (see Claim D.6).

**Nash Equilibrium.** Our main goal is to compute a joint policy profile that is an (approximate) *Nash equilibrium*, a standard equilibrium concept in game theory formalized below.

**Definition 2.1** (Nash equilibrium)**.** *A joint policy profile* $\left(\boldsymbol{\pi}^\star_{team}, \boldsymbol{\pi}^\star_{adv}\right) \in \Pi$ *is an* $\varepsilon$*-approximate Nash equilibrium, for* $\epsilon \geq 0$*, if*

$$\begin{cases} V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{team}, \boldsymbol{\pi}^\star_{adv}) & \leq V_{\boldsymbol{\rho}}((\boldsymbol{\pi}'_k, \boldsymbol{\pi}^\star_{-k}), \boldsymbol{\pi}^\star_{adv}) + \varepsilon, \quad \forall k \in [n], \forall \boldsymbol{\pi}'_k \in \Pi_k, \\ V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{team}, \boldsymbol{\pi}^\star_{adv}) & \geq V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{team}, \boldsymbol{\pi}'_{adv}) - \varepsilon, \qquad \forall \boldsymbol{\pi}'_{adv} \in \Pi_{adv}. \end{cases}$$

That is, a joint policy profile is an (approximate) Nash equilibrium if no unilateral deviation from a player can result in a non-negligible—more than additive $\epsilon$—improvement for that player. Nash equilibria always exist in multi-agent stochastic games (Fink, 1964); our main result implies an (efficient) constructive proof of that fact for the special case of adversarial team Markov games.

### 2.3 Adversarial Markov Potential Games

A recent line of work has extended the fundamental class of potential normal-form games (Monderer & Shapley, 1996) to *Markov potential games* (Marden, 2012; Macua et al., 2018; Leonardos et al., 2021; Ding et al., 2022; Zhang et al., 2021b; Chen et al., 2022; Maheshwari et al., 2022; Fox et al., 2022). Importantly, our results readily carry over even if players in the team are not necessarily identically interested, but instead, there is some underlying potential function for the team; we will refer to such games as *adversarial Markov potential games*, formally introduced below.

**Definition 2.2.** *An adversarial Markov potential game* $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \mathcal{A}, \mathcal{B}, \{r_k\}_{k\in[n]}, \mathbb{P}, \gamma, \rho)$ *is a multi-agent discounted MDP that shares all the properties of adversarial team Markov games (Section 2.1), with the exception that (2) is relaxed in that there exists a potential function* $\Phi_s$*,* $\forall s \in \mathcal{S}$*, such that for any* $\boldsymbol{\pi}_{adv} \in \Pi_{adv}$*,*

$$\Phi_s(\boldsymbol{\pi}_k, \boldsymbol{\pi}_{-k}; \boldsymbol{\pi}_{adv}) - \Phi_s(\boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}; \boldsymbol{\pi}_{adv}) = V_{k,s}(\boldsymbol{\pi}_k, \boldsymbol{\pi}_{-k}; \boldsymbol{\pi}_{adv}) - V_{k,s}(\boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}; \boldsymbol{\pi}_{adv}),$$

*for every agent* $k \in [n]$*, every state* $s \in \mathcal{S}$*, and all policies* $\boldsymbol{\pi}_k, \boldsymbol{\pi}_{k'} \in \Pi_k$ *and* $\boldsymbol{\pi}_{-k} \in \Pi_{-k}$*.*

## 3 Main Result

In this section, we sketch the main pieces required in the proof of our main result, Theorem 1.1. We begin by describing our algorithm in Section 3.1. Next, in Section 3.2, we characterize the strategy $\hat{\boldsymbol{x}} \in \mathcal{X}$ for the team returned by IPGMAX, while Section 3.3 completes the proof by establishing that $\hat{\boldsymbol{x}}$ can be efficiently extended to an approximate Nash equilibrium. The formal proof of Theorem 1.1 is deferred to the Appendix.

### 3.1 Our Algorithm

In this subsection, we describe in detail our algorithm for computing $\epsilon$-approximate Nash equilibria, IPGMAX, in adversarial team Markov games (Algorithm 1). IPGMAX takes as input a precision parameter $\epsilon > 0$ (Line 1) and an initial strategy for the team $(\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_n^{(0)}) = \boldsymbol{x}^{(0)} \in \mathcal{X} := \bigtimes_{k=1}^n \mathcal{X}_k$ (Line 2). The algorithm then proceeds in two phases:

- In the first phase the team players are performing independent policy gradient steps (Line 7) with learning rate $\eta$, as defined in Line 3, while the adversary is then best responding to their strategy (Line 6). This process is repeated for $T$ iterations, with $T$ as defined in Line 4. We note that $\mathrm{Proj}\,(\cdot)$ in Line 7 stands for the Euclidean projection, ensuring that each player selects a valid strategy. The first phase is completed in Line 9, where we set $\hat{\boldsymbol{x}}$ according to the iterate at time $t^\star$, for some $0 \leq t^\star \leq T-1$. As we explain in Section 3.2, selecting uniformly at random is a practical and theoretically sound way of setting $t^\star$.
- In the second phase we are fixing the strategy of the team $\hat{\boldsymbol{x}} \in \mathcal{X}$, and the main goal is to determine a strategy $\hat{\boldsymbol{y}} \in \mathcal{Y}$ so that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $O(\epsilon)$-approximate Nash equilibrium. This is accomplished in the subroutine `AdvNashPolicy`$(\hat{\boldsymbol{x}})$, which consists of solving a linear program—from the perspective of the adversary—that has polynomial size. Our analysis of the second phase of IPGMAX can be found in Section 3.3.

**Algorithm 1** Independent Policy GradientMax (IPGMAX)

1: Precision $\epsilon > 0$
2: Initial Strategy $\boldsymbol{x}^{(0)} \in \mathcal{X}$
3: Learning rate $\eta := \frac{\epsilon^2(1-\gamma)^9}{32S^4D^2\left(\sum_{k=1}^n A_k+B\right)^3}$
4: Number of iterations $T := \frac{512S^8D^4\left(\sum_{k=1}^n A_k+B\right)^4}{\epsilon^4(1-\gamma)^{12}}$
5: **for** $t \leftarrow 1, 2, \ldots, T$ **do**
6: $\quad \boldsymbol{y}^{(t)} \leftarrow \arg\max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}\left(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}\right)$
7: $\quad \boldsymbol{x}_k^{(t)} \leftarrow \mathrm{Proj}_{\mathcal{X}_k}\left(\boldsymbol{x}_k^{(t-1)} - \eta\nabla_{\boldsymbol{x}_k} V_{\boldsymbol{\rho}}\left(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t)}\right)\right)$ ▷ for all agents $i \in [n]$
8: **end for**
9: $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x}^{(t^\star)}$
10: $\hat{\boldsymbol{y}} \leftarrow \texttt{AdvNashPolicy}(\hat{\boldsymbol{x}})$ ▷ defined in Algorithm 2
11: **return** $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$

**Remark 1** (Approx. & exact best-response)**.** *Our main theoretical result applies directly even in the case that each iterate $\boldsymbol{y}^{(t)}$ is an approximate maximizer of $V_\rho(\boldsymbol{x}^{(t-1)}, \cdot)$. This means that Line 6 can be implemented by an efficient reinforcement learning algorithm (e.g., q-learning (Jin et al., 2018)) as long as the approximation error is small enough. Nevertheless, there are algorithms (Ye, 2011) that can compute exact maximizers in polynomial time. For a longer discussion in Remark 4.*

### 3.2 Analyzing Independent Policy GradientMax

In this subsection, we establish that IPGMAX finds an $\epsilon$-nearly stationary point $\hat{\boldsymbol{x}}$ of $\phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ in a number of iterations $T$ that is polynomial in the natural parameters of the game, as well as $1/\epsilon$; this is formalized in Proposition 3.1.

First, we note the by-now standard property that the value function $V_{\boldsymbol{\rho}}$ is $L$-Lipschitz continuous and $\ell$-smooth, where $L := \frac{\sqrt{\sum_{k=1}^n A_k+B}}{(1-\gamma)^2}$ and $\ell := \frac{2\left(\sum_{k=1}^n A_k+B\right)}{(1-\gamma)^3}$ (Lemma C.1). An important observation for the analysis is that IPGMAX is essentially performing gradient descent steps on $\phi(\boldsymbol{x})$. However, the challenge is that $\phi(\boldsymbol{x})$ is not necessarily differentiable; thus, our analysis relies on the *Moreau envelope* of $\phi$, defined as follows.

**Definition 3.1** (Moreau Envelope)**.** *Let $\phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$. For any $0 < \lambda < \frac{1}{\ell}$ the Moreau envelope $\phi_\lambda$ of $\phi$ is defined as*

$$\phi_\lambda(\boldsymbol{x}) := \min_{\boldsymbol{x}'\in\mathcal{X}} \left\{\phi(\boldsymbol{x}') + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2\right\}. \tag{4}$$

*We will let $\lambda := \frac{1}{2\ell}$.*

Crucially, the Moreau envelope $\phi_\lambda$, as introduced in (4), is $\ell$-strongly convex; this follows immediately from the fact that $\phi(\boldsymbol{x})$ is $\ell$-*weakly convex*, in the sense that $\phi(\boldsymbol{x}) + \frac{\ell}{2}\|\boldsymbol{x}\|^2$ is convex (see Lemma A.1). A related notion that will be useful to measure the progress of IPGMAX is the *proximal mapping* of a function $f$, defined as $\mathrm{prox}_f : \mathcal{X} \ni \boldsymbol{x} \mapsto \arg\min_{x'\in\mathcal{X}} \left\{f(\boldsymbol{x}') + \frac{1}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|^2\right\}$; the proximal of $\phi/(2\ell)$ is well-defined since $\phi$ is $\ell$-weakly convex (Proposition A.1). We are now ready to state the convergence guarantee of IPGMAX.

**Proposition 3.1.** Consider any $\epsilon > 0$. If $\eta = 2\epsilon^2(1-\gamma)$ and $T = \frac{(1-\gamma)^4}{8\epsilon^4(\sum_{k=1}^n A_k+B)^2}$, there exists an iterate $t^\star$, with $0 \le t^\star \le T-1$, such that $\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|_2 \le \epsilon$, where $\tilde{\boldsymbol{x}}^{(t^\star)} := \mathrm{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t^\star)})$.

The proof relies on the techniques of Lin et al. (2020), and it is deferred to Appendix C. The main takeaway is that $O(1/\epsilon^4)$ iterations suffice in order to reach an $\epsilon$-nearly stationary point of $\phi$—in the sense that it is $\epsilon$-far in $\ell_2$ distance from its proximal point. A delicate issue here is that Proposition 3.1 only gives a best-iterate guarantee, and identifying that iterate might introduce a substantial computational overhead. To address this, we also show in Corollary C.1 that by randomly selecting $\lceil\log(1/\delta)\rceil$ iterates over the $T$ repetitions of IPGMAX, we are guaranteed to recover an $\epsilon$-nearly stationary point with probability at least $1-\delta$, for any $\delta > 0$.

### 3.3 EFFICIENT EXTENSION TO NASH EQUILIBRIA

In this subsection, we establish that any $\epsilon$-nearly stationary point $\hat{\boldsymbol{x}}$ of $\phi$, can be *extended* to an $O(\epsilon)$-approximate Nash equilibrium $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ for any adversarial team Markov game, where $\hat{\boldsymbol{y}} \in \mathcal{Y}$ is the strategy for the adversary. Further, we show that $\hat{\boldsymbol{y}}$ can be computed in polynomial time through a carefully constructed linear program. This "extendibility" argument significantly extends a seminal characterization of Von Stengel & Koller (1997), and it is the crux in the analysis towards establishing our main result, Theorem 1.1.

To this end, the techniques we leverage are more involved compared to (Von Stengel & Koller, 1997), and revolve around nonlinear programming. Specifically, in the spirit of (Filar & Vrieze, 2012, Chapter 3), the starting point of our argument is the following nonlinear program with variables $(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{X} \times \mathbb{R}^S$:

$$
\text{(Q-NLP)} \quad
\begin{aligned}
&\min \sum_{s \in \mathcal{S}} \rho(s) v(s) + \ell \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \\
&\text{s.t. } r(s, \boldsymbol{x}, b) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' | s, \boldsymbol{x}, b) v(s') \leq v(s), \quad \forall (s, b) \in \mathcal{S} \times \mathcal{B}; && (Q1) \\
&\qquad \boldsymbol{x}_{k,s}^\top \mathbf{1} = 1, \quad \forall (k, s) \in [n] \times \mathcal{S}; \text{ and} && (Q2) \\
&\qquad x_{k,s,a} \geq 0, \quad \forall k \in [n], (s, a) \in \mathcal{S} \times \mathcal{A}_k. && (Q3)
\end{aligned}
$$

Here, we have overloaded notation so that $r(s, \boldsymbol{x}, b) := \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{x}_s}[r(s, \boldsymbol{a}, b]$ and $\mathbb{P}(s' | s, \boldsymbol{x}, b)) := \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{x}_s}[\mathbb{P}(s' | s, \boldsymbol{a}, b)]$. For a fixed strategy $\boldsymbol{x} \in \mathcal{X}$ for the team, this program describes the (discounted) MDP faced by the adversary. A central challenge in this formulation lies in the nonconvexity-nonconcavity of the constraint functions, witnessed by the multilinear constraint ($Q1$). Importantly, unlike standard MDP formulations, we have incorporated a quadratic regularizer in the objective function; this term ensures the following property.

**Proposition 3.2.** For any fixed $\boldsymbol{x} \in \mathcal{X}$, there is a unique optimal solution $\boldsymbol{v}^\star$ to (Q-NLP). Further, if $\tilde{\boldsymbol{x}} := \text{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$, and $\tilde{\boldsymbol{v}} \in \mathbb{R}^S$ is the corresponding optimal, then $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is the global optimum of (Q-NLP).

The uniqueness of the associated value vector is a consequence of Bellman's optimality equation, while the optimality of the proximal point follows by realizing that (Q-NLP) is an equivalent formulation of the proximal mapping. These steps are formalized in Appendix B.2. Having established the optimality of $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$, the next step is to show the existence of nonnegative Lagrange multipliers satisfying the KKT conditions (recall Definition A.2); this is non-trivial due to the nonconvexity of the feasibility set of (Q-NLP).

To do so, we leverage the so-called *Arrow-Hurwicz-Uzawa constraint qualification* (Theorem A.1)—a form of "regularity condition" for a nonconvex program. Indeed, in Lemma B.3 we show that any feasible point of (Q-NLP) satisfies that constraint qualification, thereby implying the existence of nonnegative Lagrange multipliers satisfying the KKT conditions for any local optimum (Corollary B.1), and in particular for $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$:

**Proposition 3.3.** There exist nonnegative Lagrange multipliers satisfying the KKT conditions at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$.

Now the upshot is that a subset of those Lagrange multipliers $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^{S \times B}$ can be used to establish the extendability of $\hat{\boldsymbol{x}}$ to a Nash equilibrium. Indeed, our next step makes this explicit: We construct a linear program whose sole goal is to identify such multipliers, which in turn will allow us to efficiently compute an admissible strategy for the adversary $\hat{\boldsymbol{y}}$. However, determining $\tilde{\boldsymbol{\lambda}}$ exactly seems too ambitious. For one, IPGMAX only granted us access to $\hat{\boldsymbol{x}}$, but not to $\tilde{\boldsymbol{x}}$. On the other hand, the Lagrange multipliers $\tilde{\boldsymbol{\lambda}}$ are induced by $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$. To address this, the constraints of our linear program are phrased in terms of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$, instead of $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$, while to guarantee feasibility we appropriately relax all the constraints of the linear program; this relaxation does not introduce a large error since $\|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| \leq \epsilon$ (Proposition 3.1), and the underlying constraint functions are Lipschitz continuous—with constants that depend favorably on the game $\mathcal{G}$; we formalize that in Lemma B.4. This leads to our main theorem, summarized below (see Theorem B.1 for a precise statement).

**Theorem 3.1.** *Let $\hat{\boldsymbol{x}}$ be an $\epsilon$-nearly stationary point of $\phi$. There exist a linear program,* ($\mathrm{LP}_{\mathrm{adv}}$)*, such that:*

*(i) It has size that is polynomial in $\mathcal{G}$, and all the coefficients depend on the (single-agent) MDP faced by the adversary when the team is playing a fixed strategy $\hat{\boldsymbol{x}}$; and*

*(ii) It is always feasible, and any solution induces a strategy $\hat{\boldsymbol{y}}$ such that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $O(\epsilon)$-approximate Nash equilibrium.*

The proof of this theorem carefully leverages the structure of adversarial team Markov games, along with the KKT conditions we previously established in Proposition 3.3. The algorithm for computing the policy for the adversary is summarized in Algorithm 2 of Appendix B. A delicate issue with Theorem 3.1, and in particular with the solution of ($\mathrm{LP}_{\mathrm{adv}}$), is whether one can indeed *efficiently simulate* the environment faced by the adversary. Indeed, in the absence of any structure, determining the coefficients of the linear program could scale exponentially with the number of players; this is related to a well-known issue in computational game theory, revolving around the exponential blow-up of the input space as the number of players increases (Papadimitriou & Roughgarden, 2008). As is standard, we bypass this by assuming access to natural oracles that ensure we can efficiently simulate the environment faced by the adversary (Remark 3).

## 4 Further Related Work

In this section, we highlight certain key lines of work that relate to our results in the context of adversarial team Markov games. We stress that the related literature on multi-agent reinforcement learning (MARL) is too vast to even attempt to faithfully cover here. For some excellent recent overviews of the area, we refer the interested reader to (Yang & Wang, 2020; Zhang et al., 2021a) and the extensive lists of references therein.

**Team Games.** The study of team games has been a prolific topic of research in economic theory and group decision theory for many decades; see, *e.g.*, (Marschak, 1955; Groves, 1973; Radner, 1962; Ho & Chu, 1972). A more modern key reference point to our work is the seminal paper of Von Stengel & Koller (1997) that introduced the notion of *team-maxmin equilibrium (TME)* in the context of normal-form games. A TME profile is a mixed strategy for each team member so that the minimal expected team payoff over all possible responses of the adversary—who potentially knows the play of the team—is the maximum possible. While TME's enjoy a number of compelling properties, being the optimal equilibria for the team given the lack of coordination, they suffer from computational intractability even in 3-player team games (Hansen et al., 2008; Borgs et al., 2010).[3] Nevertheless, practical algorithms have been recently proposed and studied for computing them in multiplayer games (Zhang & An, 2020a;b; Basilico et al., 2017). It is worth pointing out that team equilibria are also useful for extensive-form two-player zero-sum games where one of the players has *imperfect recall* (Piccione & Rubinstein, 1997).

The intractability of TME has motivated the study of a relaxed equilibrium concept that incorporates a *correlation device* (Farina et al., 2018; Celli & Gatti, 2018; Basilico et al., 2017; Zhang & An, 2020b; Zhang & Sandholm, 2021; Zhang et al., 2022; Carminati et al., 2022); namely, *TMECor*. In TMECor players are allowed to select *correlated strategies*. Despite the many compelling aspects of TMECor as a solution concept in team games, even *ex ante* coordination or correlated randomization—beyond the structure of the game itself—can be overly expensive or even infeasible in many applications (Von Stengel & Koller, 1997). Further, even TMECor is NP-hard to compute (in the worst-case) for *imperfect-information* extensive-form games (EFGs) (Chu & Halpern, 2001), although fixed-parameter-tractable (FPT) algorithms have recently emerged for natural classes of EFGs (Zhang & Sandholm, 2021; Zhang et al., 2022).

