UNIVERSITY OF CALIFORNIA,
IRVINE

Structure and Computation of Equilibria in Markov Games

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Fivos Kalogiannis

Dissertation Committee:
Assistant Professor Ioannis Panageas, Chair
Assistant Professor Roy Fox
Distinguished Professor Padhraic Smyth

2024

DEDICATION

To Eva

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

No man is an island, and the quality of one's research is significantly influenced by the caliber of their collaborators. I would like to extend my gratitude to my collaborators over the past years: Ioannis Anagnostides, Vaggos Chatziafratis, Ioannis Panageas, Nikolas Patris, Stelios Stavroulakis, Manolis Vlatakis, Jingming Yan, and Rose Zhang. I am especially thankful to Professor Ioannis Panageas for all the opportunities he has offered me and his research guidance and to Manolis Vlatakis for his mentorship during my early steps into research. I also want to express my appreciation to the members of my thesis committee, Roy Fox and Padhraic Smyth, for their support and assistance.

Additionally, I am grateful to the Archimedes Research Center in Athens, Greece, for the Summer Fellowship in 2023 and the exceptional research experience it provided. Part of the research in the current thesis was carried out there. During this time, I had the opportunity to collaborate with Panayotis Mertikopoulos, whose support I am thankful for.

My transition to the United States was made smooth by the support and love of my friends, family, and partner, Eva. Their encouragement helped me adapt quickly and continue my work seamlessly. I am deeply grateful to my parents, Yorgos and Angeliki, for their unwavering love, support, and assistance, and to my siblings, Despina, Iasonas, and Vasilis. I owe a great debt to my partner, Eva, who consistently provides me with a sense of security and whose keen insight and sound judgment I am thankful for.

The friends I made here gave me the sense of home away from home. They provided a lived experience about the significance of friendship and how it transforms across state and cultural borders; yet, it is the case that concern for one another, trust, and honesty seem to be a common denominator among cultures. I am thankful to my friends from Greece—Elli, Jiorgos, Nasos, Nikos, Thomas, and Vangelis—as well as to those I made in the States—Ale, Andrés, Bea, César, Gabriel, Jingming, Luís, Miguel, Nickole (the first one I met), and Santiago—for their invaluable support, care, honesty, and love. I am grateful for the lessons they have taught me, the amount of precious time they have spent with me, and how joyful they could turn my days in Irvine. Finally, I thank them for the homemade arepas, dumplings, tacos, pizzas, and ice creams, all of which are essential for someone who loves to indulge in the first deadly sin, gluttony.

ABSTRACT OF THE DISSERTATION

Structure and Computation of Equilibria in Markov Games

By

Fivos Kalogiannis

Master of Science in Computer Science

University of California, Irvine, 2024

Assistant Professor Ioannis Panageas, Chair

A Nash equilibrium is an important solution concept in most forms of strategic interactions. We are interested in computing Nash equilibria in Markov games. In turn, Markov games are a family of games that apart from instantaneous reward incorporate a dynamically changing environment whose state changes according to the transition dynamics which depend on the decisions of the agents. Of course, this computational problem in its full generality is known to be intractable. Even more, relaxed notions of equilibria were recently (Deng et al. '21, Daskalakis et al. '22, Jin '22) proven comparably intractable to compute.

For this purpose, we examine several structural assumptions on the games themselves in the hope of being able to provide favorable finite-time computational guarantees. We examine reward-potential, zero-sum polymatrix, and adversarial team Markov games.

These assumptions along with certain others placed on the transition functions of the game allow us to obtain favorable results with regards to equilibrium computation. Further, we observe that without assumptions on the transition dynamics of the game, the task of equilibrium computation remains as hard as the general case even in place of very strong assumptions on the reward functions.

Chapter 1

Introduction

Algorithmic game theory is the discipline concerned with the computational aspect of multiagent *games*. In this context, the term *game* signifies any strategic interactions between multiple self-interested agents. Agents are assumed to be rational in the sense that they are always pursuing the maximization of their own utility. A cornerstone concept of game theory is the concept of the *Nash equilibrium* (NE). A Nash equilibrium of the game, describes a profile of collective (possibly randomized) behavior —better said, *strategies*— from which no agent has an incentive to deviate. In a sense, it is a concept of a stable state of the game as no agent will be motivated to change their strategy. It is crucial to note that in an NE the possibly randomized individual strategies do not share a common source of randomness — figuratively, every player would have to roll their own dice in order to decide which action to take.

For the sake of this introductory discussion, let us provide some additional informal definitions. A normal-form game is defined as a collection of a finite number, n, of players (or agents), each of whom is equipped with a finite set of actions, and an individual utility function for every player. The game itself does not change even if it is repeated for more than one

round. On the other hand, a Markov game is played over a (possibly infinite) time horizon H. In each time step, the game can find itself in any state from the set $\mathcal{S}$ which is common for all players; every state changes the reward function r_k of each player $k \in \{1, \ldots, n\}$, and most importantly, the players' joint action define the probability of transitioning from the current state to the next one through the transition rule $\mathbb{P}$. Players in a Markov game strive to maximize their expected sum of rewards over the time horizon of the game. Finally, a two player normal-form (or Markov game) is said to be zero-sum when the utilities (or rewards) of the two players always sum to zero; otherwise, the game is said to be of general-sum.