On the other hand, the computational aspects of the standard Nash equilibrium (NE) in adversarial team games is not well-understood, even in normal-form games. In fact, it is worth pointing out that Von Neumann's celebrated *minimax theorem* (von Neumann & Morgenstern, 2007) does not apply in team games, rendering traditional techniques employed in two-player zero-sum games of

[3] Hansen et al. (2008); Borgs et al. (2010) establish FNP-hardness and inapproximability for general 3-player games, but their argument readily applies to 3-player team games as well.

little use. Indeed, Schulman & Vazirani (2017) provided a precise characterization of the *duality gap* between the two teams based on the natural parameters of the problem, while Kalogiannis et al. (2021) showed that standard no-regret learning dynamics such as gradient descent and optimistic Hedge could fail to stabilize to mixed NE even in binary-action adversarial team games. Finally, we should also point out that although from a complexity-theoretic standpoint our main result (Theorem 1.1) establishes a *fully polynomial time approximate scheme (*$\mathsf{FPTAS}$*)*, since the dependence on the approximation error $\epsilon$ is $\mathsf{poly}(1/\epsilon)$, an improvement to $\mathsf{poly}(\log(1/\epsilon))$ is precluded even in normal-form games unless $\mathsf{CLS} \subseteq \mathsf{P}$ (an unlikely event); this follows as adversarial team games capture potential games (Kalogiannis et al., 2021), wherein computing mixed Nash equilibria is known to be complete for the class $\mathsf{CLS} = \mathsf{PPAD} \cap \mathsf{PLS}$ (Babichenko & Rubinstein, 2021).

**Multi-agent RL.** Computing Nash equilibria has been a central endeavor in multi-agent RL. While some algorithms have been proposed, perhaps most notably the Nash-Q algorithm (Hu & Wellman, 1998; 2003), convergence to Nash equilibria is only guaranteed under severe restrictions on the game. More broadly, the long-term behavior of independent policy gradient methods (Schulman et al., 2015) is still not well-understood. Before all else, from the impossibility result of Hart & Mas-Colell, universal convergence to Nash equilibria is precluded even for normal-form games; this is aligned with the computational intractability ($\mathsf{PPAD}$-completeness) of Nash equilibria even in two-player general-sum games (Daskalakis et al., 2009; Chen et al., 2009). Surprisingly, recent work has also established hardness results in turn-based stochastic games, rendering even the weaker notion of (stationary) CCEs intractable (Daskalakis et al., 2022; Jin et al., 2022).

As a result, the existing literature has inevitably focused on specific classes of games, such as Markov potential games (Leonardos et al., 2021; Ding et al., 2022; Zhang et al., 2021b; Chen et al., 2022; Maheshwari et al., 2022; Fox et al., 2022) or two-player zero-sum Markov games (Daskalakis et al., 2020; Wei et al., 2021; Sayin et al., 2021; Cen et al., 2021; Sayin et al., 2020). As we pointed out earlier, adversarial Markov team games can unify and extend those settings (Section 2.3). More broadly, identifying multi-agent settings for which Nash equilibria are provably efficiently computable is recognized as an important open problem in the literature (see, *e.g.*, (Daskalakis et al., 2020)), boiling down to one of the main research question of this paper (Question (★)). We also remark that certain guarantees for convergence to Nash equilibria have been recently obtained in a class of symmetric games (Emmons et al., 2022)—including symmetric team games. Finally, weaker solution concepts relaxing either the Markovian or the stationarity properties have also recently attracted attention (Daskalakis et al., 2022; Jin et al., 2021).

## 5 Conclusions

Our main contribution in this paper is the first polynomial algorithm for computing (stationary) Nash equilibria in adversarial team Markov games, an important class of games in which a team of uncoordinated but identically-interested players is competing against an adversarial player. We argued that this setting serves as a step towards modeling more realistic multi-agent applications that feature both competing and cooperative interests. There exists an array of possible future directions of research. One caveat of our main algorithm (IPGMAX) is that it requires a separate subroutine for computing the optimal policy of the adversary. It is plausible that a carefully designed two-timescale policy gradient method can efficiently reach a Nash equilibrium, effectively yielding fully model-free algorithms for solving adversarial team Markov games. Another approach towards an algorithm that would not require prior knowledge of the environment could be developed by utilizing techniques from the literature of constrained Markov decision processes in order to compute the policy of the adversary, given that IPGMAX has converged to a stationary point; ultimately, eliminating the need of solving a linear program and assuming knowledge of the environment. Exploring different solution concepts— beyond Nash equilibria—could also be a fruitful avenue for the future. Indeed, allowing some limited form of correlation between the players in the team could lead to more efficient algorithms; whether that form of coordination is justified (arguably) depends to a large extent on the application at hand. In such a scenario, one could be able to further argue about the efficiency of Nash equilibria against other solution concepts in this class of games. Finally, returning to Question (★), a more ambitious agenda revolves around understanding the fundamental structure of games for which computing Nash equilibria is provably computationally tractable.

## REFERENCES


Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66. PMLR, 2020.

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

Anonymous. Algorithms and complexity for computing nash equilibria in adversarial team games. *Personal Communication*, 2022.

Kenneth J Arrow, Leonid Hurwicz, and Hirofumi Uzawa. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly*, 8(2):175–191, 1961.

Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

Yakov Babichenko and Aviad Rubinstein. Settling the complexity of nash equilibrium in congestion games. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, 2021*, pp. 1426–1437. ACM, 2021. doi: 10.1145/3406325.3451039.

Nicola Basilico, Andrea Celli, Giuseppe De Nittis, and Nicola Gatti. Team-maxmin equilibrium: Efficiency bounds and algorithms. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017*, pp. 356–362. AAAI Press, 2017.

MS Bazaraa, JJ Goode, and CM Shetty. Constraint qualifications revisited. *Management Science*, 18(9):567–573, 1972.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Christian Borgs, Jennifer T. Chayes, Nicole Immorlica, Adam Tauman Kalai, Vahab S. Mirrokni, and Christos H. Papadimitriou. The myth of the folk theorem. *Games Econ. Behav.*, 70(1):34–43, 2010. doi: 10.1016/j.geb.2009.04.016.

Michael Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 89–94. Morgan Kaufmann, 2000.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015. doi: 10.1126/science.1259433.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019. doi: 10.1126/science.aay2400.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng 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.

Luca Carminati, Federico Cacciamani, Marco Ciccone, and Nicola 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*, pp. 2638–2657. PMLR, 2022.

Andrea Celli and Nicola Gatti. Computational results for extensive-form adversarial team games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Shicong Cen, Yuting Wei, and Yuejie 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*, pp. 27952–27964, 2021.

Dingyang Chen, Qi Zhang, and Thinh 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.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *J. ACM*, 56(3):14:1–14:57, 2009. doi: 10.1145/1516512.1516516.

Francis C. Chu and Joseph 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. doi: 10.1007/s001820100066.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98*, pp. 746–752. AAAI Press / The MIT Press, 1998.

Anne Condon. On algorithms for simple stochastic games. In *Advances in Computational Complexity Theory, volume 13 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 51–73. American Mathematical Society, 1993.

Constantinos Daskalakis, Alex Fabrikant, and Christos H. Papadimitriou. The game world is flat: The complexity of nash equilibria in succinct games. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006*, volume 4051 of *Lecture Notes in Computer Science*, pp. 513–524. Springer, 2006. doi: 10.1007/11786986_45.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM J. Comput.*, 39(1):195–259, 2009. doi: 10.1137/070699652.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540, 2020.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *CoRR*, abs/2204.03991, 2022. doi: 10.48550/arXiv.2204.03991.

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

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo 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.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart 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*, pp. 5924–5943. PMLR, 2022.

Kousha Etessami and Mihalis Yannakakis. On the complexity of nash equilibria and other fixed points. *SIAM J. Comput.*, 39(6):2531–2597, 2010. doi: 10.1137/080720826.

Gabriele Farina, Andrea Celli, Nicola Gatti, and Tuomas 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*, pp. 9661–9671, 2018.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

A. M. Fink. Equilibrium in a stochastic $n$-person game. *Journal of Science of the Hiroshima University, Series A-I (Mathematics)*, 28(1):89 – 93, 1964. doi: 10.32917/hmj/1206139508.

Roy Fox, Stephen M. McAleer, Will Overman, and Ioannis 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*, pp. 4414–4425. PMLR, 2022.

Giorgio Giorgi et al. A guided tour in constraint qualifications for nonlinear programming under differentiability assumptions. Technical report, University of Pavia, Department of Economics and Management, 2018.

Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973.

Kristoffer Arnsfelt Hansen, Thomas Dueholm Hansen, Peter Bro Miltersen, and Troels Bjerre Sørensen. Approximability and parameterized complexity of minmax values. In *International Workshop on Internet and Network Economics*, pp. 684–695. Springer, 2008.

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

Yu-Chi Ho and K'ai-Ching Chu. Team decision theory and information structures in optimal control problems–part i. *IEEE Transactions on Automatic Control*, 17(1):15–22, 1972. doi: 10.1109/TAC.1972.1099850.

Wassily Hoeffding and J. Wolfowitz. Distinguishability of Sets of Distributions. *The Annals of Mathematical Statistics*, 29(3):700 – 718, 1958.

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

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

Wan Huang and Bernhard von Stengel. Computing an extensive-form correlated equilibrium in polynomial time. In *Internet and Network Economics, 4th International Workshop, WINE 2008*, volume 5385 of *Lecture Notes in Computer Science*, pp. 506–513. Springer, 2008. doi: 10.1007/978-3-540-92185-1_56.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pp. 4868–4878, 2018.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4880–4889. PMLR, 2020.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*, 2022.

Fivos Kalogiannis, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ioannis Panageas. Teamwork makes von neumann work: Min-max optimization in two-team zero-sum games. *arXiv preprint arXiv:2111.04178*, 2021.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Yuanzhi Li, Ruosong Wang, and Lin F. Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021*, pp. 965–976. IEEE, 2021. doi: 10.1109/FOCS52979.2021.00097.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. Independent and decentralized learning in markov potential games, 2022.

Olvi L Mangasarian. *Nonlinear programming*. SIAM, 1994.

Jason R Marden. State based potential games. *Automatica*, 48(12):3075–3088, 2012.

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

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

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960.

H. Moulin and J.-P. Vial. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4):201–221, 1978.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. Technical report, 1985.

John Nash. Non-cooperative games. *Annals of mathematics*, pp. 286–295, 1951.

Christos H. Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *J. ACM*, 55(3):14:1–14:29, 2008. doi: 10.1145/1379759.1379762.

Julien Perolat, Bart de Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning, 2022.

Michele Piccione and Ariel Rubinstein. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, 20(1):3–24, 1997. doi: https://doi.org/10.1006/game.1997.0536.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

R. Radner. Team Decision Problems. *The Annals of Mathematical Statistics*, 33(3):857 – 881, 1962. doi: 10.1214/aoms/1177704455.

R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. *SIGecom Exch.*, 15(2):45–49, 2017. doi: 10.1145/3055589.3055596.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized q-learning in zero-sum markov games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18320–18334. Curran Associates, Inc., 2021.

Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv preprint arXiv:2010.04223*, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp 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*, pp. 1889–1897. JMLR.org, 2015.

Leonard Schulman and Umesh 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.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pp. 5192–5202, 2018.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359, 2017. doi: 10.1038/nature24270.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nat.*, 575(7782):350–354, 2019. doi: 10.1038/s41586-019-1724-z.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 2007. doi: doi:10.1515/9781400829460.

Bernhard von Stengel and Daphne Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21(1):309–321, 1997. doi: https://doi.org/10.1006/game.1997.0527.

Bernhard Von Stengel and Daphne Koller. Team-maxmin equilibria. *Games and Economic Behavior*, 21(1-2):309–321, 1997.

Okko Jan Vrieze. Stochastic games with finite state and action spaces. *CWI tracts*, 1987.

Xiaofeng Wang and Tuomas Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002*, pp. 1571–1578. MIT Press, 2002.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4259–4299. PMLR, 15–19 Aug 2021.

Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, 2011.

Brian Hu Zhang and Tuomas Sandholm. Team correlated equilibria in zero-sum extensive-form games via tree decompositions. *CoRR*, abs/2109.05284, 2021.

Brian Hu Zhang, Gabriele Farina, and Tuomas Sandholm. Team belief DAG form: A concise representation for team-correlated game-theoretic decision making. *CoRR*, abs/2202.00789, 2022.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Youzhi Zhang and Bo 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, pp. 11033–11043. PMLR, 2020a.

Youzhi Zhang and Bo An. Computing team-maxmin equilibria in zero-sum multiplayer extensive-form games. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 2318–2325. AAAI Press, 2020b.

# A Additional Preliminaries

## A.1 Background on Nonlinear Programming

In this subsection, we provide additional background on the theory of nonlinear programming (Mangasarian, 1994). For the purposes of our work, given a constrained minimization problem, we are interested in conditions that establish the existence of (nonnegative) Lagrange multipliers that satisfy the *Kaursh-Kuhn-Tucker (KKT)* conditions. In the unconstrained case, this property corresponds to the nullification of the gradient at point of a local optimum (Fermat's Theorem). In the constrained case however, further regularity conditions have to be met with respect to the underlying feasible set; this is formalized via so-called *constraint qualifications* (Bazaraa et al., 1972; Giorgi et al., 2018). For our purposes, we will use the so-called *Arrow-Hurwicz-Uzawa constraint qualification* (Arrow et al., 1961; Mangasarian, 1994) (see Theorem A.1) to show that the set of (local) optima of a particular constrained optimization problem is contained within the set of KKT points (Lemma B.3).

We first define the *nonlinear program* that encodes a *constrained minimization problem*. Then, we state the Karush-Kuhn-Tucker optimality conditions for a given feasible point of the problem.

**Constrained optimization problems.** In a *constrained optimization problem* on a Euclidean space $\mathbb{R}^d$, for $d \in \mathbb{N}$, we are interested in optimizing a given function $f : \mathbb{R}^d \to \mathbb{R}$ over a given nonempty set $\mathcal{D} \subseteq \mathbb{R}^d$. The function $f$ is called the *objective function*, and the set $\mathcal{D}$ the *constraint* or *feasibility* set. Notationally, we will represent such problems by

"Minimize $f(\boldsymbol{z})$ subject to $\boldsymbol{z} \in \mathcal{D}$", or more compactly as "$\min\{f(\boldsymbol{z}) \mid \boldsymbol{z} \in \mathcal{D}\}$."

A global solution to such a problem is a point $\boldsymbol{z}^*$ in $\mathcal{D}$ such that $f(\boldsymbol{z}^*) \leq f(\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathcal{D}$; the existence of such a solution is typically guaranteed by Weierstrass' theorem. Relaxing the requirement of global optimality, below we clarify the meaning of a local minimum with following definition:

**Definition A.1** (Local minimum)**.** *Let a function $f : \mathbb{R}^d \to \mathbb{R}$ and a point $\boldsymbol{z}_0$. A* constrained local minimum *occurs at $\boldsymbol{z}_0 \in \mathcal{D}$, with $\mathcal{D} \subseteq \mathbb{R}^d$, if there exists $\delta > 0$ such that*

$$f(\boldsymbol{z}_0) \leq f(\boldsymbol{z}),\ \forall \boldsymbol{z} \in \{B(\boldsymbol{z}_0, \delta)\} \cap \mathcal{Z},$$

*where $B(\boldsymbol{z}_0, \delta)$ denotes the set of all points belonging to the open ball with radius $\delta$ and center at $\boldsymbol{z}_0$.*

We now turn to study constrained optimization problems with feasible sets defined by inequality constraints. Namely, the constraint set will have the form

$$\mathcal{D} = \{\boldsymbol{z} \in U \mid g_i(\boldsymbol{z}) \leq 0, \forall i = 1, \ldots, m\}, \tag{6}$$

where $U \subseteq \mathbb{R}^d$ is an open set in $\mathbb{R}^d$, and $m$ is the number of the necessary inequalities to describe the feasible set $\mathcal{D}$. The minimization problem can now be written as follows.

$$\begin{array}{ll} \min & f(\boldsymbol{z}) \\ \text{s.t.} & g_i(\boldsymbol{z}) \leq 0,\ \forall i \in [m]. \end{array} \tag{MP}$$

In the sequel, we say that an inequality constraint $g_i(\boldsymbol{z}) \leq 0$ is *active* at a point $\boldsymbol{z}^*$ if the constraint holds as an equality at $\boldsymbol{z}^*$, that is, we have $g_i(\boldsymbol{z}^*) = 0$; otherwise, it is called *inactive*. Below we introduce the KKT conditions (*e.g.*, see (Boyd et al., 2004, Chapter 5.5.3)).

**Definition A.2** (Karush-Kuhn-Tucker Conditions)**.** *Suppose that $f : U \to \mathbb{R}$ and $g_i : U \to \mathbb{R}$ are differentiable functions, for any $i = 1, \ldots, m$. Further, let $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\lambda}) \coloneqq f(\boldsymbol{z}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x})$ be the associated Lagrangian function. We say that a point $(\boldsymbol{z}^*, \lambda^*)$ satisfies the KKT conditions if*

$$\begin{aligned} \lambda_i^* g_i(\boldsymbol{z}^*) &= 0, \quad \forall i = 1, \ldots, m; && \textit{(Complementary Slackness)} \\ g_i(\boldsymbol{z}^*) &\leq 0, \quad \forall i = 1, \ldots, m; && \textit{(Primal Feasibility)} \\ \lambda_i^* &\geq 0, \quad \forall i = 1, \ldots, m;\ \textit{and} && \textit{(Dual Feasibility)} \\ \nabla_{\boldsymbol{z}} \mathcal{L}(\boldsymbol{z}^*, \boldsymbol{\lambda}^*) = \nabla_{\boldsymbol{z}} f(\boldsymbol{z}^*) &+ \sum_{i=1}^m \lambda_i^* \nabla_{\boldsymbol{z}} g_i(\boldsymbol{z}^*) = \mathbf{0}. && \textit{(First-Order Stationarity)} \end{aligned} \tag{KKT}$$

In general, while these conditions are necessary for optimality, they are not necessarily sufficient. We also remark that for the unconstrained case, *i.e.*, $\{g_i(\boldsymbol{z}) \equiv 0\}$, the (KKT) conditions generalize the necessary condition of a gradient equal to zero.

**The Arrow-Hurwicz-Uzawa constraint qualification.** The establish the KKT conditions under nonconvex constraints, a number of different constraint qualifications have been developed (Bazaraa et al., 1972; Giorgi et al., 2018). We recall that constraint qualifications ensure that all the local minimizers acquire a set of (nonnegative) Lagrange multipliers that (jointly) satisfy the KKT conditions (Definition A.2). For our purposes, we will use the Arrow-Hurwicz-Uzawa constraint qualification (henceforth AHU-CQ for brevity), which is recalled below (see (Mangasarian, 1994, Ch. 7)).

**Theorem A.1** (AHU-CQ (Mangasarian, 1994))**.** *Consider a constrained minimization problem with a feasibility set $\mathcal{D}$ given in (6). Further, let $\boldsymbol{z}_0$ be a feasible point and let $A(\boldsymbol{z}_0)$ denote the set of* active constraints *at $\boldsymbol{z}$. We differentiate between concave $A'(\boldsymbol{z}_0)$ and nonconcave $A''(\boldsymbol{z}_0)$ active constraints, so that $A(\boldsymbol{z}_0) = A'(\boldsymbol{z}_0) \cup A''(\boldsymbol{z}_0)$. If there exists a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that*

$$\begin{cases} \boldsymbol{w}^\top \nabla_{\boldsymbol{z}} g_i(\boldsymbol{z}_0) \geq 0, & \forall i \in A'; \text{ and} \\ \boldsymbol{w}^\top \nabla_{\boldsymbol{z}} g_i(\boldsymbol{z}_0) > 0, & \forall i \in A'', \end{cases} \tag{7}$$

*then, the Arrow-Hurwicz-Uzawa constraint qualification at point $\boldsymbol{z}_0$ is satisfied.*

The importance of this theorem lies in the following implication, which provides sufficient conditions for the satisfaction of the KKT conditions.

**Corollary A.1.** Consider a local minimum $\boldsymbol{z}_0$ of (MP). If the Arrow-Hurwicz-Uzawa constraint qualification is satisfied at $\boldsymbol{z}_0$, there exist (nonnegative) Lagrange multipliers satisfying the (KKT) conditions of Definition A.2.

It is important to stress that the Arrow-Hurwicz-Uzawa constraint qualification—see (7)—does *not* involve the objective function; this is the case more broadly for constraint qualifications.