In both multiagent normal-form and Markov games an NE is guaranteed to exist (Nash, 1951; Fink, 1964). The NE is broadly considered a *solution concept*; nevertheless, a NE is intractable, relaxed notions of equilibria are considered which are also guaranteed to exist and usually enjoy a more favorable complexity of being computing. Intractability of an NE in even a two-player general-sum game was a celebrated result in algorithmic game theory (Daskalakis et al., 2009; Chen and Deng, 2006). It lead economists turn skeptic over the notion of an NE as relevant to real-world markets in the spirit that is captured by the quote: "*If your laptop cannot find it, neither can the market*". To be a bit more precise, the problem of computing an NE in a general-sum is complete for the complexity class PPAD (Papadimitriou, 1994). In turn, the complexity class PPAD lies within FNP, the search problem counterpart of the class NP. Even more, PPAD is the class of computational problems which a solution is guaranteed to exist by virtue of the Brouwer fixed point theorem (or —as recently proven— the Kakutani fixed point theorem (Papadimitriou et al., 2022)).

In (Deng et al., 2021), authors pose the question of whether computing an NE in a Markov game can be harder than a normal-form game. They answer in the negative and demonstrate that computing an NE in a Markov games is PPAD-complete — *i.e.*, it is as hard and no more so than computing it in a normal-form game. This means that the complexity of computing an approximate NE in a Markov game is no harder than computing an approximate

NE in a general-sum normal-form game. As we will discuss, most common assumptions (which come from normal-form games) for the structure of the reward r_k function do not manage to make the complexity of computing equilibria more favorable. Our main message is that the outcome of the game is overwhelmingly defined by decisions regarding the state transitions. In fact, PPAD-completeness of the problem of computing a NE survives most of the assumption on the rewards which would otherwise make solving a normal-form game a decisively more tractable task. Even more so, the reward functions of each state can be assumed independent of actions for each state and player, and still, computing a NE would still be PPAD-hard; *i.e.*, as hard as the general case with no assumptions on the reward functions. Although the matter is more nuanced, we feel the urge to exclaim that, *a Markov game is the game of state transitions.*

1.1 Multi-agent Reinforcement Learning and Markov Games

Markov games (MGs) — or stochastic games — (Shapley, 1953) are a generalization of multi-agent Markov decision processes (MDPs). The joint action of all players affects the transitions of the process and not just the individual instantaneous rewards of each agent. MGs have long stood as the theoretical framework used to formulate and address questions in field of multi-agent reinforcement learning (MARL) (Littman, 1994). A computational issue which has been encountered by MARL literature is the *curse of multiagents*. Effectively, the curse of multiagents signifies an algorithmic complexity of achieving a given objective (*e.g.* computing an equilibrium) that depends exponentially on the number of agents and/or each agent's actions.

1.2 Searching for Structure

In general, characterizing the complexity of a computational problem results from the proof of the existence of a hard worst-case instance. Nevertheless, it is conventional wisdom that the worst-case instance might never be encountered in practice. Indeed, frequently used approaches of tackling a problem usually exhibit favorable empirical performance. To the end of rigorously arguing about performance guarantees of conventional methods beyond the worst-case one can either perform *smoothed analysis* (Spielman and Teng, 2004) of the problem, or investigate common ways in which the problem instances are *structured*. *I.e.,* one might want to recognize certain characteristics of classes of the problem instances that make the computational complexity of the given task provably more favorable than the worst-case.

In light of the robustness of the hardness of computing a Nash equilibrium in the smoothed analysis setting (Chen and Deng, 2006; Boodaghians et al., 2020), we aim our efforts in examining classes of Markov games.

1.3 Our Contribution

We experimented with a multitude of structural assumptions on reward functions, transitions functions, and both. Namely, we showed some very favorable results for adversarial team Markov games (*i.e.,* games where a team of identically interest agents competes a single player) — in this setting, there was no need of making assumptions on the transition function. Then, we investigated games where the reward functions in every state follow *monotone* and *potential* game structures; in those games we concluded that it is necessary to make assumptions on the transition functions. Otherwise, the computational problem of computing a NE is as hard as the general case of computing a NE in a general-sum Markov game.

Chapter 2

Preliminaries

2.1 Normal-Form Games

A normal-form game is the tuple $\Gamma\left(n, \{\mathcal{A}_k\}_{k\in[n]}\{u_k\}_{k\in[n]}\right)$; every player i is endowed with pure strategies $A_k \in \mathcal{A}_k$; their mixed strategies are denoted as $\boldsymbol{x}_i \in \Delta(\mathcal{A}_k)$, and we mark $\boldsymbol{x} \coloneqq (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The utility of player i is denoted as $u_i(\boldsymbol{x})$. Depending on the assumptions on utility functions u_i we retrieve different classes of games.

2.2 Markov Games

Formally, we define a Markov game (MG) with n finite numbers players as the tuple $\Gamma(H, \mathcal{S}, \{\mathcal{A}_k\}_{k\in[n]}, \mathbb{P}, \{r_k\}_{k\in[n]}, \gamma, \boldsymbol{\rho})$, where:

- $H \in \mathbb{N}_+$ denotes the *time horizon*, or the length of each episode,
- $\mathcal{S}$, with cardinality $S \coloneqq |\mathcal{S}|$, represents the state space,

- $\{\mathcal{A}_k\}_{k\in[n]}$ is the collection of each player's action space, with $\mathcal{A} \coloneqq \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denoting the *joint action space*; an element of this set, a joint action, is generally noted as $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathcal{A}$,
- $\mathbb{P} \coloneqq \{\mathbb{P}_h\}_{h\in[H]}$ is the set of all *transition matrices*, with $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$; $\mathbb{P}_h(\cdot|s, \boldsymbol{a})$ indicates the probability of transitioning to each state given that the joint action $\boldsymbol{a}$ is selected at time h in state s — in infinite-horizon games $\mathbb{P}$ does not depend on h and the index is dropped,
- $r_k \coloneqq \{r_{k,h}\}$ is the reward function of player k at time h; $r_{k,h} : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ yields the reward of player k at a given state and joint action — in infinite-horizon games, $r_{k,h}$ is the same for every h and the index is dropped,
- $\gamma > 0$ is the discount factor, which is generally set to 1 when $H < \infty$, and $\gamma < 1$ when $H \to \infty$,
- $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ is the initial state distribution.

Policies and Value Functions. We will define stationary and nonstationary Markov policies. When the horizon H is finite, a stationary policy equilibrium need not necessarily exist even for a single-agent MG, *i.e.*, a Markov decision process; in this case, we seek nonstationary policies. For the case of infinite-horizon games, it is folklore that a stationary Markov policy Nash equilibrium always exists.

We note that a policy is *Markovian* when it depends only on the present state. A *nonstationary* Markov policy $\boldsymbol{\pi}_k$ for player k is defined as $\boldsymbol{\pi}_k \coloneqq \{\boldsymbol{\pi}_{k,h} : \mathcal{S} \to \Delta(\mathcal{A}_k),\ \forall h \in [H]\}$. It is a sequence of mappings of states s to a distribution over actions $\Delta(\mathcal{A}_k)$ for every timestep h. By $\boldsymbol{\pi}_{k,h}(a|s)$ we will denote the probability of player k taking action a in timestep h and state s. A Markov policy is said to be *stationary* if it outputs an identical probability distribution over actions whenever a particular state is visited, regardless of the corresponding

timestep h.

Further, we define a nonstationary Markov *joint policy* $\boldsymbol{\sigma} \coloneqq \{\boldsymbol{\pi}_h,\ \forall h \in [H]\}$ to be a sequence of mappings from states to distributions over joint actions $\Delta(\mathcal{A}) \equiv \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_n)$ for all time steps h in the time horizon. In this case, the players can be said to share a common source of randomness, or that the joint policy is correlated.

By fixing a joint policy $\boldsymbol{\pi}$ we can define the value function of any given state s and timestep h for every player k as the expected cumulative reward they get from that state and timestep h onward,

$$V_{k,h}^{\boldsymbol{\pi}}(s_1) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{\tau=h}^{H} \gamma^{\tau-1} r_{k,\tau}(s_\tau, \boldsymbol{a}_\tau)\big|s_1\right] = \boldsymbol{e}_{s_1}^\top \sum_{\tau=h}^{H}\left(\gamma^{\tau-1}\prod_{\tau'=h}^{\tau}\mathbb{P}_{\tau'}(\boldsymbol{\pi}_{\tau'})\right)\boldsymbol{r}_{k,\tau}(\boldsymbol{\pi}_\tau).$$

Depending on whether the game is of finite or infinite horizon we get the followin displays,

- In finite-horizon games, $\gamma = 1$, the value function reads,

$$V_{k,h}^{\boldsymbol{\pi}}(s_1) = \boldsymbol{e}_{s_1}^\top \sum_{\tau=h}^{H}\left(\prod_{\tau'=h}^{\tau}\mathbb{P}_{\tau'}(\boldsymbol{\pi}_{\tau'})\right)\boldsymbol{r}_{k,\tau}(\boldsymbol{\pi}_\tau),$$

- In infinite-horizon games, the value function of each state is,

$$V_{k}^{\boldsymbol{\pi}}(s_1) = \boldsymbol{e}_{s_1}^\top \left(\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi})\right)^{-1}\boldsymbol{r}(\boldsymbol{\pi}).$$

Where $\mathbb{P}_h(\boldsymbol{\pi}_h), \mathbb{P}(\boldsymbol{\pi})$ and $\boldsymbol{r}_h(\boldsymbol{\pi}_h), \boldsymbol{r}(\boldsymbol{\pi})$ denote the state-to-state transition probability matrix and expected per-state reward vector for a given policy $\boldsymbol{\pi}_h$ or $\boldsymbol{\pi}$ accordingly. Additionally, $\boldsymbol{e}_{s_1}$ is an all-zero vector apart of a value of 1 in its s_1-th position. Also, we denote $V_{k,h}^{\boldsymbol{\pi}}(\boldsymbol{\rho}) = \sum_{s\in\mathcal{S}}\rho(s)V_{k,h}^{\boldsymbol{\pi}}(s)$.