### A.2 Weak Convexity, the Moreau Envelope, and Near-Stationarity

In this subsection, we provide some necessary background on optimizing nonsmooth functions. We refer the interested reader to (Davis & Drusvyatskiy, 2019) for a more complete discussion on the subject.

Throughout this subsection, we will tacitly assume that $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, convex and compact subsets of a Euclidean space. We will also denote by $\mathrm{dist}$ the distance between a vector $\boldsymbol{x}$ and $\mathcal{Y}$, defined as follows.

$$\mathrm{dist}(\boldsymbol{x}; \mathcal{Y}) = \min_{\boldsymbol{y} \in \mathcal{Y}} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

**Definition A.3** (Weak Convexity)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ and any $t \in [0, 1]$, it holds that $f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$. Additionally, a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be $\lambda$-weakly convex if the function $f(\boldsymbol{x}) + \frac{\lambda}{2}\|\boldsymbol{x}\|^2$ is convex.*

The following corollary is an immediate consequence of the definition of weak convexity, and the fact that the function $\frac{\lambda}{2}\|\boldsymbol{x}\|^2$ is $\lambda$-strongly convex.

**Corollary A.2.** Let $f : \mathcal{X} \ni \boldsymbol{x} \mapsto \mathbb{R}$ be a $\lambda$-weakly convex function. Then, the function $f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|^2$ is $\lambda$-strongly convex.

A notion closely related to weak convexity within optimization literature is the *Moreau envelope* (also knwon as Moreau-Yosida regularization). Namely, the Moreau envelope of a function is defined as follows for $\lambda > 0$.

$$f_\lambda(\boldsymbol{x}) \coloneqq \min_{\boldsymbol{x}' \in \mathcal{X}} \left\{ f(\boldsymbol{x}') + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \right\}.$$

Moreover, when $\lambda < \frac{1}{\ell}$, with $\ell$ being the corresponding parameter of weak convex, the Moreau envelope $f_\lambda$ is $C^1$-smooth, and its gradient given by $\nabla f_\lambda = \lambda^{-1}(\boldsymbol{x} - \mathrm{prox}_{\lambda f}(\boldsymbol{x}))$ (Rockafellar, 1970, Theorem 31.5), where $\mathrm{prox}_{\lambda f}(\cdot)$ is the *proximal mapping*. Namely, for a convex and continuous function $f : \mathcal{X} \to \mathbb{R}$ we define its proximal operator $\mathrm{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ as follows.

$$\mathrm{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}' \in \mathcal{X}} \left\{ f(\boldsymbol{x}') + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \right\}. \tag{8}$$

The point $\tilde{\boldsymbol{x}} := \operatorname{prox}_f(\boldsymbol{x})$ that results by applying the proximal operator (8) on $\boldsymbol{x}$ is called the *proximal point* of $\boldsymbol{x}$. The proximal point of the scaled function $\lambda f$ coincides with the solution of the minimization problem needed in order to determine the Moreau envelope of $f$ at $\boldsymbol{x}$. The proximal operator of an $\ell$-weakly convex function is well-defined, as as long as $\lambda$ is sufficiently small:

**Proposition A.1** (Lin et al. (2020))**.** Let $\phi$ be a $\ell$-weakly convex function. Then, $\operatorname{prox}_{\phi/(2\ell)}(\boldsymbol{x})$ is well-defined.

**Minimization of weakly convex functions.** Generally, in a minimization problem we are interested in computing minima of a function subject to constraints. If no convexity assumption holds for the objective function, even computing local minima is NP-hard (Murty & Kabadi, 1985). Instead, one is often interested in computing an approximate stationary point of the objective function.

More precisely, an $\epsilon$-approximate stationary point $\boldsymbol{x}_0$ of a nondifferentiable function is a point such that $\operatorname{dist}(\mathbf{0}; \partial f(\boldsymbol{x}_0)) \leq \epsilon$ where $\partial f(\boldsymbol{x}_0)$ is the *subdifferential* of $f$ at $\boldsymbol{x}_0$ (see (Davis & Drusvyatskiy, 2019, Sec. 2.2)). However, such a measure of stationarity for nonsmooth objective functions is so restrictive that, in fact, it can be shown as difficult as solving the optimization problem exactly—*e.g.*, if we let $f(x) = |x|$ then $x = 0$ is the only $\epsilon$-approximate stationary point for $\epsilon \in [0, 1)$.

The alternative notion of *near stationarity* for a nonsmooth function $f(\boldsymbol{x})$, contributed by Davis & Drusvyatskiy (2019), has become standard (see Propositions 4.11 and 4.12 in (Lin et al., 2020)) for optimization of weakly convex functions. (For a more in depth discussion see (Drusvyatskiy & Paquette, 2019, Section 4.1).) More precisely, we measure stationarity by means of the proximal operator:

**Definition A.4** ($\epsilon$-nearly stationary point)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a continuous, nonsmooth function, and some $\epsilon > 0$. We say that a point $\boldsymbol{x}_0 \in \mathcal{X}$ is $\epsilon$-nearly stationary if*

$$\|\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0\|_2 \leq \epsilon,$$

*where $\tilde{\boldsymbol{x}}_0 := \operatorname{prox}_{\lambda f}(\boldsymbol{x}_0)$ is the proximal point of $\boldsymbol{x}_0$.*

The Moreau envelope of $f$ offers a number of useful properties for the analysis of convergence to near stationarity, as formalized below.

**Fact A.1** ((Davis & Drusvyatskiy, 2019))**.** Let $f : \mathcal{X} \to \mathbb{R}$ be an $\ell$-weakly convex function and $\lambda < \frac{1}{\ell}$. Further, let $\boldsymbol{x} \in \mathcal{X}$ and $\tilde{\boldsymbol{x}} := \operatorname{prox}_{\lambda f}(\boldsymbol{x})$ be its proximal point. Then,

$$\begin{cases} \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 & \leq & \lambda \|\nabla f_\lambda(\boldsymbol{x})\|; \\ f(\tilde{\boldsymbol{x}}) & \leq & f(\boldsymbol{x}); \\ \operatorname{dist}(\mathbf{0}; \partial f(\boldsymbol{x})) & \leq & \|\nabla f_\lambda(\boldsymbol{x})\|. \end{cases}$$

**Remark 2.** *An $\frac{\epsilon}{\lambda}$-approximate first-order order stationary point of $f_\lambda$ is an $\epsilon$-near stationary point of $f$.*

**Properties of the max function.** In our analysis of IPGMAX, we will measure progress based on the function $\phi(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$, where $f$ corresponds to the value function in our setting; using $\phi$ is fairly common in the context of min-max optimization. The following lemma points out some useful properties of $\phi$.

**Lemma A.1** (Lin et al. (2020))**.** Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be $L$-Lipschitz and $\ell$-smooth. Then, the function $\phi(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$ is

- $L$-Lipschitz continuous; and
- $\ell$-weakly convex.

### A.3 Further Background on Markov Decision Processes

Additionally, we will need some further preliminaries on Markov decision processes (MDPs). First, the *(discounted) state visitation measure* effectively measures the "discounted" expected amount of time that the Markov chain—induced by fixing the players' policies—spends at a state $s$ given that it starts from an initial state $\bar{s}$. That is, every visit is multiplied by a discount factor $\gamma^t$, where $t$ is the time of the visit. We note that Agarwal et al. (2021) use the definition that makes it a

probability measure, in the sense that for a given initial state distribution $\boldsymbol{\rho}$ the discounted state visitation distribution sums to 1. For convenience, we will work with the unnormalized definition found in (Puterman, 2014, Chapter 6.10) that instead sums to $\frac{1}{1-\gamma}$; this is the reason why we use the term *measure* instead of *distribution*.

**Definition A.5.** *Consider an initial state distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ and a stationary joint policy $\boldsymbol{\pi} \in \Pi$. The state visitation measure $d_{\overline{s}}^{\boldsymbol{\pi}}$ is defined as*

$$d_{\overline{s}}^{\boldsymbol{\pi}}(s) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}(s^{(t)} = s | \boldsymbol{\pi}, s^{(0)} = \overline{s}).$$

*Further, overloading notation, we let*

$$d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\overline{s} \sim \boldsymbol{\rho}} \left[ d_{\overline{s}}^{\boldsymbol{\pi}}(s) \right].$$

With a slight abuse of notation, we will also write $d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s)$ to denote the state visitation measure induced by strategies $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$.

**Definition A.6** (Distribution Mismatch Coefficient)**.** *Let $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ be a full-support distribution over states, and $\Pi$ be the joint set of policies. We define the distribution mismatch coefficient $D$ as*

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

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

To further enhance the paper's readability, we include below a table listing our main notation.

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 |
| $\mathcal{A}_k$ | Action space of player $k$ of the team |
| $\mathcal{A}$ | Team's joint action space |
| $\mathcal{B}$ | Action space of the adversary |
| $A_k$ | Number of actions available to player $k$ of the team |
| $B$ | Number of actions available to the adversary |
| $\mathcal{X}_k$ | The set of feasible directly parameterized policies of player $k$: $\mathcal{X}_k := \Delta(\mathcal{A}_k)^S$ |
| $\mathcal{X}$ | The set of feasible directly parameterized policies of the team: $\mathcal{X} := \bigtimes_{k=1}^n \mathcal{X}_k$ |
| $\mathcal{Y}$ | The set of feasible directly parameterized policies of the adversary: $\mathcal{Y} := \Delta(\mathcal{B})^S$ |
| $\gamma$ | Discount factor |
| $\mathbb{P}(s'\|s, \boldsymbol{a}, b)$ | 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 $(\boldsymbol{x}, \boldsymbol{y})$ |
| $\mathbf{V}(\boldsymbol{x}, \boldsymbol{y}), V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ | The value vector per-state, the expected value under initial distribution $\boldsymbol{\rho}$ |
| Parameters: | |
| $L$ | Lipschitz constant of the value function $V_{\boldsymbol{\rho}}(\cdot, \cdot)$ |
| $\ell$ | Smoothness constant of the value function $V_{\boldsymbol{\rho}}(\cdot, \cdot)$ |
| $D$ | Distribution mismatch coefficient |
| NLP: | |
| $\boldsymbol{\rho}$, $\rho(s)$ | Initial state distribution, probability that $s$ is the initial state |
| $\boldsymbol{v}$, $v(s)$ | value vector, value of state $s$ |
| $r(s, \boldsymbol{x}, b)$ | Expected reward at state $s$ under $(\boldsymbol{x}, b) \in \mathcal{X} \times \mathcal{B}$ |
| $\mathbb{P}(s'\|s, \boldsymbol{x}, b)$ | Expected probability of transitioning to state $s'$ from $s$ under $(\boldsymbol{x}, b) \in \mathcal{X} \times \mathcal{B}$ |
| $\boldsymbol{x}_k$ | Strategy of player $k$ of the team |
| $\boldsymbol{x}_{k,s}$ | Strategy of player $k$ of the team at state $s$ |
| Additional notation: | |
| $\phi(\boldsymbol{x})$ | Maximum of the value function w.r.t. $\boldsymbol{x}$: $\phi(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ |
| $\phi_{1/2\ell}(\boldsymbol{x})$ | Moreau envelope of $\phi$ with parameter $\lambda := \frac{1}{2\ell}$ (Definition 3.1) |
| $\hat{\boldsymbol{x}}$ | $\epsilon$-nearly stationary point of $\phi(\cdot)$ (Definition A.4) |
| $\tilde{\boldsymbol{x}}$ | proximal point of $\frac{1}{2\ell}\phi$ w.r.t. $\hat{\boldsymbol{x}}$ (Equation (8)) |

# B PROOF OF EXTENDIBILITY TO NASH EQUILIBRIA

In this section, we demonstrate how a nearly stationary point $\hat{\boldsymbol{x}}$ of $\phi(\cdot) := \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\cdot, \boldsymbol{y})$, returned by IPGMAX, can be extended to an approximate Nash equilibrium.

Our extension argument uses a nonlinear program that is in spirit similar to the one found in (Filar & Vrieze, 2012, Chapter 3.9). But, unlike the program in (Filar & Vrieze, 2012, Chapter 3.9), ours is designed to capture adversarial team Markov games. In this context, there are two main challenges in the proof. First, even if we had an *exact* stationary point of $\phi$, establishing the existence of nonnegative Lagrange multipliers that satisfy the KKT conditions is particularly challenging. This is unfortunate since it turns out that establishing the KKT conditions is crucial, and is at the heart of our extendibility argument. Indeed, the upshot is that an admissible policy for the adversary *can be derived from a subset of the Lagrange multipliers*. Further, our algorithm only has access to an *approximate* stationary point. As a result, our argument needs to be robust in terms of approximation errors.

To address the first issue, we consider a modified nonlinear program—namely, (Q-NLP) introduced earlier in Section 3.3—that incorporates an additional quadratic term to the objective function. This allows us to show that the proximal point $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ is part of a global optimum for our new program. In turn, this is crucial to establish the existence of nonnegative Lagrange multipliers at that point. Moreover, we bypass the second issue we discussed above by studying a relaxed linear program, which serves as a proxy for the ideal linear program that uses knowledge of the global optimum of (Q-NLP). Our main argument establishes that any solution to the proxy linear program is basically as good as solving the ideal one—modulo factors that depend polynomially on the natural parameters of the game. In turn, that solution—which incidentally can be computed efficiently—induces a strategy profile $\hat{\boldsymbol{y}} \in \mathcal{Y}$ so that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $O(\epsilon)$-approximate Nash equilibrium.

**Outline of the proof.** Below we sketch the main steps in our proof.

(i) In Appendix B.1 we consider (Q-NLP), a nonlinear program that incorporates an additional quadratic term to the objective function of the natural MDP formulation ($\mathrm{NLP}_{\mathcal{G}}$).

(ii) In Appendix B.2 we show that (Q-NLP) attains a global optimum at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ (Lemma B.2), where $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ and $\tilde{\boldsymbol{v}}$ is the unique value vector associated with $\tilde{\boldsymbol{x}}$ (Proposition B.1).

(iii) In Appendix B.3.1 we show that any feasible point of (Q-NLP) satisfies the Arrow-Hurwicz-Uzawa constraint qualification (Lemma B.3). In turn, this implies the existence of nonnegative Lagrange multipliers at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ satisfying the KKT conditions (Corollary B.1).

(iv) In Appendix B.4 we introduce a linear program, namely ($\mathrm{LP}_{\mathrm{adv}}$), to formulate the optimization problem faced by the adversary; ($\mathrm{LP}_{\mathrm{adv}}$) will be eventually used to compute an admissible policy for the adversary.

(v) In Lemma B.4 we show that ($\mathrm{LP}_{\mathrm{adv}}$) is always feasible. This is shown by first constructing an "ideal" linear program ($\mathrm{LP}'_{\mathrm{adv}}$), and arguing that the ideal program is feasible (Lemma B.5) using the KKT conditions. The transition to ($\mathrm{LP}_{\mathrm{adv}}$) leverages the fact that $\|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\| \leq \epsilon$ and the Lipschitz continuity of the underlying constraint functions to show that the introduced error is only $O(\epsilon)$.

(vi) Finally, this section is culminated in Lemma B.6 and Theorem B.1, which establish that any solution of ($\mathrm{LP}_{\mathrm{adv}}$) induces a policy for the advesrary $\hat{\boldsymbol{y}} \in \mathcal{Y}$ so that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $O(\epsilon)$-approximate Nash equilibrium.

## B.1 THE QUADRATIC NLP

In this subsection, we describe in more detail the nonlinear program (Q-NLP) we introduced earlier in Section 3.3. For completeness, let us first describe the perhaps most natural nonlinear formulation used to solve the min-max problem $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\rho}\big(\boldsymbol{x}, \boldsymbol{y}\big)$ (see (Filar & Vrieze, 2012, Chapter 3)), introduced below.

$$
\text{(NLP}_{\mathcal{G}}\text{)} \quad \begin{aligned}
&\min \sum_{s\in\mathcal{S}} \rho(s)v(s) \\
&\text{s.t. } r(s,\boldsymbol{x},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\boldsymbol{x},b)v(s') \leq v(s), \quad \forall (s,b) \in \mathcal{S}\times\mathcal{B}; \\
&\qquad\qquad \boldsymbol{x}_{k,s}^\top \mathbf{1} = 1, \quad \forall (k,s) \in [n]\times\mathcal{S}; \text{ and} \\
&\qquad\qquad x_{k,s,a} \geq 0, \quad \forall k \in [n], (s,a) \in \mathcal{S}\times\mathcal{A}_k.
\end{aligned}
$$

The variables of this program correspond to a strategy profile for the team players $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \in \mathcal{X}$, while the value vector $\boldsymbol{v}$ captures the value at each state when the adversary is best responding. Before we proceed further, it will be useful to note that, for any $(s,b) \in \mathcal{S}\times\mathcal{B}$ and $s' \in \mathcal{S}$, the functions $r(s,\boldsymbol{x},b)$ and $\mathbb{P}(s'|s,\boldsymbol{x},b)$ are multinear in $\boldsymbol{x}$, so that

$$
\left\{ \begin{array}{lcl}
r\big(s,(\boldsymbol{x}_k;\boldsymbol{x}_{-k}),b\big) & = & \sum_{a\in\mathcal{A}_k} x_{k,s,a} r\big(s,(\boldsymbol{e}_{k,s,a};\boldsymbol{x}_{-k}),b\big); \text{ and} \\
\mathbb{P}\left(s'|s,(\boldsymbol{x}_k;\boldsymbol{x}_{-k}),b\right) & = & \sum_{a\in\mathcal{A}_k} x_{k,s,a}\, \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\boldsymbol{x}_{-k}),b\right),
\end{array} \right.
$$

where $\boldsymbol{e}_{k,s,a} \in \Delta(\mathcal{A}_k)$ is such that its unique nonzero element corresponds to the action $a \in \mathcal{A}_k$ of agent $k \in [n]$. An additional immediate consequence that will be useful in the sequel is the following property.

$$
\left\{ \begin{array}{lcl}
\frac{\partial}{\partial x_{k,s,a}} r(s,\boldsymbol{x},b) & = & r\big(s,(\boldsymbol{e}_{k,s,a};\boldsymbol{x}_{-k},b)\big); \text{ and} \\
\frac{\partial}{\partial x_{k,s,a}} \mathbb{P}(s'|s,\boldsymbol{x},b) & = & \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\boldsymbol{x}_{-k}),b\right).
\end{array} \right.
$$

Those multilinear (nonconvex-nonconcave) functions are part of the source of the complexity in our problem. We clarify that when all team players select a fixed strategy, (NLP$_{\mathcal{G}}$) retrieves the linear-programming formulation of the Bellman equation for the single-agent MDP (Puterman, 2014)—as seen from the perspective of the adversary.

Nevertheless, for our analysis it will be convenient to use a formulation that perturbs the objective function of (NLP$_{\mathcal{G}}$) with a quadratic term. In particular, let $\phi(\cdot) = \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\cdot,\boldsymbol{y})$ and $\hat{\boldsymbol{x}} \in \mathcal{X}$ be a point such that $\|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| \leq \epsilon$, where $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/2\ell}(\hat{\boldsymbol{x}})$ is its proximal point; such a point $\hat{\boldsymbol{x}}$ will be available after the termination of the first phase of IPGMAX, as implied by Proposition 3.1. Now the program we consider still has variables $(\boldsymbol{x},\boldsymbol{v})$, but its objective function incorporates an additional quadratic term. This program was first introduced in Section 3.3, but we include it below for the convenience of the reader.

$$
\text{(Q-NLP)} \quad \begin{aligned}
&\min \sum_{s\in\mathcal{S}} \rho(s)v(s) + \ell\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \\
&\text{s.t. } r(s,\boldsymbol{x},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\boldsymbol{x},b)v(s') \leq v(s), \quad \forall (s,b) \in \mathcal{S}\times\mathcal{B}; \\
&\qquad\qquad \boldsymbol{x}_{k,s}^\top \mathbf{1} = 1, \quad \forall (k,s) \in [n]\times\mathcal{S}; \text{ and} \\
&\qquad\qquad x_{k,s,a} \geq 0, \quad \forall k \in [n], (s,a) \in \mathcal{S}\times\mathcal{A}_k.
\end{aligned}
$$

As we show in the following subsection, (Q-NLP) attains a global minimum in the proximal point $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$. First, let us point out that (Q-NLP)—and subsequently (NLP$_{\mathcal{G}}$)—has nonempty feasibility set.