Best-response policies. Given an arbitrary joint policy $\boldsymbol{\sigma}$, we define the *best-response policy* of a player k to be a policy $\boldsymbol{\pi}_k^\dagger \coloneqq \{\boldsymbol{\pi}_{k,h}^\dagger,\ \forall h \in [H]\}$, such that it is a maximizer of $\max_{\boldsymbol{\pi}_k'} V_{k,1}^{\boldsymbol{\pi}_k'\times\boldsymbol{\sigma}_{-k}}(s_1)$. Additionally, we will use the following notation $V_{k,h}^{\dagger,\boldsymbol{\sigma}_{-k}}(s) \coloneqq$

$\max_{\boldsymbol{\pi}'_k} V_{k,h}^{\boldsymbol{\pi}'_k \times \boldsymbol{\sigma}_{-k}}(s)$.

Notions of equilibria — Finite Horizon. Having defined what a best-response is, it is then quite direct to define different notions of equilibria for Markov games.

Definition 2.1 (CCE)**.** *We will say that a joint (potentially correlated) policy* $\boldsymbol{\sigma} \in \Delta(\mathcal{A})^{H \times S}$ *is an* ϵ*-approximate coarse-correlated equilibrium if it holds that, for an* $\epsilon > 0$*,*

$$V_{k,1}^{\dagger, \boldsymbol{\sigma}_{-k}}(s_1) - V_{k,1}^{\boldsymbol{\sigma}}(s_1) \leq \epsilon, \ \forall k \in [n]. \tag{CCE}$$

Further, we will define a Nash equilibrium policy,

Definition 2.2 (NE)**.** *A joint, product policy* $\boldsymbol{\pi} \in \prod_{k \in [n]} \Delta(\mathcal{A}_k)^{H \times S}$ *is an* ϵ*-approximate Nash equilibrium if it holds that, for an* $\epsilon > 0$*,*

$$V_{k,1}^{\dagger, \boldsymbol{\pi}_{-k}}(s_1) - V_{k,1}^{\boldsymbol{\pi}}(s_1) \leq \epsilon, \ \forall k \in [n]. \tag{NE}$$

It is quite evident that an approximate Nash equilibrium is also an approximate coarse-correlated equilibrium while the converse is not generally true. For infinite-horizon games the definitions are analogous and are deferred to the appendix.

Notions of equilibria — Infinite Horizon. Analogous to the finite-horizon MGs, infinite-horizon MGs assert an array of equilibria that are guaranteed to exist. We will define the notions that are relevant, namely approxiamte CCEs and approximate NEs.

Definition 2.3 (CCE—stationary)**.** *For an* $\epsilon \geq 0$*, a joint product policy* $\boldsymbol{\pi} \in \Delta(\mathcal{A})^{\mathcal{S}}$ *is*

- *an ϵ-approximate Markov-perfect coarse correlated equilibrium if,*

$$V_k^{\dagger,\boldsymbol{\pi}_{-k}}(s) - V_k^{\boldsymbol{\pi}}(s) \leq \epsilon, \ \forall k \in [n],$$

- *an ϵ-approximate (Markov) coarse correlated equilibrium if,*

$$V_k^{\dagger,\boldsymbol{\pi}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

Definition 2.4 (NE—stationary)**.** *For an $\epsilon \geq 0$, a joint product policy $\boldsymbol{\pi} \in \prod_{k=1}^{n} \Delta(\mathcal{A}_k)^{\mathcal{S}}$ is*

- *an ϵ-approximate Markov-perfect Nash equilibrium if,*

$$V_k^{\dagger,\boldsymbol{\pi}_{-k}}(s) - V_k^{\boldsymbol{\pi}}(s) \leq \epsilon, \ \forall k \in [n],$$

- *an ϵ-approximate (Markov) Nash equilibrium if,*

$$V_k^{\dagger,\boldsymbol{\pi}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

2.2.1 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 $\overline{s}$. That is, every visit is multiplied by a discount factor γ^t, where t is the time of the visit. We note that the authors of (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 2.5. *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 2.6 (Distribution Mismatch Coefficient)**.** *Let $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ be a full-support distribution over states, and Π be the joint set of policies. We define the distribution mismatch coefficient D as*

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

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

2.2.2 Properties of the Value Function

The value function of each player, for the case of direct parametrization, asserts some quite favorable properties as demonstrated in (Agarwal et al., 2020). First, it is the case that is smooth. Second, it holds that the value function satisfies what is known as a gradient dominance condition, or, a KŁ- condition (Karimi et al., 2016).