**Lemma B.1.** The program Equation (Q-NLP) is feasible.

*Proof.* Let $\boldsymbol{x} \in \mathcal{X}$ be any directly parameterized policy for the team and $\boldsymbol{v} := \frac{1}{1-\gamma}\mathbf{1}$, where recall that $\mathbf{1}$ is the all-ones vector (with dimension $S$). Clearly, $\boldsymbol{x}_{k,s}^\top \mathbf{1} = 1$, for all $(k,s) \in [n]\times\mathcal{S}$, and $x_{k,s,a} \geq 0$ for all $k \in [n], (s,b) \in \mathcal{S}\times\mathcal{B}$. Further, for any $(s,b) \in \mathcal{S}\times\mathcal{B}$, we have

$$
r(s,\boldsymbol{x},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\boldsymbol{x},b)\frac{1}{1-\gamma} = r(s,\boldsymbol{x},b) + \gamma\frac{1}{1-\gamma} \leq 1 + \gamma\frac{1}{1-\gamma} \leq \frac{1}{1-\gamma}.
$$

□

### B.2 THE GLOBAL MINIMUM OF EQUATION (Q-NLP)

Here we demonstrate that Equation (Q-NLP) attains a global minimum under $\boldsymbol{x} = \tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$. To do so, we first show that fixing $\boldsymbol{x}$ yields a unique optimal value vector $\boldsymbol{v}$ such that $\boldsymbol{\rho}^\top \boldsymbol{v} = \phi(\boldsymbol{x})$, where recall that $\phi$ is defined as $\phi(\cdot) = \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\cdot, \boldsymbol{y})$. Next, we prove that the objective function of Equation (Q-NLP) is lower bounded by the minimum of the function $\Psi(\boldsymbol{w}) = \phi(\boldsymbol{w}) + \ell\|\boldsymbol{w} - \hat{\boldsymbol{x}}\|^2$; the latter function is $\ell$-strongly convex, which means that it has a unique minimizer, namely $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$. In turn, this implies that the objective function of (Q-NLP) is at least $\Psi(\boldsymbol{x})$ for any fixed $\boldsymbol{x} \in \mathcal{X}$. Finally, we conclude the proof by showing that $\tilde{\boldsymbol{x}}$ is part of a feasible solution of (Q-NLP).

First, we relate the optimal vector $\boldsymbol{v}$ that arises by fixing $\boldsymbol{x}$ in Equation (Q-NLP) and the function $\phi(\boldsymbol{x})$:

**Proposition B.1.** Suppose that $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ is full support. For any $\boldsymbol{x} \in \mathcal{X}$ there exists a unique optimal vector $\boldsymbol{v}^\star$ in Equation (Q-NLP). Further,

$$\boldsymbol{\rho}^\top \boldsymbol{v}^\star = \phi(\boldsymbol{x}).$$

*Proof.* First, we observe that by fixing a feasible point $\boldsymbol{x} \in \mathcal{X}$ in Equation (Q-NLP) we recover a linear program with variable $\boldsymbol{v} \in \mathbb{R}^S$, which incidentally corresponds to the formulation of a single-agent MDP (Puterman, 2014, Chapter 6). The reward function of this MDP is the expected reward of the adversary given that team plays $\boldsymbol{x}$, and the transition function is the expected transition function conditioned on the team playing $\boldsymbol{x} \in \mathcal{X}$. Formally, we introduce this linear program below.

$$\begin{aligned} \min \quad & \boldsymbol{\rho}^\top \boldsymbol{v} \\ \text{s.t.} \quad & r(s, \boldsymbol{x}, b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, \boldsymbol{x}, b) v(s') \leq v(s), \ \forall (s, b) \in \mathcal{S} \times \mathcal{B}. \end{aligned}$$

We claim that the optimal solution $\boldsymbol{v}^\star$ is unique for any given $\boldsymbol{x} \in \mathcal{X}$. Indeed, this is a consequence of the fact that—when $\boldsymbol{\rho}$ is full-support—it is equivalent to the Bellman optimality equation, whose solutions can be in turn expressed as the fixed point of a contraction operator (Puterman, 2014, Chapter 6.2 & 6.4). Further, let us consider its dual linear program with variables $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$:

$$\begin{aligned} \max \quad & \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} r(s, \boldsymbol{x}, b)\lambda(s, b) \\ \text{s.t.} \quad & \rho(\overline{s}) + \sum_{s\in\mathcal{S}} \sum_{b\in\mathcal{B}} \lambda(s, b)\gamma\mathbb{P}(\overline{s}|s, \boldsymbol{x}, b) - \sum_{b\in\mathcal{B}} \lambda(\overline{s}, b) = 0, \forall \overline{s} \in \mathcal{S}; \text{ and} \\ & \qquad\qquad \lambda(s, b) \geq 0, \forall (s, b) \in \mathcal{S} \times \mathcal{B}. \end{aligned}$$

The dual linear program is both feasible and bounded (Puterman, 2014, Chapter 6.9). As such, it admits at least one optimal vector $\boldsymbol{\lambda}^\star$, with the additional property that $\sum_{b\in\mathcal{B}} \lambda^\star(s, b) > 0$; the latter follows since $\boldsymbol{\rho}$ is full-support. Moreover, by (Puterman, 2014, Theorem 6.9.1), we know that

(i) Any $\boldsymbol{y} \in \mathcal{Y}$ defines a feasible vector $\boldsymbol{\lambda}$ for the dual linear program; namely,

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

(ii) Any feasible vector of the dual linear program $\boldsymbol{\lambda}$ defines a feasible $\boldsymbol{y} \in \mathcal{Y}$; namely,

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

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

An implication of this theorem is a "1–1" correspondence between $\boldsymbol{y} \in \mathcal{Y}$ and feasible solutions $\boldsymbol{\lambda}$ of the dual program. Further, for a pair $(\boldsymbol{\lambda}, \boldsymbol{y})$, the associated discounted state visitation measure is such that $d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s) = \sum_{b\in\mathcal{B}} \lambda(s,b),\ \forall s \in \mathcal{S}$. Moreover, strong duality of linear programming implies that

$$\boldsymbol{\rho}^\top \boldsymbol{v}^\star = \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda^\star(s,b) r(s,\boldsymbol{x},b) = \sum_{s\in\mathcal{S}} d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}^\star}(s) r(s,\boldsymbol{x},\boldsymbol{y}^\star).$$

But, by Claim D.4 we know that

$$V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{s\in\mathcal{S}} d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s) r(s,\boldsymbol{x},\boldsymbol{y}).$$

Thus, for an optimal pair $(\boldsymbol{\lambda}^\star, \boldsymbol{y}^\star)$, it holds that

$$V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}^\star) = \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda^\star(s,b) r(s,\boldsymbol{x},b) = \boldsymbol{\rho}^\top \boldsymbol{v}^\star.$$

Finally, the optimality of $\boldsymbol{\lambda}^\star$ in the dual program implies that for any correspondence pair $(\boldsymbol{\lambda}, \boldsymbol{y})$,

$$\begin{aligned}\boldsymbol{\rho}^\top \boldsymbol{v}^\star &= \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda^\star(s,b) r(s,\boldsymbol{x},b) \\ &\geq \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda(s,b) r(s,\boldsymbol{x},b) \\ &= V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}).\end{aligned}$$

□

**Lemma B.2.** Let $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$, and $\tilde{\boldsymbol{v}}$ be the unique minimizer for (Q-NLP) under a fixed $\boldsymbol{x} = \tilde{\boldsymbol{x}}$. Then, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is a global minimum of (Q-NLP).

*Proof.* Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. By Proposition B.1, we know that there is a unique optimal vector $\boldsymbol{v}^\star$ in (Q-NLP), which also satisfies the equality

$$\boldsymbol{\rho}^\top \boldsymbol{v}^\star = \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}). \tag{13}$$

Now let us consider the function $\Psi(\boldsymbol{w}) := \phi(\boldsymbol{w}) + \ell\|\boldsymbol{w} - \hat{\boldsymbol{x}}\|^2$. $\Psi$ is $\ell$-strongly convex and its unique minimum value is attained at $\tilde{\boldsymbol{x}} := \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ (Corollary A.2). By (13), it follows that for any feasible $(\boldsymbol{x}, \boldsymbol{v})$,

$$\boldsymbol{\rho}^\top \boldsymbol{v} + \ell\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 \geq \min_{\boldsymbol{x}\in\mathcal{X}} \Psi(\boldsymbol{x}).$$

Finally, the value $\min_{\boldsymbol{x}\in\mathcal{X}} \Psi(\boldsymbol{x})$ is indeed attained by (Q-NLP) when we set $\boldsymbol{x} = \tilde{\boldsymbol{x}}$, which is feasible for (Q-NLP) (see Lemma B.1 and Proposition B.1). □

### B.3 KKT Conditions for a Minimizer of Equation (Q-NLP)

As we have shown in the previous subsection, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is a minimum of the program Equation (Q-NLP). In this subsection, we leverage this fact to establish the existence of nonnegative Lagrange multipliers at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ that satisfy the KKT conditions; this will be crucial for our extendibility argument. First, let us write the Lagrangian $\mathcal{L}\Big((\boldsymbol{x},\boldsymbol{v}), (\boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\zeta})\Big)$ of the constrained minimization problem associated with (Q-NLP):

$$\begin{aligned}\mathcal{L} := \boldsymbol{\rho}^\top \boldsymbol{v} + \ell\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda(s,b) \left( r(s,\boldsymbol{x},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\boldsymbol{x},b) v(s') - v(s) \right) \\ + \sum_{(k,s)} \omega(k,s) \left( \boldsymbol{x}_{k,s}^\top \mathbf{1} - 1 \right) + \sum_{(k,s)} \psi(k,s) \left( 1 - \boldsymbol{x}_{k,s}^\top \mathbf{1} \right) + \sum_{(k,s,a)} \zeta(k,s,a) \left( -x_{k,s,a} \right),\end{aligned} \tag{14}$$

where

$$\{\lambda(s,b)\}_{(s,b)} \cup \{\omega(k,s)\}_{(k,s)} \cup \{\psi(k,s)\}_{(k,s)} \cup \{\zeta(k,s,a)\}_{(k,s,a)}$$

are the associated Lagrange multipliers. Let us denote by $I$ set indexing the constraints of (Q-NLP). Before we proceed, we partition the set of constraints $I$ into $I = I_1 \cup I_2 \cup I_2' \cup I_3$, such that:

- The constraints of Equation ($Q1$), corresponding to the subset of Lagrange multipliers $\{\lambda(s,b)\}_{(s,b)}$, are assumed to lie in set $I_1$, so that every index $i \in I_1$ is uniquely associated with a pair $(s,b) \in \mathcal{S} \times \mathcal{B}$. In particular, for all $i \in I_1$, and the uniquely associated pair $(s,b) \in \mathcal{S} \times \mathcal{B}$, we let

$$g_i(\boldsymbol{x}, \boldsymbol{v}) \coloneqq r(s, \boldsymbol{x}, b) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \boldsymbol{x}, b) v(s') - v(s).$$

  For any index $i \in I_1$, and the associated pair $(s,b) \in \mathcal{S} \times \mathcal{B}$, we have that
  - For any $\overline{s} \in \mathcal{S}$,

$$\frac{\partial}{\partial v(\overline{s})} g_i(\boldsymbol{x}, \boldsymbol{v}) = \begin{cases} \gamma\, \mathbb{P}(\overline{s}|s, \boldsymbol{x}, b), & \text{if } \overline{s} \neq s; \text{ and} \\ -1 + \gamma\, \mathbb{P}(s|s, \boldsymbol{x}, b), & \text{if } \overline{s} = s. \end{cases}$$

  - For any $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}_k$, the partial derivative $\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} g_i(\boldsymbol{x}, \boldsymbol{v})$ is equal to

$$\begin{cases} 0, & \text{if } \overline{s} \neq s; \text{ and} \\ r\big(s, (\boldsymbol{e}_{\overline{k},s,\overline{a}}; \boldsymbol{x}_{-\overline{k},s}), b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s'|s, (\boldsymbol{e}_{\overline{k},s,\overline{a}}; \boldsymbol{x}_{-\overline{k},s}), b\right) v(s') & \text{if } \overline{s} = s. \end{cases}$$

- The constraints described by ($Q2$), corresponding to the subset of Lagrange multipliers $\{\omega(k,s)\}_{(k,s)} \cup \{\psi(k,s)\}_{(k,s)}$, are assumed to lie in the set $I_2 \cup I_2'$ as follows. Every equality constraint ($Q2$) is converted to a pair of inequality constraints corresponding to the sets $I_2$ and $I_2'$, respectively, so that every index $i \in I_2$ or $i \in I_2'$ is uniquely associated with a pair $(k,s) \in [n] \times \mathcal{S}$. In particular, for all $i \in I_2$, and the associated pair $(k,s) \in [n] \times \mathcal{S}$, we let

$$g_i(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \boldsymbol{x}_{k,s}^\top \mathbf{1} - 1,$$

  and for all $i \in I_2'$

$$g_i'(\boldsymbol{x}, \boldsymbol{v}) \coloneqq 1 - \boldsymbol{x}_{k,s}^\top \mathbf{1}.$$

  For any index $i \in I_2$ and the associated pair $(k,s) \in [n] \times \mathcal{S}$, we have that
  - For any $\overline{s} \in \mathcal{S}$,

$$\frac{\partial}{\partial v(\overline{s})} g_i(\boldsymbol{x}, \boldsymbol{v}) = 0.$$

  - For any $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} g_i(\boldsymbol{x}, \boldsymbol{v}) = \begin{cases} 1, & \text{if } (k,s) = (\overline{k}, \overline{s}); \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

  For any index $i \in I_2'$ and the associated pair $(k,s) \in [n] \times \mathcal{S}$, we have that
  - For any $\overline{s} \in \mathcal{S}$,

$$\frac{\partial}{\partial v(\overline{s})} g_i'(\boldsymbol{x}, \boldsymbol{v}) = 0.$$

  - For any $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} g_i'(\boldsymbol{x}, \boldsymbol{v}) = \begin{cases} -1, & \text{if } (k,s) = (\overline{k}, \overline{s}); \\ 0, & \text{otherwise.} \end{cases}$$

- Finally, the constraints described by ($Q3$), corresponding to the subset of Lagrangian multipliers $\{\zeta(k,s,a)\}_{(k,s,a)}$, are assumed to lie in the set $I_3$, so that every index $i \in I_3$ is uniquely associated with a triple $(k,s,a)$. In particular, for each $i \in I_3$, and the associated triple $(k,s,a)$, we let

$$g_i(\boldsymbol{x}, \boldsymbol{v}) \coloneqq -x_{k,s,a}.$$

  For any index $i \in I_3$ and the associated triple $(k,s,a)$, we have that

– For any $\overline{s} \in \mathcal{S}$,

$$\frac{\partial}{\partial v(\overline{s})} g_i(\boldsymbol{x}, \boldsymbol{v}) = 0.$$

– For any $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} g_i(\boldsymbol{x}, \boldsymbol{v}) = \begin{cases} -1, & \text{if } (k, s, a) = (\overline{k}, \overline{s}, \overline{a}); \\ 0, & \text{otherwise.} \end{cases}$$

We are now ready to determine the partial derivatives of the Lagrangian, as formalized below.

**Claim B.1.** Consider the Lagrangian function $\mathcal{L}$ of (Q-NLP), as introduced in (14). Then, for any $\overline{s} \in \mathcal{S}$, the partial derivative of $\mathcal{L}$ with respect to $v(\overline{s})$ reads

$$\frac{\partial}{\partial v(\overline{s})} \mathcal{L} = \rho(\overline{s}) + \sum_{s \in \mathcal{S}} \sum_{b \in \mathcal{B}} \Big[ \lambda(s, b) \gamma \, \mathbb{P}\left(\overline{s} | s, \boldsymbol{x}, b\right) \Big] - \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b). \tag{15}$$

Further, for any $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}_k$, the partial derivative $\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} \mathcal{L}$ is equal to

$$\begin{aligned} 2\ell(x_{\overline{k},\overline{s},\overline{a}} - \hat{x}_{\overline{k},\overline{s},\overline{a}}) + \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b) \Bigg[ r\left(\overline{s}, (\boldsymbol{e}_{\overline{k},\overline{s},\overline{a}}; \boldsymbol{x}_{-\overline{k}}), b\right) + \gamma \sum_{s \in \mathcal{S}} \mathbb{P}\left(s | \overline{s}, (\boldsymbol{e}_{\overline{k},\overline{s},\overline{a}}; \boldsymbol{x}_{-\overline{k}}), b\right) v(s) \Bigg] \\ + \omega(\overline{k}, \overline{s}) - \psi(\overline{k}, \overline{s}) - \zeta(\overline{k}, \overline{s}, \overline{a}). \end{aligned} \tag{16}$$

*Proof.* Let us first establish (15). Fix any $\overline{s} \in \mathcal{S}$. The partial derivative of the objective function of (Q-NLP) with respect to $v(\overline{s})$ reads

$$\frac{\partial}{\partial v(\overline{s})} \left( \boldsymbol{\rho}^\top \boldsymbol{v} + \ell \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 \right) = \rho(\overline{s}).$$

Further, ($Q1$) is the only constraint that involves the variable $v(\overline{s})$, and we previously showed that for any $i \in I_1$,

$$\frac{\partial}{\partial v(\overline{s})} g_i(\boldsymbol{x}, \boldsymbol{v}) = \begin{cases} \gamma \, \mathbb{P}(\overline{s} | s, \boldsymbol{x}, b), & \text{if } \overline{s} \neq s; \\ -1 + \gamma \, \mathbb{P}(s | s, \boldsymbol{x}, b), & \text{if } \overline{s} = s, \end{cases}$$

where $(s, b) \in \mathcal{S} \times \mathcal{B}$ is the pair associated with index $i \in I_1$. Thus,

$$\begin{aligned} \sum_{(s,b) \in \mathcal{S} \times \mathcal{B}} \frac{\partial}{\partial v(\overline{s})} g_i(\boldsymbol{x}, \boldsymbol{v}) &= \sum_{b \in \mathcal{B}} \sum_{s \neq \overline{s}} \left[ \lambda(s, b) \gamma \, \mathbb{P}(\overline{s} | s, \boldsymbol{x}, b) \right] + \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b) \left( -1 + \mathbb{P}(\overline{s} | \overline{s}, \boldsymbol{x}, b) \right) \\ &= \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}} \left[ \lambda(s, b) \gamma \, \mathbb{P}(\overline{s} | s, \boldsymbol{x}, b) \right] - \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b). \end{aligned}$$

As a result, we conclude that

$$\frac{\partial}{\partial v(\overline{s})} \mathcal{L} = \rho(\overline{s}) + \sum_{s \in \mathcal{S}} \sum_{b \in \mathcal{B}} \lambda(s, b) \gamma \, \mathbb{P}(\overline{s} | s, \boldsymbol{x}, b) - \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b),$$

establishing (15). Next, we show (16). We first calculate the partial derivative of the objective function:

$$\frac{\partial}{\partial x_{\overline{k},\overline{s},\overline{a}}} \left( \boldsymbol{\rho}^\top \boldsymbol{v} + \ell \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 \right) = 2\ell(x_{\overline{k},\overline{s},\overline{a}} - \hat{x}_{\overline{k},\overline{s},\overline{a}}). \tag{17}$$

Moreover, the summation of all the partial derivatives with respect to $x_{\overline{k},\overline{s},\overline{a}}$, for a fixed triple $(\overline{k}, \overline{s}, \overline{a})$, of the constraints ($Q1$), ($Q2$), and ($Q3$), multiplied by their respective Lagrange multipliers reads

$$\begin{aligned} \sum_{b \in \mathcal{B}} \lambda(\overline{s}, b) \Big( r\big(\overline{s}, (\boldsymbol{e}_{\overline{k},\overline{s},\overline{a}}; \boldsymbol{x}_{-\overline{k},s}), b\big) + \gamma \sum_{s \in \mathcal{S}} \mathbb{P}\left(s | \overline{s}, (\boldsymbol{e}_{\overline{k},\overline{s},\overline{a}}; \boldsymbol{x}_{-\overline{k},s}), b\right) v(s) \Big) \\ + \omega(\overline{k}, \overline{s}) - \psi(\overline{k}, \overline{s}) - \zeta(\overline{k}, \overline{s}, \overline{a}). \end{aligned} \tag{18}$$

Combining (17) and (18) implies (16), concluding the proof. □

### B.3.1 LOCAL OPTIMA SATISFY THE KKT CONDITIONS

Here we will show that for $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X} \times \mathbb{R}^S$, a global minimum of (Q-NLP), there exist (nonnegative) Lagrange multipliers that jointly satisfy the KKT conditions. We will first argue in Lemma B.3 below that any feasible point of (Q-NLP) satisfies the Arrow-Hurwicz-Uzawa constraint qualification. Then, we will leverage Corollary A.1 to show that any local minimizer of (Q-NLP)—and in particular $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$—attains Lagrange multipliers that satisfy the KKT conditions. The following proof is analogous to (Vrieze, 1987, Ch. 4.4).