Lemma 2.1 (Value Function Lipschitz Continuity)**.** *For any initial distribution $\boldsymbol{\rho}$, the value function $V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho})$ is $\frac{\sqrt{\sum_k |\mathcal{A}_k|}}{(1-\gamma)^2}$-Lipschitz continuous and $\frac{2\left(\sum_k |\mathcal{A}_k|\right)}{(1-\gamma)^3}$-smooth:*

$$|V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}'}(\boldsymbol{\rho})| \leq \frac{\sqrt{\sum_{k=1}^n |\mathcal{A}_k|}}{(1-\gamma)^2} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|; \text{ and}$$
$$\left\|\nabla V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) - \nabla V_k^{\boldsymbol{\pi}'}(\boldsymbol{\rho})\right\| \leq \frac{2\left(\sum_{k=1}^n |\mathcal{A}_k|\right)}{(1-\gamma)^3} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|,$$

for all $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \prod_{k=1}^n \Delta(\mathcal{A}_k)^S$.

The following property effectively tells us that stationarity implies optimality and gives bound on the optimality gap depending on the accuracy of an approximately stationary point.

Lemma 2.2 (Gradient Dominance)**.** *Let Γ be an infinite-horizon Markov game. It is the case that for any joint policy $\boldsymbol{\pi} := (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$ and every player k,*

$$\max_{\boldsymbol{\pi}_k^\star \in \Pi_k} V_k^{\boldsymbol{\pi}_k^\star, \boldsymbol{\pi}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq \frac{1}{1-\gamma} D \max_{\boldsymbol{\pi}_k' \in \Pi_k} (\boldsymbol{\pi}_k' - \boldsymbol{\pi}_k)^\top \nabla_{\boldsymbol{\pi}_k} V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}).$$

Thanks to the latter condition, an NE computation problem can be cast as a variational inqeuality problem of the form,

$$(\boldsymbol{\pi}_k' - \boldsymbol{\pi}_k)^\top \nabla_{\boldsymbol{\pi}_k} V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq 0, \quad \forall \boldsymbol{\pi}_k' \in \Pi_k, \forall k \in [n].$$

Chapter 3

Structure and Equilibrium Computation: Normal-Form Games

3.1 Monotone Normal-Form Games

We have to begin by stating that *monotone games* (Rosen, 1965) include games that are not of normal-form, namely, continuous games. Monotone games are an important class of games that include two-player zero-sum normal-form games, convex-concave games, socially concave games (Even-Dar et al., 2009), polymatrix zero-sum games (Bregman and Fokin, 1987), *etc.* Here, we will discuss monotone normal-form games and specifically, two-player zero-sum and polymatrix zero-sum games.

In a monotone game the gradient operator of utilities $F(\boldsymbol{x}) \coloneqq (\nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x}))_{i\in[n]}$ satisfies the following inequality:

$$\langle F(\boldsymbol{x}) - F(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle \leq 0, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}.$$

3.1.1 Two-Player Zero-Sum Games

Zero-sum games are a fundamental family of games in game theory, where one participant's gain is balanced by another's loss. John von Neumann initiated game theory by formalizing zero-sum games and developing the minimax theorem. This theorem states that in a zero-sum game, each player can minimize their maximum possible loss by choosing an optimal strategy. Von Neumann's work, along with economist Oskar Morgenstern, laid the foundation for modern game theory in their book (von Neumann and Morgenstern, 2007).

An array of favorable properties make NE computation easy to compute as well as to learn as well. Arguably, the most crucial property is the fact that the duality gap is equal to zero.

Theorem 3.1 (Sion's Minimax Theorem)**.** *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets and a real-valued function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for which:*

- *$f(\cdot, \boldsymbol{y})$ is lower-semicontinuous and quasi-convex for every fixed $\boldsymbol{y}$*
- *$f(\boldsymbol{x}, \cdot)$ is upper-semicontinuous and quasi-concave for every fixed $\boldsymbol{x}$.*

Then:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) = \sup_{\boldsymbol{y}\in\mathcal{Y}} \inf_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y}).$$

We can also deduce the following corollary for convex-concave functions:

Corollary 3.1. *Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets and function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function that is convex-concave, i.e.:*

- *$f(\cdot, \boldsymbol{y})$ is convex for every fixed $\boldsymbol{y}$*
- *$f(\boldsymbol{x}, \cdot)$ is concave for every fixed $\boldsymbol{x}$.*

Then:

$$\min_{\mathcal{X}} \max_{\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) = \max_{\mathcal{Y}} \min_{\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y}).$$

This corollary which is also known as the Von Neumann Minimax Theorem effectively tells as that in a zero-sum game it does not matter which one of the two players commits first to their strategy and which one has the opportunity to adapt.

Computing and Learning NE in Two-Player Zero-Sum Games. As we discussed, a NE equilibrium can efficiently be computed in two-player zero-sum games. In fact, computing a NE is equivalent to solving a linear program (Adler et al., 2009).

Further, researchers have showed interest in learning in two-player zero-sum games and showed promising results of learning processes that are independently followed by the players (Robinson, 1951). Research in this direction has culminated to learning processes with finite-time convergence guarantees of independent learning processes that achieve optimal convergence rates (Syrgkanis et al., 2015).

3.1.2 Zero-Sum Polymatrix Games

Another class of monotone normal-form games is known as zero-sum polymatrix games or zero-sum network separable games. For this class of games, Daskalakis and Papadimitriou (2009) observe that the time-averages of the strategies of no-regret dynamics converge to a NE. A similar property had been demonstrated in (Even-Dar et al., 2009) for socially concave games which are monotone games as well. Following works (Cai and Daskalakis, 2011; Cai et al., 2016) demonstrated that it is the case that all coarse-correlated equilibria *collapse* to the set of Nash equilibria.

By collapse we mean that, given a CCE $\boldsymbol{\sigma} \in \Delta\left(\prod_{k=1}^{n} \mathcal{A}_k\right)$, then the joint mixed strategy $\boldsymbol{x}^{\sigma}$ is a NE where $x_k(a_k) = \sum_{\boldsymbol{a}_{-k} \in \mathcal{A}_{-k}} \sigma(a_k, \boldsymbol{a}_{-k})$.

Computing and Learning NE in Zero-Sum Polymatrix Games. Further, the more general results on monotone games readily apply to these games that span the full spectrum of centralized computation approaches to learning approaches with bandit feedback (Nemirovski, 2004; Bravo et al., 2018; Cai et al., 2022).

3.2 Potential Games

A potential game (Monderer and Shapley, 1996; Rosenthal, 1973) is a game that asserts a function $\psi : \prod_{k=1}^{n} \mathcal{A}_k \to \mathbb{R}$, such that $\forall \boldsymbol{x} \in \prod_{k=1}^{n} \Delta(\mathcal{A}_k), \forall k \in [n], \forall \boldsymbol{x}'_k \in \Delta(\mathcal{A}_k)$

$$\psi(\boldsymbol{x}'_k, \boldsymbol{x}_{-k}) - \psi(\boldsymbol{x}) = u_k(\boldsymbol{x}'_k, \boldsymbol{x}_{-k}) - u_k(\boldsymbol{x}).$$

Computing and Learning NE in Potential Games. For this class of games, numerous algorithms guarantee convergence to a Nash equilibrium, namely, best-response dynamics (Monderer and Shapley, 1996), no-regret dynamics (Anagnostides et al., 2022), *etc.*

3.3 Adversarial Team Games

An *adversarial team game*, represented in normal form, is defined by a tuple $\Gamma(\mathcal{N}, \mathcal{M}, \mathcal{A}, \mathcal{B}, U)$. Γ consists of a finite set of $n := |\mathcal{N}|$ *players* belonging to the same team A, and a single *adversarial player*, B. Each player from team A has a finite and nonempty set of available *actions* $\mathcal{A}_k$, so that $\mathcal{A} \coloneqq \prod_{k=1}^{n} \mathcal{A}_k$ denotes the ensemble of all possible action profiles of

team A. The adversary, B, has a finite and nonempty set of actions $\mathcal{B}$. We will denote by $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathcal{A}$ the action profile of team A, and $b \in \mathcal{B}$ the action of the player in team B. Each team's *payoff* function is denoted by $U_A, U_B : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, so that the *individual utility* of a player is identical to their teammates: $U_k(\boldsymbol{a}, b) = U_A(\boldsymbol{a}, b)$ for all joint action profiles $(\boldsymbol{a}, b) \in \mathcal{A} \times \mathcal{B}$ and for all players $i \in \mathcal{N}$. The utility of the adversary player is $U_B(\boldsymbol{a}, b) = U(\boldsymbol{a}, b)$. Further, the game is assumed to be zero-sum, in the sense that $U_B(\boldsymbol{a}, b) = -U_A(\boldsymbol{a}, b) = U(\boldsymbol{a}, b)$. As a result, the adversary player, B, aims to maximize U—thereby referred to as the maximizer, while players in team A aim to minimize U (hereinafter, minimizers).

Nash Equilibrium. In these games, a NE gets the particular form of,

$$U(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \leq U(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_{-i}, \hat{\boldsymbol{y}}) + \epsilon \text{ and } U(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \geq U(\hat{\boldsymbol{x}}, \boldsymbol{y}) - \epsilon. \tag{NE}$$

Computing and Learning NE in Adversarial Team Games. In (Anagnostides et al., 2023), among some complexity results, it was shown that Algorithm 1 can compute an approximate NE in time $\mathsf{poly}(1/\epsilon, \Gamma)$. Also, the arguments we use in Appendix D.4 for adversarial team Markov games can directly be used to prove the extendibility result found in (Anagnostides et al., 2023). We further note that the two time-scale gradient descent/ascent approach of (Lin et al., 2020) can be used along adding a smoothing strongly-concave function $h(\boldsymbol{y})$ for the adversary guaranteeing convergence to a NE.

Chapter 4

Structure and Equilibrium Computation: Markov Games

4.1 Structured transitions

As we intend to demonstrate, the transition function can essentially be used to simulate any general-sum normal form game even when the reward function form a potential game. This goes to show that computing approximate stationary equilibria is not only hard in infinite-horizon games; transition functions in their full generality can make even finite-horizon nonstationary equilibria intractable. As such, we will present several assumptions that are standard in the literature of MGs and we shall see that under those, approximating equilibria is a tractable problem. We will highlight the structural assumptions of (i) *a single controller*, (ii) *switching-control*, and (iii) *additive transitions*. Each of these assumptions is strictly contained to the one that follows it.

single controller $\subseteq$ switching control $\subseteq$ additive transitions.