**Lemma B.3.** Let $(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{X} \times \mathbb{R}^S$ be any feasible point of (Q-NLP). Then, the Arrow-Hurwicz-Uzawa constraint qualification is satisfied at $(\boldsymbol{x}, \boldsymbol{v})$.

*Proof.* Suppose that $A(\boldsymbol{x}, \boldsymbol{v}) \subseteq I$ is the set of active constraints at a feasible point $(\boldsymbol{x}, \boldsymbol{v})$. Let us further denote by $d$ the dimension of $(\boldsymbol{x}, \boldsymbol{v})$. To apply Theorem A.1, we have to establish the existence of a vector $\boldsymbol{w} \in \mathbb{R}^d$, such that for any $i \in A(\boldsymbol{x}, \boldsymbol{v})$,

$$\begin{cases} \boldsymbol{w}^\top \nabla_{(\boldsymbol{x},\boldsymbol{v})} g_i(\boldsymbol{x}, \boldsymbol{v}) > 0, & \text{if } g_i \text{ is nonconcave; and} \\ \boldsymbol{w}^\top \nabla_{(\boldsymbol{x},\boldsymbol{v})} g_i(\boldsymbol{x}, \boldsymbol{v}) \geq 0, & \text{if } g_i \text{ concave.} \end{cases}$$

For convenience, we will index the entries of $\boldsymbol{w}$ so that $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{v}})$. For reasons that will shortly become clear, we set $\boldsymbol{w}_{\boldsymbol{x}} = \boldsymbol{0}$. Now consider any active constraint $i$ (if any exists) from the set $I_2 \cup I_2' \cup I_3$. The corresponding constraint function $g_i$ is affine, and in particular concave. Further, it holds that for any $s \in \mathcal{S}$,

$$\frac{\partial}{\partial v(s)} g_i(\boldsymbol{x}, \boldsymbol{v}) = 0.$$

As a result, for our choice of vector $\boldsymbol{w} = (\boldsymbol{0}, \boldsymbol{w}_{\boldsymbol{v}})$, it immediately follows that

$$\boldsymbol{w}^\top \nabla_{(\boldsymbol{x},\boldsymbol{v})} g_i(\boldsymbol{x}, \boldsymbol{v}) = 0,$$

for any $i \in I_2 \cup I_2' \cup I_3$. Let us now treat (if any) active constraints $i \in I_1$. In particular, let $(s, b) \in \mathcal{S} \times \mathcal{B}$ be the pair associated with $i$, so that

$$g_i(\boldsymbol{x}, \boldsymbol{v}) = r(s, \boldsymbol{x}, b) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \boldsymbol{x}, b) v(s') - v(s).$$

Then,

$$\begin{aligned} \boldsymbol{w}^\top \nabla g_i(\boldsymbol{x}, \boldsymbol{v}) &= \boldsymbol{w}_{\boldsymbol{v}}^\top \nabla_{\boldsymbol{v}} \Big[ r(s, \boldsymbol{x}, b) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \boldsymbol{x}, b) \boldsymbol{v}(s') - \boldsymbol{v}(s) \Big] \Big|_{(\boldsymbol{x},\boldsymbol{v})} \\ &= \sum_{\overline{s} \neq s} w_{v(\overline{s})} \gamma \, \mathbb{P}(\overline{s}|s, \boldsymbol{x}, b) + w_{v(s)} \Big( -1 + \gamma \, \mathbb{P}(s|s, \boldsymbol{x}, b) \Big) \\ &= \sum_{\overline{s} \in \mathcal{S}} w_{v(\overline{s})} \gamma \, \mathbb{P}(\overline{s}|s, \boldsymbol{x}, b) - w_{v(s)}. \end{aligned}$$

By virtue of Theorem A.1, it suffices to show that there exists $\boldsymbol{w}_{\boldsymbol{v}}$ so that for any $(s, b) \in \mathcal{S} \times \mathcal{B}$,

$$\gamma \sum_{s' \in \mathcal{S}} w_{v(s')} \, \mathbb{P}(s'|s, \tilde{\boldsymbol{x}}, b) - w_{v(s)} > 0.$$

We will show that this property holds for $\boldsymbol{w}_{\boldsymbol{v}} := -\boldsymbol{v}$. Indeed, since $(\boldsymbol{x}, \boldsymbol{v})$ is feasible, we get that

$$\gamma \sum_{s' \in \mathcal{S}} w_{v(s')} \, \mathbb{P}(s'|s, \tilde{\boldsymbol{x}}, b) - w_{v(s)} = -\gamma \sum_{s' \in \mathcal{S}} v(s') \, \mathbb{P}(s'|s, \tilde{\boldsymbol{x}}, b) + v(s) \geq r(s, \boldsymbol{x}, b) > 0,$$

since we have assumed that $r(s, \boldsymbol{a}, b) > 0$ for any $(\boldsymbol{a}, b) \in \mathcal{A} \times \mathcal{B}$. This concludes the proof. □

Next, leveraging this lemma and Corollary A.1, we conclude that $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$—in fact, any local minimum of (Q-NLP)—attains nonnegative Lagrange multipliers that satisfy the KKT conditions.

**Corollary B.1.** For any local minimum $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X} \times \mathbb{R}^S$ of (Q-NLP), there exists (nonnegative) Lagrange multipliers satisfying the KKT conditions.

In particular, by the first-order stationarity condition and the complementary slackness condition (recall Definition A.2) with respect to $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$, we have

$$\nabla_{(\boldsymbol{x},\boldsymbol{v})}\mathcal{L}\Big((\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}), (\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\zeta}})\Big) = \mathbf{0}; \tag{19a}$$

$$\begin{aligned}
&\tilde{\lambda}(s,b)\Big(r(s,\tilde{\boldsymbol{x}},b) + \gamma \textstyle\sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\tilde{\boldsymbol{x}},b)\tilde{v}(s') - \tilde{v}(s)\Big) = 0, \quad \forall (s,b) \in \mathcal{S}\times\mathcal{B};\\
&\tilde{\omega}(k,s)\Big(\tilde{\boldsymbol{x}}_{k,s}^\top \mathbf{1} - 1\Big) = 0, \quad \forall (k,s) \in [n]\times\mathcal{S};\\
&\tilde{\psi}(k,s)\Big(1 - \tilde{\boldsymbol{x}}_{k,s}^\top \mathbf{1}\Big) = 0, \quad \forall (k,s) \in [n]\times\mathcal{S};\\
&\tilde{\zeta}(k,s,a)\Big(-\tilde{x}_{k,s,a}\Big) = 0, \quad \forall k \in [n], \forall (s,a) \in \mathcal{S}\times\mathcal{A}_k; \text{ and}
\end{aligned} \tag{19b}$$

$$\tilde{\omega}(k,s), \tilde{\psi}(k,s), \tilde{\zeta}(k,s,a) \geq 0, \ \forall (k,s) \in [n]\times\mathcal{S}, \text{ and } \forall k \in [n], (s,a) \in \mathcal{S}\times\mathcal{A}_k. \tag{19c}$$

### B.3.2 Connecting the Lagrange Multipliers with the Visitation Measure

Here we establish an important connection between a subset of the Lagrange multipliers and the *visitation measure* under a specific policy of the adversary. This fact will be crucial later in the proof of Lemma B.6 for controlling the approximation error.

**Proposition B.2.** Suppose that the initial distribution $\boldsymbol{\rho}$ is full support. Let also $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}_{\geq 0}^{S\times B}$ be the associated vector of Lagrange multipliers at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X}\times\mathbb{R}^S$ that satisfy (19). Then, it holds that $\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b) > 0$, for any $s \in \mathcal{S}$. Further, if

$$\tilde{y}_{s,b} := \frac{\tilde{\lambda}(s,b)}{\sum_{b'\in\mathcal{B}} \tilde{\lambda}(s,b')},$$

for any $(s,b) \in \mathcal{S}\times\mathcal{B}$, then it holds that

$$\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b) = d_{\boldsymbol{\rho}}^{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}}(s), \quad \forall s \in \mathcal{S},$$

where $d_{\boldsymbol{\rho}}^{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}}(s)$ defines the visitation measure at state $s \in \mathcal{S}$ induced by $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$.

*Proof.* First of all, it follows directly from (15) and the fact that the Langrange multipliers are nonnegative that $\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b) > 0$. Next, for convenience, let us define a vector $\boldsymbol{d} \in \mathbb{R}_{>0}^S$ such that

$$d(s) = \sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b), \tag{20}$$

for all $s \in \mathcal{S}$. Then, starting from (15), we have that for any $\overline{s} \in \mathcal{S}$,

$$\rho(\overline{s}) + \sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}} \left[\frac{d(s)}{d(s)}\tilde{\lambda}(s,b)\gamma\,\mathbb{P}\left(\overline{s}|s,\tilde{\boldsymbol{x}},b\right)\right] - d(\overline{s}) = 0 \tag{21}$$

$$\rho(\overline{s}) + \sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}} \left[\frac{d(s)}{\sum_{b'\in\mathcal{B}} \tilde{\lambda}(s,b')}\tilde{\lambda}(s,b)\gamma\,\mathbb{P}\left(\overline{s}|s,\tilde{\boldsymbol{x}},b\right)\right] - d(\overline{s}) = 0$$

$$\rho(\overline{s}) + \sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}} \left[d(s)\frac{\tilde{\lambda}(s,b)}{\sum_{b'\in\mathcal{B}} \tilde{\lambda}(s,b')}\gamma\,\mathbb{P}\left(\overline{s}|s,\tilde{\boldsymbol{x}},b\right)\right] - d(\overline{s}) = 0$$

$$\rho(\overline{s}) + \gamma\sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}} \left[d(s)\tilde{y}_{s,b}\,\mathbb{P}\left(\overline{s}|s,\tilde{\boldsymbol{x}},b\right)\right] - d(\overline{s}) = 0 \tag{22}$$

$$\rho(\overline{s}) + \gamma\sum_{s\in\mathcal{S}} \left[d(s)\,\mathbb{P}\left(\overline{s}|s,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}\right)\right] - d(\overline{s}) = 0, \tag{23}$$

where (21) uses the definition of $\boldsymbol{d}$ given in (20); (22) follows from the definition of strategy $\boldsymbol{y}$ in the statement of the proposition; and (23) is derived since $\mathbb{P}(\overline{s}|s,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) = \sum_{b\in\mathcal{B}} \tilde{y}_{s,b}\mathbb{P}(\overline{s}|s,\tilde{\boldsymbol{x}},b)$

(law of total probability). Next, we observe that (23) can be compactly expressed as $\boldsymbol{\rho}^\top = \boldsymbol{d}^\top (\mathbf{I} - \gamma\, \mathbb{P}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}))$ (recall the definition of matrix $\mathbb{P}$), in turn implying that

$$\boldsymbol{d}^\top = \boldsymbol{\rho}^\top (\mathbf{I} - \gamma\, \mathbb{P}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}))^{-1}.$$

We note that $(\mathbf{I} - \gamma\, \mathbb{P}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}))$ is invertible (Claim D.1). As a result, by virtue of Claim D.3 we conclude that $\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) = d_{\boldsymbol{\rho}}^{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}}(s)$, for all $s \in \mathcal{S}$. This concludes the proof. □

We also provide an additional auxiliary claim that will be useful in the sequel. The proof follows by carefully leveraging the KKT conditions, as we formalize below.

**Claim B.2.** Let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X} \times \mathbb{R}^S$ be a local optimum of the (Q-NLP), and $\{\tilde{\lambda}(s, b)\}, \{\tilde{\psi}(k, s)\}, \{\tilde{\omega}(k, s)\}$ be the associated Lagrange multipliers defined in Equation (19). Then, for any player $k \in [n]$,

$$\tilde{v}(s) - \frac{2\ell(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s})^\top \tilde{\boldsymbol{x}}_{k,s}}{\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b)} = \frac{\tilde{\psi}(k, s) - \tilde{\omega}(k, s)}{\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b)}, \quad \forall s \in \mathcal{S}.$$

*Proof.* First, multiplying Equation (16) by $\tilde{x}_{k,s,a}$ we get that

$$\begin{aligned}
&-2\ell(\tilde{x}_{k,s,a} - \hat{x}_{k,s,a})\tilde{x}_{k,s,a} \\
&\quad + \tilde{x}_{k,s,a} \sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \Big[ r\big(s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\right) v(s') \Big] \\
&\quad + \tilde{x}_{k,s,a} \Big( \tilde{\omega}(k, s) - \tilde{\psi}(k, s) \Big) - \tilde{x}_{k,s,a} \tilde{\zeta}(k, s, a) = 0, \quad \forall k \in [n], (s, a) \in \mathcal{S} \times \mathcal{A}.
\end{aligned}$$

By complementary slackness, it follows that $-\tilde{x}_{k,s,a}\tilde{\zeta}(k, s, a) = 0$, for all $k \in [n], (s, a) \in \mathcal{S} \times \mathcal{A}_k$. Thus, the previously displayed equation can be simplified as

$$\begin{aligned}
&-2\ell(\tilde{x}_{k,s,a} - \hat{x}_{k,s,a})\tilde{x}_{k,s,a} \\
&\quad + \tilde{x}_{k,s,a} \sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \Big[ r\big(s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\right) v(s') \Big] \\
&\quad + \tilde{x}_{k,s,a} \Big( \tilde{\omega}(k, s) - \tilde{\psi}(k, s) \Big) = 0, \quad \forall k \in [n], (s, a) \in \mathcal{S} \times \mathcal{A}_k.
\end{aligned}$$

Next, summing the previous equation over all $a \in \mathcal{A}_k$ it follows that for any $(k, s) \in [n] \times \mathcal{S}$,

$$\begin{aligned}
&\sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a} \sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \Big[ r\big(s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\right) \tilde{v}(s') \Big] \\
&\qquad -2\ell \sum_{a \in \mathcal{A}_k} (\tilde{x}_{k,s,a} - \hat{x}_{k,s,a})\tilde{x}_{k,s,a} + \sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a} \Big( \tilde{\omega}(k, s) - \tilde{\psi}(k, s) \Big) = 0 \\
&\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a} \Big[ r\big(s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\right) \tilde{v}(s') \Big] \\
&\qquad -2\ell(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s})^\top \tilde{\boldsymbol{x}}_{k,s} + \Big( \tilde{\omega}(k, s) - \tilde{\psi}(k, s) \Big) = 0,
\end{aligned}$$

where the last derivation uses that $\sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a} = 1$ since $\boldsymbol{x}_{k,s} \in \Delta(\mathcal{A}_k)$. Further, using that

(i) $\sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a} r\big(s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\big) = r\big(s, \tilde{\boldsymbol{x}}, b\big)$, and

(ii) $\sum_{a \in \mathcal{A}_k} \tilde{x}_{k,s,a}\, \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \tilde{\boldsymbol{x}}_{-k}), b\right) = \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right)$,

it follows that for any $(k, s) \in [n] \times \mathcal{S}$,

$$\begin{aligned}
&\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \Big[ r\big(s, \tilde{\boldsymbol{x}}, b\big) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) \tilde{v}(s') \Big] \\
&\qquad -2\ell(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s})^\top \tilde{\boldsymbol{x}}_{k,s} + \Big( \tilde{\omega}(k, s) - \tilde{\psi}(k, s) \Big) = 0. \qquad (24)
\end{aligned}$$

Further, we know from the complementary slackness condition (19) that for any $(s, b) \in \mathcal{S} \times \mathcal{B}$,

$$\tilde{\lambda}(s,b)\left(r\big(s,\tilde{\boldsymbol{x}}_s,b\big)+\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s'|s,\tilde{\boldsymbol{x}}_s,b\right)\tilde{v}(s')-\tilde{v}(s)\right)=0.$$

In turn, summing over all actions $b \in \mathcal{B}$ we get that for any $s \in \mathcal{S}$,

$$\tilde{v}(s)\sum_{b\in\mathcal{B}}\tilde{\lambda}(s,b)=\sum_{b\in\mathcal{B}}\tilde{\lambda}(s,b)\Big[r\big(s,\tilde{\boldsymbol{x}},b\big)+\gamma\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s'|s,\tilde{\boldsymbol{x}},b\right)\tilde{v}(s')\Big].$$

Combining this equation with (24), and recalling that $\sum_{b\in\mathcal{B}}\tilde{\lambda}(s,b)>0$ for any $s \in \mathcal{S}$ (by Proposition B.2), leads to the desired conclusion. □

### B.4 Efficient Extension to Nash Equilibria

This subsection completes the proof that an $\epsilon$-near stationary point $\hat{\boldsymbol{x}}$ of $\phi$ can be extended to a strategy profile $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ that is an $O(\epsilon)$-approximate Nash equilibrium. Further, we provide a computationally efficient way for computing $\hat{\boldsymbol{y}}$ based on an appropriate linear program, ($\mathrm{LP}_{\mathrm{adv}}$) introduced below. The upshot is that feasible solutions of ($\mathrm{LP}_{\mathrm{adv}}$) induce the appropriate strategy for the adversary $\hat{\boldsymbol{y}} \in \mathcal{Y}$. In this context, we are ready to introduce ($\mathrm{LP}_{\mathrm{adv}}$), a linear program with free variables $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$:

$$(\mathrm{LP}_{\mathrm{adv}})\quad\begin{aligned}
\max\quad & \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}}\lambda(s,b)r\left(s,\hat{\boldsymbol{x}},b\right)\\
\text{s.t.}\quad & \sum_b\lambda(s,b)\left[r\left(s,(\boldsymbol{e}_{k,s,a};\hat{\boldsymbol{x}}_{-k}),b\right)+\gamma\sum_{s'}\mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\hat{\boldsymbol{x}}_{-k}),b\right)\hat{v}(s')-\hat{v}(s)\right]\\
&\qquad\qquad\geq -c_1\cdot\epsilon,\quad\forall s\in\mathcal{S}; && (\mathrm{LP}_{\mathrm{adv}}.1)\\
&\lambda(s,b)\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right)+\gamma\textstyle\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right]-\hat{v}(s)\right)\leq c_2\cdot\epsilon,\\
&\qquad\qquad\forall(s,b)\in\mathcal{S}\times\mathcal{B}; && (\mathrm{LP}_{\mathrm{adv}}.2)\\
&\lambda(s,b)\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right)+\gamma\textstyle\sum_{s'\in\mathcal{S}}\mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right]-\hat{v}(s)\right)\geq -c_2\cdot\epsilon\\
&\qquad\qquad\forall(s,b)\in\mathcal{S}\times\mathcal{B}; && (\mathrm{LP}_{\mathrm{adv}}.3)\\
&\textstyle\sum_{b\in\mathcal{B}}\lambda(s,b)\geq\rho(s),\forall s\in\mathcal{S}; && (\mathrm{LP}_{\mathrm{adv}}.4)\\
&\sum_{b\in\mathcal{B}}\lambda(s,b)\leq\frac{1}{1-\gamma},\ \forall s\in\mathcal{S}. && (\mathrm{LP}_{\mathrm{adv}}.5)
\end{aligned}$$

Here,

$$\begin{aligned}
c_2 &\coloneqq \frac{1}{1-\gamma}\left(\sqrt{\textstyle\sum_{k=1}^n A_k}+\gamma S\sqrt{\textstyle\sum_{k=1}^n A_k}\frac{1}{1-\gamma}+\gamma SL+L\right),\\
c_1 &\coloneqq 4\ell+c_2.
\end{aligned}$$

Before we proceed, a few remarks are in order. First, let us relate ($\mathrm{LP}_{\mathrm{adv}}$) with (Q-NLP). As alluded to by our notation, the free variables of ($\mathrm{LP}_{\mathrm{adv}}$) are related to a subset of the Lagrange multipliers introduced in (14). In light of this, ($\mathrm{LP}_{\mathrm{adv}}.2$) and ($\mathrm{LP}_{\mathrm{adv}}.3$) are related to the complementary slackness condition given in (19b), while ($\mathrm{LP}_{\mathrm{adv}}.1$) is related to the first-order stationary condition (19a). An important point is that we previously established the KKT conditions only with respect to the pair $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$, instead of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$. This partially explains the "slackness" we introduced in ($\mathrm{LP}_{\mathrm{adv}}.1$), ($\mathrm{LP}_{\mathrm{adv}}.2$) and ($\mathrm{LP}_{\mathrm{adv}}.3$). Correspondingly, the slackness parameters $c_1$ and $c_2$ were introduced to "transfer" the constraints from $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ to $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$, in a sense that will become clear in the sequel. We stress that expressing ($\mathrm{LP}_{\mathrm{adv}}$) in terms of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$ is crucial since $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is not actually available to the algorithm. We also remark that the objective function of ($\mathrm{LP}_{\mathrm{adv}}$) is not relevant for out argument; even a constant objective would suffice for our purposes.

But first, we need to show that ($\mathrm{LP}_{\mathrm{adv}}$) is feasible. To do so, we construct an auxiliary linear program that, unlike ($\mathrm{LP}_{\mathrm{adv}}$), depends on $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$, an *exact* minimum of (Q-NLP). As such, the feasibility of this program, ($\mathrm{LP}'_{\mathrm{adv}}$), is established using the Lagrange multipliers $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^{S\times B}$ associated with $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$.

**Lemma B.4.** The linear program Equation ($\mathrm{LP}_{\mathrm{adv}}$) with variables $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$ is feasible.

*Proof.* We introduce the following auxiliary linear program with variables $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$:

$$
(\mathrm{LP}'_{\mathrm{adv}})\quad
\begin{aligned}
\max \quad & \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} \lambda(s,b) r\left(s, \tilde{\boldsymbol{x}}, b\right) \\
\text{s.t.} \quad & \rho(\bar{s}) + \sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}} \Big[\lambda(s,b)\gamma\, \mathbb{P}\left(\bar{s}|s,\tilde{\boldsymbol{x}},b\right)\Big] - \sum_{b\in\mathcal{B}} \lambda(\bar{s},b) = 0, && (\mathrm{LP}'_{\mathrm{adv}}.1)\\
& \qquad\qquad \forall (s,b)\in\mathcal{S}\times\mathcal{B}; \\
& \sum_{b\in\mathcal{B}} \lambda(s,b)\left[r\left(s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k}),b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k}),b\right)\tilde{v}(s') - \tilde{v}(s)\right] && (\mathrm{LP}'_{\mathrm{adv}}.2)\\
& \qquad\qquad \geq -4\epsilon\ell,\ \forall k\in[n], \forall (s,a)\in\mathcal{S}\times\mathcal{A}_k; \\
& \lambda(s,b)\left(\left[r\left(s,\tilde{\boldsymbol{x}},b\right) + \gamma\textstyle\sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\tilde{\boldsymbol{x}},b\right)\tilde{v}(s')\right] - \tilde{v}(s)\right) = 0, && (\mathrm{LP}'_{\mathrm{adv}}.3)\\
& \qquad\qquad \forall (s,b)\in\mathcal{S}\times\mathcal{B}; \\
& \lambda(s,b) \geq 0,\ \forall (s,b)\in\mathcal{S}\times\mathcal{B}. && (\mathrm{LP}'_{\mathrm{adv}}.4)
\end{aligned}
$$

Again, the objective function of ($\mathrm{LP}'_{\mathrm{adv}}$) is not relevant for our argument. For our purposes, it suffices to show that ($\mathrm{LP}'_{\mathrm{adv}}$) is feasible.

**Lemma B.5.** Let $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^{S\times B}$ be a subset of Lagrange multipliers associated with $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X}\times\mathbb{R}^S$ of (Q-NLP). Then, $\tilde{\boldsymbol{\lambda}}$ satisfies all the constraints of Equation ($\mathrm{LP}'_{\mathrm{adv}}$).

*Proof.* First, ($\mathrm{LP}'_{\mathrm{adv}}.1$) is satisfied by the first-order stationarity condition Equation (19a); ($\mathrm{LP}'_{\mathrm{adv}}.3$) is satisfied by the complementary slackness condition Equation (19b); and Equation ($\mathrm{LP}'_{\mathrm{adv}}.4$) by the nonnegative of the Lagrange multipliers Equation (19c). The rest of the proof is devoted to showing that $\tilde{\boldsymbol{\lambda}}$ also satisfies ($\mathrm{LP}'_{\mathrm{adv}}.2$). To this end, we first recall that, by Claim B.2, we have that

$$
\tilde{\omega}(k,s) - \tilde{\psi}(k,s) = -\tilde{v}(s)\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b) + 2\ell(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s})^\top \tilde{\boldsymbol{x}}_{k,s},
$$

for any $s\in\mathcal{S}$. Combing this relation with Equation (16) we get that for any $k\in[n]$, $(s,a)\in\mathcal{S}\times\mathcal{A}_k$,

$$
\begin{aligned}
\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b)\Big[r\left(s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) + \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) v(s')\Big] + 2\ell(\tilde{x}_{k,s,a} - \hat{x}_{k,s,a}) \\
-\tilde{v}(s)\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b) + 2\ell\tilde{\boldsymbol{x}}_{k,s}^\top(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s}) - \tilde{\zeta}(k,s,a) = 0 \\
\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b)\Big[r\left(s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) + \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) v(s')\Big] - \sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b)\tilde{v}(s) \\
+2\ell(\tilde{x}_{k,s,a} - \hat{x}_{k,s,a}) - 2\ell\tilde{\boldsymbol{x}}_{k,s}^\top(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s}) = \tilde{\zeta}(k,s,a).
\end{aligned}
$$

As a result, we conclude that

$$
\sum_{b\in\mathcal{B}} \tilde{\lambda}(s,b)\Big[r\left(s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) + \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,(\boldsymbol{e}_{k,s,a};\tilde{\boldsymbol{x}}_{-k,s}),b\right) v(s') - \tilde{v}(s)\Big] \geq -4\epsilon\ell,
$$

since

(i) $\tilde{\zeta}(k,s,a) \geq 0$ by Equation (19c);

(ii) $2\ell(\hat{x}_{k,s,a} - \tilde{x}_{k,s,a}) \geq -2\ell|\hat{x}_{k,s,a} - \tilde{x}_{k,s,a}| \geq -2\ell\epsilon$ given that $\|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|_\infty \leq \|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|_2 \leq \epsilon$; and

(iii) $2\ell\tilde{\boldsymbol{x}}_{k,s}^\top(\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s}) \geq -2\ell\|\tilde{\boldsymbol{x}}_{k,s}\|_2\|\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s}\|_2 \geq -2\ell\epsilon$, by Cauchy-Schwarz inequality and the fact that $\|\tilde{\boldsymbol{x}}_{k,s}\|_2 \leq 1$ since $\tilde{\boldsymbol{x}}_{k,s} \in \Delta(\mathcal{A}_k)$.

This concludes the proof of the lemma. □

We next leverage this lemma to establish that the original linear program is also feasible. To do so, we will leverage the Lipschitz continuity of the constraint functions. In particular, consider any $(s, b) \in \mathcal{S} \times \mathcal{B}$. We observe that

$$\begin{aligned}
&r\left(s, \tilde{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) \tilde{v}(s') - \tilde{v}(s) \\
&\quad = r\left(s, \tilde{\boldsymbol{x}}, b\right) + r\left(s, \hat{\boldsymbol{x}}, b\right) - r\left(s, \hat{\boldsymbol{x}}, b\right) \\
&\quad + \gamma \sum_{s' \in \mathcal{S}} \Big( \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) + \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) - \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \Big) \Big( \tilde{v}(s') + \hat{v}(s') - \hat{v}(s') \Big) \\
&\quad - \tilde{v}(s) + \hat{v}(s) - \hat{v}(s).
\end{aligned}$$

Thus,

$$\begin{aligned}
r\left(s, \tilde{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) \hat{v}(s') - \tilde{v}(s) = & r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') - \hat{v}(s) \\
& + r\left(s, \tilde{\boldsymbol{x}}, b\right) - r\left(s, \hat{\boldsymbol{x}}, b\right) \\
& + \gamma \sum_{s' \in \mathcal{S}} \Big( \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) - \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \Big) \tilde{v}(s') \\
& + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \left( \tilde{v}(s') - \hat{v}(s') \right) \\
& - \tilde{v}(s) + \hat{v}(s). \qquad (27)
\end{aligned}$$

As a result, given that

$$\tilde{\lambda}(s, b) \left( \left[ r\left(s, \tilde{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) \tilde{v}(s') \right] - \tilde{v}(s) \right) = 0,$$

it follows that from (27) and the triangle inequality that

$$\begin{aligned}
&\left| \tilde{\lambda}(s, b) \left( \left[ r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s'} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) \right| \leq \\
&\qquad \frac{1}{1-\gamma} \left( \sqrt{\textstyle\sum_{k=1}^n A_k} + \gamma S \sqrt{\textstyle\sum_{k=1}^n A_k} \frac{1}{1-\gamma} + \gamma S L + L \right) \epsilon.
\end{aligned}$$

This inequality uses that $\|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\| \leq \epsilon$; the fact that $\tilde{\lambda}(s, b) \leq \frac{1}{1-\gamma}$ Equation ($\text{LP}_{\text{adv}}$.5); and the Lipschitz continuity bounds provided in Claim D.9:

$$\begin{cases}
\left| r\left(s, \tilde{\boldsymbol{x}}, b\right) - r\left(s, \hat{\boldsymbol{x}}, b\right) \right| \leq \sqrt{\sum_{k=1}^n A_k} \epsilon; \\
\left| \sum_{s' \in \mathcal{S}} \Big( \mathbb{P}\left(s' | s, \tilde{\boldsymbol{x}}, b\right) - \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \Big) \tilde{v}(s') \right| \leq S \sqrt{\sum_{k=1}^n A_k} \frac{1}{1-\gamma} \epsilon; \\
\left| \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \left( \tilde{v}(s') - \hat{v}(s') \right) \right| \leq S L \epsilon; \text{ and} \\
\left| \tilde{v}(s) - \hat{v}(s) \right| \leq L \epsilon.
\end{cases}$$

We proceed in a similar manner for ($\text{LP}_{\text{adv}}$.1), yielding that

$$\begin{aligned}
&\sum_{b \in \mathcal{B}} \tilde{\lambda}(s, b) \left[ r\left(s, (\boldsymbol{e}_{k,s,a}; \hat{\boldsymbol{x}}_{-k}), b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{e}_{k,s,a}; \hat{\boldsymbol{x}}_{-k}), b\right) \hat{v}(s') - \hat{v}(s) \right] \geq \\
&\qquad \geq -4\epsilon\ell - \frac{1}{1-\gamma} \left( \sqrt{\textstyle\sum_{k=1}^n A_k} + \gamma S \sqrt{\textstyle\sum_{k=1}^n A_k} \frac{1}{1-\gamma} + \gamma S L + L \right) \epsilon.
\end{aligned}$$

Thus, $\tilde{\boldsymbol{\lambda}}$ satisfies ($\text{LP}_{\text{adv}}$.1). Finally, $\tilde{\boldsymbol{\lambda}}$ also satisfies ($\text{LP}_{\text{adv}}$.4) and ($\text{LP}_{\text{adv}}$.5), implied directly by Proposition B.2 and Claim D.7. □

**Lemma B.6.** Let $\hat{\boldsymbol{x}}$ be an $\epsilon$-nearly stationary point of $\phi(\cdot) = \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\cdot, \boldsymbol{y})$. Any feasible solution $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$ of Equation (LP$_{\text{adv}}$) induces an $O(\epsilon)$-approximate Nash equilibrium for the adversarial team Markov game.

*Proof.* Consider any feasible solution $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$ of (LP$_{\text{adv}}$), and the induced strategy for the adversary defined as

$$\hat{y}_{s,b} := \frac{\lambda(s,b)}{\sum_{b\in\mathcal{B}} \lambda(s,b)},$$

for any $(s,b) \in \mathcal{S}\times\mathcal{B}$; this is indeed well-defined since $\sum_{b\in\mathcal{B}} \lambda(s,b) \geq \rho(s) > 0$, which in turn follows since $\boldsymbol{\rho}$ has full support. We will show that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $O(\epsilon)$-approximate Nash equilibrium. Our proof proceeds in two parts. First, we show that, if the team is responding according to $\hat{\boldsymbol{x}}$, then $\hat{y}$ is an $O(\epsilon)$-approximate best response for the adversary. Analogously, in the second part of the proof we argue about deviations from team players.

**Controlling deviations of the adversary.** Fix any $\boldsymbol{y} \in \mathcal{Y}$. Given that $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$ is a feasible solution of Equation (Q-NLP), it follows that for any $(s,b) \in \mathcal{S}\times\mathcal{B}$,

$$y_{s,b}\left(r(s,\hat{\boldsymbol{x}},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\hat{\boldsymbol{x}},b)\hat{v}(s')\right) \leq \hat{v}(s) y_{s,b},$$

Summing over all $b \in \mathcal{B}$ yields that

$$\sum_{b\in\mathcal{B}} y_{s,b}\left(r(s,\hat{\boldsymbol{x}},b) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\hat{\boldsymbol{x}},b)\hat{v}(s')\right) \leq \hat{v}(s),$$

in turn implying that

$$r\left(s,\hat{\boldsymbol{x}},\boldsymbol{y}\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},\boldsymbol{y}\right)\hat{v}(s') \leq \hat{v}(s),$$

for any $s \in \mathcal{S}$. The last inequality can be succinctly expressed in the following vector (element-wise) inequality:

$$\boldsymbol{r}\left(\hat{\boldsymbol{x}},\boldsymbol{y}\right) + \gamma\, \mathbb{P}\left(\hat{\boldsymbol{x}},\boldsymbol{y}\right)\hat{\boldsymbol{v}} \leq \hat{\boldsymbol{v}}.$$

From this inequality it follows that

$$\hat{\boldsymbol{v}} \geq \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}^t(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\boldsymbol{r}(\hat{\boldsymbol{x}},\boldsymbol{y}) = (\mathbf{I} - \gamma\, \mathbb{P}(\hat{\boldsymbol{x}},\boldsymbol{y}))^{-1}\boldsymbol{r}(\hat{\boldsymbol{x}},\boldsymbol{y}) = \boldsymbol{V}(\hat{\boldsymbol{x}},\boldsymbol{y}), \tag{28}$$

where we used Claims D.1, D.2 and D.5, and the notation $\boldsymbol{V}(\hat{\boldsymbol{x}},\boldsymbol{y})$ to represent the value vector under $(\hat{\boldsymbol{x}},\boldsymbol{y})$—recall (3). Moreover, given that $\boldsymbol{\lambda}$ is a feasible solution of (LP$_{\text{adv}}$), manipulating (LP$_{\text{adv}}$.3) yields that for any $(s,b) \in \mathcal{S}\times\mathcal{B}$,

$$\lambda(s,b)\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right] - \hat{v}(s)\right) \geq -c_2\epsilon$$

$$\frac{\lambda(s,b)}{\sum_{b'\in\mathcal{B}} \lambda(s,b')}\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right] - \hat{v}(s)\right) \geq -\frac{c_2\epsilon}{\sum_{b'\in\mathcal{B}} \lambda(s,b')} \tag{29}$$

$$\hat{y}_{s,b}\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right] - \hat{v}(s)\right) \geq -\frac{c_2\epsilon}{\sum_{b'\in\mathcal{B}} \lambda(s,b')}, \tag{30}$$

where (29) follows since $\sum_{b'\in\mathcal{B}} \lambda(s,b') > 0$, while (30) follows from the definition of $\hat{y}_{s,b}$. Summing over all $b \in \mathcal{B}$,

$$\sum_{b\in\mathcal{B}} \hat{y}_{s,b}\left(\left[r\left(s,\hat{\boldsymbol{x}},b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right] - \hat{v}(s)\right) \geq -\sum_{b\in\mathcal{B}} \frac{c_2\epsilon}{\sum_{b'\in\mathcal{B}} \lambda(s,b')}$$

$$\sum_{b\in\mathcal{B}} \hat{y}_{s,b}\left(r\left(s,\hat{\boldsymbol{x}},b\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},b\right)\hat{v}(s')\right) - \hat{v}(s) \geq -B\frac{c_2\epsilon}{\sum_{b\in\mathcal{B}} \lambda(s,b)}$$

$$r\left(s,\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\right) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}\left(s'|s,\hat{\boldsymbol{x}},,\hat{\boldsymbol{y}}\right)\hat{v}(s') \geq \hat{v}(s) - B\frac{c_2\epsilon}{\sum_{b\in\mathcal{B}} \lambda(s,b)}. \tag{31}$$

Let us set $\xi_s := \frac{c_2 \cdot \epsilon}{\sum_b \lambda(s,b)}$ for each $s \in \mathcal{S}$. Continuing from (31), we have that

$$\boldsymbol{r}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) + \gamma \mathbb{P}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \hat{\boldsymbol{v}} \geq \hat{\boldsymbol{v}} - B\boldsymbol{\xi},$$

which in turn implies that

$$\boldsymbol{V}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \geq \boldsymbol{v} - B\left(\mathbf{I} - \gamma \mathbb{P}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})\right)^{-1} \boldsymbol{\xi},$$

by Claims D.2 and D.5. Thus,

$$\begin{aligned}
V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) &\geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - B\boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma \mathbb{P}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})\right)^{-1} \boldsymbol{\xi} \\
&\geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - B\boldsymbol{\xi}^\top \boldsymbol{d}_{\boldsymbol{\rho}}^{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}} && (32) \\
&\geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - c_2 B \sum_{s \in \mathcal{S}} \frac{d_{\boldsymbol{\rho}}^{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}}(s)}{\sum_{b \in \mathcal{B}} \lambda(s,b)} \epsilon \\
&\geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - c_2 B \sum_{s \in \mathcal{S}} \frac{d_{\boldsymbol{\rho}}^{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}}(s)}{\rho(s)} \epsilon && (33) \\
&\geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - c_2 BSD\epsilon, && (34)
\end{aligned}$$

where (32) follows from Claim D.3; (33) follows from the feasibility constraint $\sum_{b \in \mathcal{B}} \lambda(s,b) \geq \rho(s)$; and (34) uses the definition of mismatch coefficient (Definition A.6). As a result, combining (28) and Equation (34), we conclude that for any $\boldsymbol{y} \in \mathcal{Y}$,

$$V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - c_2 BSD\epsilon \geq V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) - c_2 BSD\epsilon. \tag{35}$$