Single controller. The single controller assumption in words translates to the fact that only one player out of the many of a MG can affect the transitions from one state to another. This assumption is one that has been studied extensively in past as well as contemporary literature (Parthasarathy and Raghavan, 1981; Sayin et al., 2020).

Switching control. A slightly more general assumption on the structure of the transitions is that of switching control (Vrieze et al., 1983; Mohan and Raghavan, 1987; Kalogiannis and Panageas, 2023). When an n-player MG is characterized by switching control, the state-space is divided into disjoint subsets $\{\mathcal{S}_i\}_{i\in[n]}$, with $\mathcal{S} = \cup_{i=1}^{n}\mathcal{S}_i$; in every such set $\mathcal{S}_i$, it is only player i that controls the transitions.

Additive transitions. Finally, the more general transition structure we will present is that of additive transitions. This structure contains all previous assumptions as special cases and has been investigated in an array of works (Raghavan et al., 1985; Flesch et al., 2007; Park et al., 2023). It can be seen as inducing an interpolation between independent (or, *product*) state-space games (Flesch et al., 2008) and standard MGs.

Definition 4.1 (Additive transitions)**.** *A Markov game is said to exhibit additive transitions when in every state s and timestep h of the horizon, it holds that,*

$$\mathbb{P}_h(s'|s,\boldsymbol{a}) = \sum_{i\in[n]} \omega_{i,s,h}\, \mathbb{P}_{i,h}(s'|s,a_i),$$

where $\omega_{i,s,h} \geq 0, \forall i \in [n]$ and $\sum_{i\in[n]} \omega_{i,s,h} = 1$.[1]

Remark 4.1. *The Markov games with decomposable state-spaces (Flesch et al., 2008; Sayin, 2023; Zhang et al., 2023; Qin and Etesami, 2023), also known as product action spaces or games with indepdent Markov chains are captured by the single controller assumption.*

[1]When, $\omega_{s,h,j} = 1$ and $\omega_{s,h,i} = 0, \forall k \neq i$ we retrieve the switching-control setting.

An Example: Turn-based MGs. *Turn-based* MGs are a class of structured MGs that has proven useful in advancing the understanding of the computational complexity of equlibria in MGs (Daskalakis et al., 2022; Jin et al., 2022; Deng et al., 2021).

Definition 4.2 (Turn-based Markov game—TBMG)**.** *In an n-player turn-based MG, the state space $\mathcal{S}$ is split into disjoint sets $\{\mathcal{S}_i\}_{i\in[n]}$. In every such set $\mathcal{S}_i$, player i (called the controller) determines entirely through their actions both the transitions and the reward functions of all players.*

One can observe that turn-based MGs are a special case of MGs with switching control. Further, correlated policies are equivalent to product policies in those games, making CCEs and NEs equivalent may they be stationary or nonstationary and perfect or not. We will refer to them as equilibria without further specification.

4.2 Two-Player Zero-Sum Markov Games

Two-player zero-sum Markov games were first defined by (Shapley, 1953) and in a way have initiated the literature of MGs. Also, it was this class of games studied which contemporary research focused on and initiated computational approaches to equilibrium computation in MGs with finite-time guarantees (Daskalakis et al., 2020).

Since we focus on two-player zero-sum Markov games, we simplify the notation by using $V^{\cdot}_{h=1}(s) \coloneqq V^{\cdot}_{2,1}(s)$—*i.e.*, player 1 is the minimizing player and player 2 is the maximizer. We show the following theorem:

Theorem 4.1 (Collapse in two-player zero-sum MG's)**.** *Let a two-player zero-sum Markov game Γ' and an ϵ-approximate CCE policy of that game $\boldsymbol{\sigma}$. Then, the marginalized product policies $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ form a 2ϵ-approximate NE.*

Proof. Since $\boldsymbol{\sigma}$ is an ϵ-approximate CCE joint policy, by definition it holds that for any $\boldsymbol{\pi}_1$ and any $\boldsymbol{\pi}_2$,

$$V_{h=1}^{\boldsymbol{\sigma}_{-2}\times\boldsymbol{\pi}_2}(s_1) - \epsilon \leq V_{h=1}^{\boldsymbol{\sigma}}(s_1) \leq V_{h=1}^{\boldsymbol{\pi}_1\times\boldsymbol{\sigma}_{-1}}(s_1) + \epsilon.$$

Due to Claim D.1, the latter is equivalent to the following inequality,

$$V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2}(s_1) - \epsilon \leq V_{h=1}^{\boldsymbol{\sigma}}(s_1) \leq V_{h=1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2^{\sigma}}(s_1) + \epsilon.$$

Plugging in $\boldsymbol{\pi}_1^{\sigma}, \boldsymbol{\pi}_2^{\sigma}$ alternatingly, we get the inequalities:

$$\begin{cases} V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2}(s_1) - \epsilon \leq V_{h=1}^{\boldsymbol{\sigma}}(s_1) \leq V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2^{\sigma}}(s_1) + \epsilon \\ V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2^{\sigma}}(s_1) - \epsilon \leq V_{h=1}^{\boldsymbol{\sigma}}(s_1) \leq V_{h=1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2^{\sigma}}(s_1) + \epsilon \end{cases}$$