**Controlling deviations of a team player.** Next, we show that any deviation from a single player can only yield a small improvement for the player. Fix any player $k \in [n]$ and strategy $\boldsymbol{x}_k \in \mathcal{X}_k$. The proof proceeds analogously to our previous argument. In particular, for any state $s \in \mathcal{S}$, multiplying Equation ($\text{LP}_{\text{adv}}$.1) by $\boldsymbol{x}_{k,s,a}$, and summing over all actions $a \in \mathcal{A}_k$ yields that

$$\sum_{b \in \mathcal{B}} \lambda(s,b) \left[ r\left(s, (\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, (\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), b\right) \hat{v}(s') \right] \geq \hat{v}(s) \sum_{b \in \mathcal{B}} \lambda(s,b) - c_1 \epsilon;$$

here, we leveraged the feasibility of $\boldsymbol{\lambda}$. Further, given that $\sum_{b \in \mathcal{B}} \lambda(s,b) > 0$,

$$r(s, (\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' | s, (\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}) \geq \hat{v}(s) - c_1 \cdot \epsilon \frac{1}{\sum_{b \in \mathcal{B}} \lambda(s,b)} \geq \hat{v}(s) - c_1 \epsilon \frac{1}{\rho(s)},$$

for any $s \in \mathcal{S}$, since $\sum_{b \in \mathcal{B}} \lambda(s,b) \geq \rho(s)$. Hence,

$$\boldsymbol{r}\left((\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}\right) + \gamma \mathbb{P}\left((\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}\right) \hat{\boldsymbol{v}} \geq \hat{\boldsymbol{v}} - c_1 \epsilon \frac{1}{\boldsymbol{\rho}}.$$

In turn, by Claim D.5, this implies that

$$\boldsymbol{V}((\boldsymbol{x}_k, \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}) \geq \hat{\boldsymbol{v}} - c_1 \cdot \epsilon (\mathbf{I} - \gamma \mathbb{P}((\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}))^{-1} \frac{1}{\boldsymbol{\rho}}.$$

Thus, we conclude that

$$V_{\boldsymbol{\rho}}(\boldsymbol{x}_k, \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}) \geq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} - c_1 DS\epsilon, \tag{36}$$

where we used Claim D.3 and Definition A.6. Next, using ($\text{LP}_{\text{adv}}$.2) we obtain that for all $(s,b) \in \mathcal{S} \times \mathcal{B}$,

$$\begin{aligned}
\lambda(s,b) \left( \left[ r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) &\leq c_2 \epsilon \\
\frac{\lambda(s,b)}{\sum_{b' \in \mathcal{B}} \lambda(s,b')} \left( \left[ r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) &\leq \frac{c_2 \epsilon}{\sum_{b' \in \mathcal{B}} \lambda(s,b')}.
\end{aligned}$$

For convenience, let us set $\xi_s \coloneqq \frac{c_2 \epsilon}{\sum_{b' \in \mathcal{B}} \lambda(s,b')}$. By definition of $\hat{\boldsymbol{y}}$, we have

$$\hat{y}_{s,b} \left( \left[ r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) \leq \xi_s$$

$$\sum_{b \in \mathcal{B}} \hat{y}_{s,b} \left( \left[ r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) \leq B\xi_s$$

$$\sum_{b \in \mathcal{B}} \hat{y}_{s,b} \left( r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right) \leq \hat{v}(s) + B\xi_s$$

$$r\left(s, \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\right) \hat{v}(s') \leq \hat{v}(s) + B\xi_s,$$

for any $s \in \mathcal{S}$. Thus,

$$\boldsymbol{r}\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\right) + \gamma \mathbb{P}\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\right) \hat{\boldsymbol{v}} \leq \hat{\boldsymbol{v}} + B\boldsymbol{\xi}$$

$$\boldsymbol{V}\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\right) \leq \hat{\boldsymbol{v}} + B\left(\mathbf{I} - \gamma \mathbb{P}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})\right)^{-1} \boldsymbol{\xi} \tag{37}$$

$$V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \leq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} + Bc_2 \sum_{s \in \mathcal{S}} \frac{d_{\boldsymbol{\rho}}^{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}}(s)}{\sum_{b \in \mathcal{B}} \lambda(s,b)} \epsilon \tag{38}$$

$$V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \leq \boldsymbol{\rho}^\top \hat{\boldsymbol{v}} + c_2 BSD\epsilon, \tag{39}$$

where (37) follows from Claim D.5; (38) follows from Claim D.3; and (39) follows from the fact that $\sum_{b \in \mathcal{B}} \lambda(s,b) \geq \rho(s)$ and Definition A.6. As a result, combining (36) and (39) we conclude that

$$V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \leq V_{\boldsymbol{\rho}}\left((\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}\right) + c_2 BSD\epsilon + c_1 DS\epsilon. \tag{40}$$

□

We state the precise version of Lemma B.6 in Theorem B.1 below. First, let us summarize `AdvNashPolicy`, the algorithm for computing the policy for the adversary. `AdvNashPolicy`, described in Algorithm 2, takes as input $\hat{\boldsymbol{x}} \in \mathcal{X}$, an $\epsilon$-nearly stationary point of $\phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$. The algorithm then computes the best-response value vector $\hat{\boldsymbol{v}}$. This is computed by fixing the strategy of the team $\hat{\boldsymbol{x}} \in \mathcal{X}$, and then solving the single-agent MDP problem so as to maximize the value at every state. Then, the pair $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{v}})$ is used in order to determine the—polynomial number of—coefficients of $\text{LP}_{\text{adv}}$, as introduced in ($\text{LP}_{\text{adv}}$). Then, any feasible solution $\boldsymbol{\lambda} \in \mathbb{R}^{S \times B}$ of ($\text{LP}_{\text{adv}}$) is used to determine the strategy of the adversary as follows.

$$\hat{y}_{s,b} \coloneqq \frac{\lambda(s,b)}{\sum_{b \in \mathcal{B}} \lambda(s,b)}, \quad \forall (s,b) \in \mathcal{S} \times \mathcal{B}.$$

**Algorithm 2** `AdvNashPolicy`

**Input:** An $\epsilon$-nearly stationary point $\hat{\boldsymbol{x}} \in \mathcal{X}$ of $\phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$

1: $\hat{\boldsymbol{v}} \leftarrow$ The best-response value vector for the adversary
2: $\text{LP}_{\text{adv}} \leftarrow$ Compute the coefficients of the linear program ($\text{LP}_{\text{adv}}$)
3: $\boldsymbol{\lambda} \leftarrow$ Any feasible solution of $\text{LP}_{\text{adv}}$
4: $\hat{y}_{s,b} \leftarrow \frac{\lambda(s,b)}{\sum_{b \in \mathcal{B}} \lambda(s,b)}$
**return** $\hat{\boldsymbol{y}}$

**Remark 3.** *In Lines 1 and 2 of Algorithm 2 we adopt the assumption of "polynomially accessible games," which is standard in computational game theory. More precisely, we assume that we have an oracle that outputs the expected reward and the expected transition probabilities in any state given the announced (mixed) strategies of every player. It is known that such an oracle for the expected utilities can be implemented in polynomial time for virtually all*[4] *interesting classes of*

[4]Nevertheless, there are certain rather artificial succinct games for which computing the expected utilities is #P-hard (Daskalakis et al., 2006, Proposition 1).

succinctly representable *normal-form games (Papadimitriou & Roughgarden, 2008), and even for extensive-form games (Huang & von Stengel, 2008). Crucially, without a succinct representation or an oracle access to the game the input scales exponentially with the number of players, trivializing the computational aspects of the problem in multiplayer games; this issue has been discussed extensively for normal-form games with regards to Nash equilibria (Daskalakis et al., 2009), and correlated equilibria (Papadimitriou & Roughgarden, 2008).*

*In our case, we will essentially assume that the adversary has a "polynomially accessible environment." More explicitly, the adversary should have access to the expected rewards and the expected transition probabilities of the single-agent MDP induced if the rest of the players have announced and fixed their policies. Further, in* IPGMAX *we also make the standard assumption that players in the team have access to the gradients of the value function.*

**Theorem B.1** (Near stationary points extend to approximate NE)**.** *Consider an adversarial team Markov game $\mathcal{G}$, and suppose that $\hat{\boldsymbol{x}} \in \mathcal{X}$ is an $\epsilon$-nearly stationary point of $\phi(\boldsymbol{x}) := \max_{\boldsymbol{y}} V_{\boldsymbol{\rho}}(\cdot, \boldsymbol{y})$, where $V_{\boldsymbol{\rho}}$ is the value function of $\mathcal{G}$ (3). Then, any feasible solution of Equation (LP$_{\text{adv}}$) $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{S\times B}_{\geq 0}$ induces a strategy $\hat{\boldsymbol{y}}$, defined as*

$$\hat{y}_{s,b} := \frac{\hat{\lambda}(s,b)}{\sum_{b\in\mathcal{B}} \hat{\lambda}(s,b)}, \quad \forall (s,b) \in \mathcal{S}\times\mathcal{B},$$

*so that for any player $k \in [n]$ and any deviations $\boldsymbol{x}_k \in \mathcal{X}_k$ and $\boldsymbol{y} \in \mathcal{Y}$,*

$$\begin{cases} V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \;\leq\; V_{\boldsymbol{\rho}}\left((\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{-k}), \hat{\boldsymbol{y}}\right) + (BSD+1)\frac{1}{1-\gamma}\left(\sqrt{\sum_{k=1}^n A_k}(1+\gamma S\frac{1}{1-\gamma}) + \gamma SL + L\right)\epsilon \\ \qquad\qquad + \frac{1}{1-\gamma}4\epsilon\ell \\ V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \;\geq\; V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) - BSD\frac{1}{1-\gamma}\left(\sqrt{\sum_{k=1}^n A_k}(1+\gamma S\frac{1}{1-\gamma}) + \gamma SL + L\right)\epsilon, \end{cases}$$

*Here, we recall that $D = \max_{\boldsymbol{\pi}\in\Pi} \left\| \frac{\boldsymbol{d}^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}}{\boldsymbol{\rho}} \right\|_\infty$ is the mismatch coefficient, $L = \frac{\sqrt{\sum_k A_k + B}}{(1-\gamma)^2}$ is a Lipschitz constant of the value function, and $\ell = \frac{2\left(\sum_k A_k + B\right)}{(1-\gamma)^3}$ is a smoothness constant of the value function (Lemma C.1).*

*Proof.* By Lemma B.4, we know that (LP$_{\text{adv}}$) is feasible. Further, $\hat{\boldsymbol{y}}$ is a well-formed strategy since for any feasible $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}_{\geq 0}$ of (LP$_{\text{adv}}$) it holds that $\sum_{b\in\mathcal{B}} \lambda(s,b) \geq \rho(s) > 0$, for any state $s \in \mathcal{S}$, where the first bound follows by feasibility of $\boldsymbol{\lambda}$ and the second since $\boldsymbol{\rho}$ is assumed to have full support. Thus, the proof of the theorem follows from Lemma B.6, and in particular (35) and (40). □

## C Convergence to a Nearly Stationary Point

In this section, we establish that IPGMAX reaches to an $\epsilon$-nearly stationary point—in the sense of Definition A.4—after a number of iterations that is polynomial in all the natural parameters of the game, as well as $1/\epsilon$. The main result here is Proposition 3.1, which was first introduced in Section 3.2. First, we need to establish that the value function $V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ is Lipschitz continuous and smooth, as formalized below. We note that this property is by now fairly standard (*e.g.*, see (Agarwal et al., 2020)), and we therefore omit the proof.

**Lemma C.1.** For any initial distribution $\boldsymbol{\rho}$, the value function $V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ is $\frac{\sqrt{\sum_k A_k + B}}{(1-\gamma)^2}$-Lipschitz continuous and $\frac{2\left(\sum_k A_k + B\right)}{(1-\gamma)^3}$-smooth:

$$|V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) - V_{\boldsymbol{\rho}}(\boldsymbol{x}', \boldsymbol{y}')| \leq \frac{\sqrt{\sum_{k=1}^n A_k + B}}{(1-\gamma)^2} \left\|(\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}')\right\|; \text{ and}$$

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

for all $(\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{X}\times\mathcal{Y}$.

We convenience, we will let $L := \frac{\sqrt{\sum_{k=1}^n A_k + B}}{(1-\gamma)^2}$ and $\ell := \frac{2\left(\sum_{k=1}^n A_k + B\right)}{(1-\gamma)^3}$. The next key result characterizes the iteration complexity required to reach an $\epsilon$-nearly stationary point of $\phi(\cdot)$. The following analysis follows (Jin et al., 2020).

**Proposition 3.1.** Consider any $\epsilon > 0$. If $\eta = 2\epsilon^2(1-\gamma)$ and $T = \frac{(1-\gamma)^4}{8\epsilon^4(\sum_{k=1}^n A_k + B)^2}$, there exists an iterate $t^\star$, with $0 \le t^\star \le T-1$, such that $\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|_2 \le \epsilon$, where $\tilde{\boldsymbol{x}}^{(t^\star)} := \mathrm{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t^\star)})$.

*Proof.* By virtue of the $\ell$-smoothness of $V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ (Lemma C.1), it follows that for any $\boldsymbol{x} \in \mathcal{X}$ and $0 \le t \le T-1$,

$$\phi(\boldsymbol{x}) \ge V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}^{(t+1)}) \ge V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) + \langle \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}), \boldsymbol{x} - \boldsymbol{x}^{(t)} \rangle - \frac{\ell}{2}\left\|\boldsymbol{x} - \boldsymbol{x}^{(t)}\right\|^2, \tag{41}$$

since $\phi(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) \ge V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}^{(t+1)})$. Now recall that

$$\tilde{\boldsymbol{x}}^{(t)} := \arg\min_{\boldsymbol{x}' \in \mathcal{X}} \left\{ \phi(\boldsymbol{x}') + \frac{1}{2\lambda}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}'\right\|^2 \right\}, \tag{42}$$

for any $0 \le t \le T-1$, where $\lambda := \frac{1}{2\ell}$. Using the definition of Moreau envelope (Definition 3.1),

$$\begin{aligned}
\phi_\lambda(\boldsymbol{x}^{(t+1)}) &\le \phi(\tilde{\boldsymbol{x}}^{(t)}) + \ell\left\|\boldsymbol{x}^{(t+1)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2 \\
&\le \phi(\tilde{\boldsymbol{x}}^{(t)}) + \ell\left\|\mathrm{Proj}_{\mathcal{X}}\left(\boldsymbol{x}^{(t)} - \eta\nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})\right) - \mathrm{Proj}_{\mathcal{X}}\left(\tilde{\boldsymbol{x}}^{(t)}\right)\right\|^2 && (43) \\
&\le \phi(\tilde{\boldsymbol{x}}^{(t)}) + \ell\left\|\boldsymbol{x}^{(t)} - \eta\nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) - \tilde{\boldsymbol{x}}^{(t)}\right\|_2^2 && (44) \\
&\le \phi(\tilde{\boldsymbol{x}}^{(t)}) + \ell\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2 + \eta^2\ell\left\|\nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})\right\|^2 \\
&\qquad + 2\eta\ell\langle \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}), \tilde{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)} \rangle && (45) \\
&\le \phi_\lambda(\boldsymbol{x}^{(t)}) + 2\eta\ell\left(\phi(\tilde{\boldsymbol{x}}^{(t)}) - \phi(\boldsymbol{x}^{(t)}) + \frac{\ell}{2}\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2\right) + \eta^2\ell L^2, && (46)
\end{aligned}$$

where

- (43) uses the fact that $\boldsymbol{x}_k^{(t+1)} := \mathrm{Proj}_{\mathcal{X}_k}\left(\boldsymbol{x}_k^{(t)} - \eta\nabla_{\boldsymbol{x}_k} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})\right)$ for all $k \in [n]$, as defined in IPGMAX, in turn implying that $\boldsymbol{x}^{(t+1)} = \mathrm{Proj}_{\mathcal{X}}\left(\boldsymbol{x}^{(t)} - \eta\nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})\right)$, as well as the fact that $\mathrm{Proj}_{\mathcal{X}}\left(\tilde{\boldsymbol{x}}^{(t)}\right) = \tilde{\boldsymbol{x}}^{(t)}$ since $\tilde{\boldsymbol{x}}^{(t)} \in \mathcal{X}$;
- (44) follows from the fact that the projection operator is nonexpansive (Fact D.2);
- (45) uses the identity $\|\boldsymbol{a} + \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 + 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$; and
- (46) follows since
  - (i) $\phi(\tilde{\boldsymbol{x}}^{(t)}) + \ell\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2 = \min_{\boldsymbol{x}' \in \mathcal{X}}\left\{\phi(\boldsymbol{x}') + \ell\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}'\right\|^2\right\} = \phi_\lambda(\boldsymbol{x}^{(t)})$ by definition of $\tilde{\boldsymbol{x}}^{(t)}$ in (42) and the definition of Moreau envelope with $\lambda = \frac{1}{2\ell}$ (Definition 3.1);
  - (ii) $V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) + \langle \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}), \tilde{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)} \rangle - \frac{\ell}{2}\left\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^{(t)}\right\|^2 \le \phi(\tilde{\boldsymbol{x}}^{(t)})$, which is an application of (41) for $\boldsymbol{x} := \tilde{\boldsymbol{x}}^{(t)}$; and
  - (iii) $\left\|\nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})\right\|^2 \le L^2$ by $L$-Lipschitz continuity of $V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})$ (Lemma C.1) combined with Fact D.1.

As a result, taking a telescopic sum of (46) for all $0 \le t \le T-1$ and rearranging the terms yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left(\phi(\boldsymbol{x}^{(t)}) - \phi(\tilde{\boldsymbol{x}}^{(t)}) - \frac{\ell}{2}\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2\right) &\le \frac{\phi_\lambda(\boldsymbol{x}^{(0)}) - \phi_\lambda(\boldsymbol{x}^{(T)})}{2\eta\ell T} + \frac{\eta L^2}{2} \\
&\le \frac{1}{2(1-\gamma)\eta\ell T} + \frac{\eta L^2}{2}, && (47)
\end{aligned}$$

since $\phi_\lambda(\boldsymbol{x}^{(T)}) \geq 0$, directly by Definition 3.1, and $\phi_\lambda(\boldsymbol{x}^{(0)}) \leq \phi(\boldsymbol{x}^{(0)}) \leq \frac{1}{1-\gamma}$, where the last inequality follows from Claim D.8. Therefore we conclude that there exists an iterate $t^\star$, with $0 \leq t^\star \leq T-1$, so that

$$\phi(\boldsymbol{x}^{(t^\star)}) - \phi(\tilde{\boldsymbol{x}}^{(t^\star)}) - \frac{\ell}{2}\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2 \leq \frac{1}{2(1-\gamma)\eta\ell T} + \frac{\eta L^2}{2}. \tag{48}$$

Further, since $\phi(\boldsymbol{x}) + \ell\left\|\boldsymbol{x} - \boldsymbol{x}^{(t^\star)}\right\|^2$ is $\ell$-strongly convex with respect to $\boldsymbol{x}$ (by Lemma A.1 and Corollary A.2), we get that

$$\phi(\boldsymbol{x}^{(t^\star)}) - \phi(\tilde{\boldsymbol{x}}^{(t^\star)}) - \ell\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2 \geq \frac{\ell}{2}\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2,$$

by definition of $\tilde{\boldsymbol{x}}^{(t^\star)}$ in (42), in turn implying that

$$\phi(\boldsymbol{x}^{(t^\star)}) - \phi(\tilde{\boldsymbol{x}}^{(t^\star)}) - \frac{\ell}{2}\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2 \geq \ell\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2.$$

Combing this bound with (48) yields that

$$\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|^2 \leq \frac{1}{2(1-\gamma)\eta\ell^2 T} + \frac{\eta L^2}{2\ell}.$$

In particular, letting

$$\eta = \epsilon^2 \cdot \frac{\ell}{L^2} = 2\epsilon^2 \cdot (1-\gamma)$$

and

$$T = \frac{1}{\epsilon^2(1-\gamma)\eta\ell^2} = \frac{(1-\gamma)^4}{8\epsilon^4(\sum_{k=1}^n A_k + B)^2}$$

implies that $\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\| \leq \epsilon$. □

**Remark 4** (IPGMAX with approx. best-responses)**.** *Even if $\boldsymbol{y}^{(t)}$ is an $\epsilon'$-approximate maximizer of $V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \cdot)$, then Equation (41) will hold with an $\epsilon$ additive error. In turn, this translates to a $2\epsilon'\eta\ell$ additive factor in Equation (46). So, as long as $\epsilon' = O(\eta)$ the rest of the argument immediately applies. Finally, arguments in Appendix B.4 still hold modulo an additive error.*

A limitation of this proposition is that it only establishes a "best-iterate" guarantee. However, as we explained in Section 3.2, determining such an iterate could introduce a substantial computational overhead in the algorithm. For this reason, we provide a stronger guarantee below, showing that even a random iterate will also be nearly stationary with constant probability, leading to a practical implementation of IPGMAX.

**Corollary C.1.** Consider any $\epsilon > 0$, and suppose that $\eta = \epsilon^2(1-\gamma)$ and $T = \frac{(1-\gamma)^4}{2\epsilon^4(\sum_k A_k + B)^2}$. For any $\delta > 0$, if we select uniformly at random (with repetitions) a set $\mathcal{T}$ of $\lceil \log(1/\delta) \rceil$ indexes from the set $\{0, 1, \ldots, T-1\}$, then with probability at least $1-\delta$ there exists a $t' \in \mathcal{T}$ such that $\left\|\boldsymbol{x}^{(t')} - \tilde{\boldsymbol{x}}^{(t')}\right\| \leq \epsilon$, where $\tilde{\boldsymbol{x}}^{(t')} := \mathrm{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t')})$.

*Proof.* First, we claim that selecting uniformly at random an index $t'$ from the set $\{0, 1, \ldots, T-1\}$ will satisfy

$$\|\boldsymbol{x}^{(t')} - \tilde{\boldsymbol{x}}^{(t')}\|^2 \leq 2\epsilon^2$$

with probability at least $\frac{1}{2}$. To show this, let us define

$$g^{(t)} := \phi(\boldsymbol{x}^{(t)}) - \phi(\tilde{\boldsymbol{x}}^{(t)}) - \frac{\ell}{2}\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2,$$

for $t = 0, 1, \ldots, T-1$. By definition of $\tilde{\boldsymbol{x}}^{(t)}$ in (42), we have

$$g^{(t)} = \phi(\boldsymbol{x}^{(t)}) - \phi(\tilde{\boldsymbol{x}}^{(t)}) - \frac{\ell}{2}\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\|^2 \geq \ell\left\|\boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)}\right\| \geq 0, \tag{49}$$

for $0 \leq t \leq T-1$. Further, by (47) we have

$$\frac{1}{T}\sum_{t=0}^{T-1} g^{(t)} \leq \epsilon^2 \ell, \tag{50}$$

where we used that $\eta = 2\epsilon^2(1-\gamma)$ and $T = \frac{(1-\gamma)^4}{8\epsilon^4(\sum_{k=1}^n A_k + B)^2}$. As a result, we conclude that at least half of the indexes $t$ are such that $g^{(t)} \leq 2\epsilon^2\ell$. Indeed, the contrary case contradicts (50) given that $g^{(t)} \geq 0$ for all $t$. In turn, this implies our claim in light of (49). Finally, the proof of the corollary follows from a standard boosting argument, as well as rescaling $\epsilon$ by $\frac{1}{\sqrt{2}}$. □

**Theorem C.1** (Computing $\epsilon$-approximate NE)**.** *Consider an adversarial team Markov game $\mathcal{G}$. Running* IPGMAX *for $T = \frac{512 S^8 D^4 \left(\sum_{k=1}^n A_k + B\right)^4}{\epsilon^4(1-\gamma)^{12}}$ number of iterations and learning rate $\eta = \frac{\epsilon^2(1-\gamma)^9}{32 S^4 D^2 \left(\sum_{k=1}^n A_k + B\right)^3}$ yields a team strategy $\hat{\boldsymbol{x}} \in \mathcal{X}$ that can be extended to an $\epsilon$-approximate Nash equilibrium in polynomial time through the routine* `AdvNashPolicy`$(\hat{\boldsymbol{x}})$*, assuming a polynomially accessible environment for the adversary (Remark 3).*

*Proof.* In place of $\epsilon$ of Proposition 3.1 we set

$$\epsilon \leftarrow \frac{\epsilon}{\frac{1}{1-\gamma}\left[4\ell + (BSD+1)\left(\sqrt{\sum_{k=1}^n A_k} + \gamma S \sqrt{\sum_{k=1}^n A_k}\frac{1}{1-\gamma} + \gamma S L + L\right)\right]},$$

which allows us to compute an $\epsilon$-approximate Nash equilibrium by virtue of Theorem B.1. Then, the number of iterations reads

$$\begin{aligned} T &= \frac{(1-\gamma)^4}{8(1-\gamma)^4\epsilon^4(\sum_{k=1}^n A_k + B)^2}\left[4\ell + (BSD+1)\left(\sqrt{\sum_{k=1}^n A_k}(1+\gamma S\frac{1}{1-\gamma}) + \gamma S L + L\right)\right]^4 \\ &\leq \frac{8^3 S^8 D^4 \left(\sum_{k=1}^n A_k + B\right)^4}{\epsilon^4(1-\gamma)^{12}}, \end{aligned}$$

with a learning rate

$$\begin{aligned} \eta &= 2\epsilon^2(1-\gamma)(1-\gamma)^2\left(\frac{1}{1-\gamma}\left[4\ell + (BSD+1)\left(\sqrt{\sum_{k=1}^n A_k}(1+\gamma S\frac{1}{1-\gamma}) + \gamma S L + L\right)\right]\right)^{-2} \\ &\geq \frac{\epsilon^2(1-\gamma)^9}{32 S^4 D^2 \left(\sum_{k=1}^n A_k + B\right)^3}. \end{aligned}$$

Further, assuming a polynomially accessible environment for the adversary, `AdvNashPolicy` can be implemented in polynomial time via linear programming (Ye, 2011). □

## D Additional Auxiliary Claims

For the sake of readability, this section contains some simple and standard claims we used earlier in our proofs, but are only stated here.

**Fact D.1.** Let $f : \mathcal{X} \ni \boldsymbol{x} \mapsto \mathbb{R}$ be an $L$-Lipschitz continuous and differentiable function. Then,

$$\max_{\boldsymbol{x}\in\mathcal{X}} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\| \leq L.$$

**Fact D.2** (Projection operator is nonexpansive)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty, convex and compact set. Further, let $\mathrm{Proj}_{\mathcal{X}}(:)\,\mathbb{R}^d \to \mathcal{X}$ be the Euclidean projection operator defined as $\mathrm{Proj}_{\mathcal{X}}(:)\,\mathbb{R}^d \ni \boldsymbol{y} \mapsto \frac{1}{2}\arg\min_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\|^2$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\|\mathrm{Proj}_{\mathcal{X}}(\boldsymbol{x}) - \mathrm{Proj}_{\mathcal{X}}(\boldsymbol{y})\| \leq \|\boldsymbol{x}-\boldsymbol{y}\|.$$

.

In the rest of the claims, we are implicitly—for the sake of readability—fixing an adversarial team Markov game $(\mathcal{S}, \mathcal{A}, \mathcal{B}, r, \mathbb{P}, \gamma, \boldsymbol{\rho})$.

**Claim D.1.** Consider any joint stationary policy $\boldsymbol{\pi} \in \Pi$. For any $\gamma \in [0, 1)$, the matrix $\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})$ is invertible.

**Claim D.2.** Let $\boldsymbol{\pi} \in \Pi$ be a joint stationary policy. The value vector $\boldsymbol{V} \in \mathbb{R}^S$ can be expressed as

$$\boldsymbol{V} = \big(\mathbf{I} - \gamma \mathbb{P}(\boldsymbol{\pi})\big)^{-1} \boldsymbol{r}(\boldsymbol{\pi}),$$

where $\boldsymbol{r}(\boldsymbol{\pi})$ denotes the per-state reward under policy $\boldsymbol{\pi}$.

*Proof.* For any state $s \in \mathcal{S}$,

$$V_s(\boldsymbol{\pi}) = \boldsymbol{r}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) + \gamma^2\, \mathbb{P}^2(\boldsymbol{\pi}) + \cdots = \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}^t(\boldsymbol{\pi}) \boldsymbol{r}(\boldsymbol{\pi}).$$

But, given that the matrix $\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})$ is invertible (Claim D.1), we have

$$\sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}^t(\boldsymbol{\pi}) = \big(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\big)^{-1},$$

and the claim follows. □

**Claim D.3.** Consider a stationary joint policy $\boldsymbol{\pi} \in \Pi$. The discounted visitation measure $d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}(s)$ can be expressed as

$$\big(\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}\big)^\top = \boldsymbol{\rho}^\top \big(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\big)^{-1}.$$

**Claim D.4.** Consider a stationary joint strategy $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, and the visitation measure $\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}$, under some initial distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$. Then, the value function can be expressed as

$$V_{\boldsymbol{\rho}} = \sum_{s \in \mathcal{S}} d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s) r(s, \boldsymbol{x}, \boldsymbol{y}).$$

*Proof.* By definition of $\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}$, we have that for any $s \in \mathcal{S}$,

$$d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s) = \sum_{\overline{s} \in \mathcal{S}} \rho(\overline{s}) \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}\left(s^{(t)} = s \mid \boldsymbol{x}, \boldsymbol{y}, s^{(0)} = \overline{s}\right).$$

Similarly, the value function can be written as

$$V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{s \in \mathcal{S}} \sum_{\overline{s} \in \mathcal{S}} \rho(\overline{s}) \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}\left(s^{(t)} = s \mid \boldsymbol{x}, \boldsymbol{y}, s^{(0)} = \overline{s}\right) r(s, \boldsymbol{x}, \boldsymbol{y}) = \sum_{s \in \mathcal{S}} d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s) r(s, \boldsymbol{x}, \boldsymbol{y}).$$

□

**Claim D.5.** Let $\boldsymbol{\pi} \in \Pi$ be a joint stationary policy, $\boldsymbol{r}(\boldsymbol{\pi})$ be the reward vector under $\boldsymbol{\pi}$, and $\boldsymbol{v}, \boldsymbol{c} \in \mathbb{R}^S$. If $\boldsymbol{r}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) \boldsymbol{v} \leq \boldsymbol{v} + \boldsymbol{c}$, then it holds that

$$\boldsymbol{V}(\boldsymbol{\pi}) \leq \boldsymbol{v} + \big(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\big)^{-1} \boldsymbol{c}.$$

Similarly, if $\boldsymbol{r}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) \boldsymbol{v} \geq \boldsymbol{v} + \boldsymbol{c}$, then it holds that

$$\boldsymbol{V}(\boldsymbol{\pi}) \geq \boldsymbol{v} + \big(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\big)^{-1} \boldsymbol{c}.$$

*Proof.* Suppose that $\boldsymbol{r}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) \boldsymbol{v} \leq \boldsymbol{v} + \boldsymbol{c}$. Applying recursively this inequality, it follows that

$$\sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}^t(\boldsymbol{\pi}) \boldsymbol{r}(\boldsymbol{\pi}) - \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}^t(\boldsymbol{\pi}) \boldsymbol{c} \leq \boldsymbol{v}.$$

Combining this bound with Claims D.1 and D.2 implies that

$$\boldsymbol{V}(\boldsymbol{\pi}) - \Big(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\Big)^{-1} \boldsymbol{c} \leq \boldsymbol{v}.$$

The case where $\boldsymbol{r}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) \boldsymbol{v} \geq \boldsymbol{v} + \boldsymbol{c}$ admits an analogous proof. □

**Claim D.6.** Consider an adversarial team Markov game $\mathcal{G}$. Altering all the rewards by adding an additive constant $c \in \mathbb{R}$ yields a *strategically-equivalent* game $\mathcal{G}'$: any $\epsilon$-approximate Nash equilibrium in $\mathcal{G}'$ is also an $\epsilon$-approximate Nash equilibrium in $\mathcal{G}$, and *vice versa*.

*Proof.* By assumption, $r'(s, \boldsymbol{a}, b) = r(s, \boldsymbol{a}, b) + c$ for any $(s, \boldsymbol{a}, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Let $V'_{\boldsymbol{\rho}}$ be the value function in $\mathcal{G}'$. Then, for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} V'_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) &= \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{x}, \boldsymbol{y})\right)^{-1} \boldsymbol{r}'(\boldsymbol{x}, \boldsymbol{y}) \\ &= \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{x}, \boldsymbol{y})\right)^{-1} \left(\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{y}) + c \cdot \mathbf{1}\right) \\ &= V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y}) + \frac{c}{1-\gamma}. \end{aligned}$$

Thus, our claim follows immediately from the definition of Nash equilibria (Definition 2.1). □

**Claim D.7.** Let $\boldsymbol{\pi} \in \Pi$ be a joint stationary policy, and $\boldsymbol{d}^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}$ be the induced visitation measure. Then, for every $s \in \mathcal{S}$,

$$\rho(s) \leq d^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}(s) \leq \frac{1}{1-\gamma}.$$

*Proof.* This is an immediate consequence of the definition of $d^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}$; in particular,

$$d^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}(s) = \sum_{\overline{s} \in \mathcal{S}} \rho(\overline{s}) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s^{(t)} = s | \boldsymbol{\pi}, s^{(0)} = \overline{s}) \leq \sum_{\overline{s} \in \mathcal{S}} \rho(\overline{s}) \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma},$$

and

$$d^{\boldsymbol{\pi}}_{\boldsymbol{\rho}}(s) = \sum_{\overline{s} \in \mathcal{S}} \rho(\overline{s}) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s^{(t)} = s | \boldsymbol{\pi}, s^{(0)} = \overline{s}) \geq \rho(s) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s^{(t)} = s | \boldsymbol{\pi}, s^{(0)} = s) \geq \rho(s).$$

□

**Claim D.8.** Suppose that the reward function takes values in $[m_r, M_r]$, for some $m_r, M_r > 0$. Then, for any stationary joint policy $\boldsymbol{\pi} \in \Pi$ and every state $s \in \mathcal{S}$,

$$\frac{m_r}{1-\gamma} \leq V_s(\boldsymbol{\pi}) \leq \frac{M_r}{1-\gamma}.$$

*Proof.* By the definition of the value function in (3), we have

$$V_s(\boldsymbol{\pi}) \leq M_r + \gamma M_r + \gamma^2 M_r + \cdots = \frac{1}{1-\gamma} M_r,$$

for any $s \in \mathcal{S}$. Similarly, we conclude that

$$V_s(\boldsymbol{\pi}) \geq \frac{1}{1-\gamma} m_r.$$

□

**Claim D.9.** Let an adversarial team Markov game $\mathcal{G}$, two team policies $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}$ and quantities $R_b(\cdot,\cdot), P_b(\cdot|s,\cdot), v(s)$ quantities defined in Equation (Q-NLP). The following inequalities hold:

1. $|r(s, \tilde{\boldsymbol{x}}, b) - r(s, \hat{\boldsymbol{x}}, b)| \leq \sqrt{\sum_{k=1}^n A_k} \|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|$, for any $(s, b) \in \mathcal{S} \times \mathcal{B}$;
2. $\left|\sum_{s' \in \mathcal{S}} \left(\mathbb{P}(s'|s, \tilde{\boldsymbol{x}}, b) - \mathbb{P}(s'|s, \hat{\boldsymbol{x}}, b)\right) \tilde{v}(s')\right| \leq \frac{S}{1-\gamma} \sqrt{\sum_{k=1}^n A_k} \|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|$, for any $(s, b) \in \mathcal{S} \times \mathcal{B}$;
3. $|\tilde{v}(s) - \hat{v}(s)| \leq L \|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|$, for any $s \in \mathcal{S}$; and
4. $\left|\sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \hat{\boldsymbol{x}}, b) \left(\tilde{v}(s') - \hat{v}(s')\right)\right| \leq SL \|\tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|$, for any $(s, b) \in \mathcal{S} \times \mathcal{B}$.

*Proof.* We briefly note how the bounds are derived:

- We first establish Item 1. Fix any pair $(s, b) \in \mathcal{S} \times \mathcal{B}$. By definition, we have

$$r(s, \tilde{\boldsymbol{x}}, b) = \mathbb{E}_{\boldsymbol{a} \sim \tilde{\boldsymbol{x}}}[r(s, \boldsymbol{a}, b)] = \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r(s, \boldsymbol{a}, b) \prod_{k=1}^{n} \tilde{x}_{k,s,a_k}.$$

As a result,

$$\begin{aligned}
|r(s, \tilde{\boldsymbol{x}}, b) - r(s, \hat{\boldsymbol{x}}, b)| &= \left| \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r(s, \boldsymbol{a}, b) \prod_{k=1}^{n} \tilde{x}_{k,s,a_k} - \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r(s, \boldsymbol{a}, b) \prod_{k=1}^{n} \hat{x}_{k,s,a_k} \right| \\
&= \left| \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r(s, \boldsymbol{a}, b) \left( \prod_{k=1}^{n} \tilde{x}_{k,s,a_k} - \prod_{k=1}^{n} \hat{x}_{k,s,a_k} \right) \right| \\
&\leq \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} \left| \prod_{k=1}^{n} \tilde{x}_{k,s,a_k} - \prod_{k=1}^{n} \hat{x}_{k,s,a_k} \right| \qquad (51) \\
&\leq \sum_{k=1}^{n} \|\tilde{\boldsymbol{x}}_{k,s} - \hat{\boldsymbol{x}}_{k,s}\|_1 \\
&= \|\tilde{\boldsymbol{x}}_s - \hat{\boldsymbol{x}}_s\|_1 \leq \left( \sqrt{\sum_{k=1}^{n} A_k} \right) \|\tilde{\boldsymbol{x}}_s - \hat{\boldsymbol{x}}_s\|_2, \qquad (52)
\end{aligned}$$

where (51) follows from the triangle inequality and the fact that $|r(s, \boldsymbol{a}, b)| \leq 1$, and (52) follows from the fact that the total variation distance between two product distributions is bounded by the sum of the total variations of each marginal distribution (Hoeffding & Wolfowitz, 1958), as well as the fact that $\|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|_2$ for a vector $\boldsymbol{x} \in \mathbb{R}^d$.

- Item 2 follows analogously to Item 1, using the fact that $\tilde{v}(s') \leq \frac{1}{1-\gamma}$ (by Claim D.8 and Proposition B.1).

- For Item 3, we begin by noting that $\hat{\boldsymbol{v}}$ and $\tilde{\boldsymbol{v}}$ are the unique optimal vectors of Equation (Q-NLP) for $\hat{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$ respectively (recall Proposition B.1). Further, by Proposition B.1, we know that $\boldsymbol{\rho}^\top \hat{\boldsymbol{v}} = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \phi(\hat{\boldsymbol{x}})$ and $\boldsymbol{\rho}^\top \tilde{\boldsymbol{v}} = \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\tilde{\boldsymbol{x}}, \boldsymbol{y}) = \phi(\tilde{\boldsymbol{x}})$, for any $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ of full support. As a result, Item 3 is a consequence of the fact that $\phi(\cdot)$ is $L$-Lipschitz continuous, which in turn follows since $V_{\boldsymbol{\rho}}$ is $L$-Lipschitz continuous (see Lemma C.1 and Lemma A.1).

- Finally, Item 4 follows from Item 3 and the fact that

$$\begin{aligned}
\left| \sum_{s' \in \mathcal{S}} \mathbb{P}(s' | s, \hat{\boldsymbol{x}}, b) \right| &= \left| \sum_{s' \in \mathcal{S}} \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} \mathbb{P}(s' | s, \boldsymbol{a}, b) \prod_{k=1}^{n} \hat{\boldsymbol{x}}_{k,s,a_k} \right| \\
&\leq \sum_{s' \in \mathcal{S}} \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} \prod_{k=1}^{n} \hat{\boldsymbol{x}}_{k,s,a_k} = S,
\end{aligned}$$

for any fixed $(s, b) \in \mathcal{S} \times \mathcal{B}$, where the last bound follows from the triangle inequality and the normalization constraint of the product distribution: $\sum_{(a_1, \ldots, a_n) \in \mathcal{A}} \prod_{k=1}^{n} \hat{\boldsymbol{x}}_{k,s,a_k} = 1$.

□