The latter leads us to conclude that for any $\boldsymbol{\pi}_1$ and any $\boldsymbol{\pi}_2$,

$$V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2}(s_1) - 2\epsilon \leq V_{h=1}^{\boldsymbol{\pi}_1^{\sigma}\times\boldsymbol{\pi}_2^{\sigma}}(s_1) \leq V_{h=1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2^{\sigma}}(s_1) + 2\epsilon,$$

which is the definition of a NE in a zero-sum game. □

4.3 Markov Potential Games

An important class of MGs that has gained traction in recent literature is the class of Markov potential games (MPGs) (Leonardos et al., 2021; Zhang et al., 2021; Mguni et al., 2021). The latter references are the ones that have provided finite-time computation of approximate NE; nevertheless, the same setting is present in other works that considered asymptotic convergence guarantees (Fudenberg and Levine, 1988; Macua et al., 2018). In this class

of games, there exists a state-dependent potential function for the *value functions* of the players, rather than just the reward functions. In (Leonardos et al., 2021) it is highlighted that an MPG can be zero-sum in the rewards of one state and potential in the rewards of another. We remark that for an MPG, it is assumed that there exists a potential function for the *value functions* of the game, rather than the rewards. One is encouraged to revise the counterexamples provided in (Leonardos et al., 2021; Zhang et al., 2021) for MGs which fail to be an MPG even though every stage game is a potential game, or MGs with stage games which are zero-sum games, yet they are MPGs.

Definition 4.3 (Markov potential game — MPG)**.** *An MG is a Markov potential game if there exists a state-dependent potential function, $\Phi^{\boldsymbol{\pi}}(s)$, such that for all players $k \in [n]$, joint policies $\boldsymbol{\pi}$, and unilateral deviations $\boldsymbol{\pi}'_k$,*

$$\Phi^{\boldsymbol{\pi}}(s) - \Phi^{\boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}}(s) = V_k^{\boldsymbol{\pi}}(s) - V_i^{\boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}}(s).$$

We include a list of conditions placed upon the reward functions that *do not* suffice to make an MG an MPG. These conditions appeared in (Zhang et al., 2021).

Proposition 4.1 ((Zhang et al., 2021))**.** *None of the following conditions imply that an MG is an MPG,*

1. *There exists a function $\phi : \mathcal{S} \times \mathcal{A}$ in for each state, such that,*

$$r_k(s, \boldsymbol{a}) - r_k(s, a'_k, \boldsymbol{a}_{-k}) = \phi(s, \boldsymbol{a}) - \phi(s, a'_k, \boldsymbol{a}_{-k}), \ \forall s \in \mathcal{S}, \forall \boldsymbol{a}, a'_i.$$

2. *There exists a function $\phi : \mathcal{S} \times \mathcal{A}$ such that,*

$$r_k(s, a'_{-k}, \boldsymbol{a}_{-k}) - r_k(s', a''_k, \boldsymbol{a}_{-k}) = \phi(s, a'_{-k}, \boldsymbol{a}_{-k}) - \phi(s', a''_k, \boldsymbol{a}_{-k}),$$
$$\forall s, s' \in \mathcal{S}, \forall \boldsymbol{a}, a'_i, a''_i.$$

3. Reward functions are independent of state s, such that,

$$r_k(\boldsymbol{a}) - r_k(a'_k, \boldsymbol{a}_{-k}) = \phi(\boldsymbol{a}) - \phi(a'_k, \boldsymbol{a}_{-k}),\ \forall \boldsymbol{a}, a'_k.$$

The referenced papers (Leonardos et al., 2021; Zhang et al., 2021; Mguni et al., 2021) do not offer an answer regarding the tractability of computing equilibria in games that satisfy any of the previous conditions; assumptions of all three items hold true in our construction in Theorem 4.5 — hence, with no assumption on the transition function, computing approximate nonstationary NEs is PPAD-hard.

4.4 Adversarial Team Markov Games

We define an *adversarial team Markov game* (or an adversarial team *stochastic* game) to be the Markov game extension of static, normal-form adversarial team games (Von Stengel and Koller, 1997). We consider the infinite-horizon discounted setting in which a team of identically-interested agents win what the adversary loses. Formally, the game Γ is defined as a tuple $\Gamma(\mathcal{S}, \mathcal{N}, \mathcal{A}, \mathcal{B}, r, \mathbb{P}, \gamma, \rho)$ whose components are:

- $\mathcal{S}$ is a finite and nonempty set of *states*, with cardinality $S \coloneqq |\mathcal{S}|$;
- $\mathcal{N}$ is the set of players, partitioned into a set of n team agents $\mathcal{N}_A \coloneqq [n]$ and a single *adversary*
- $\mathcal{A}_k$ is the action space of each player in the team $k \in [n]$, so that $\mathcal{A} \coloneqq \bigtimes_{k\in[n]} \mathcal{A}_k$, while $\mathcal{B}$ is the action space of the adversary. We also let $A_k \coloneqq |\mathcal{A}_k|$ and $B \coloneqq |\mathcal{B}|$;[2]
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to (0, 1)$ is the (deterministic) instantaneous *reward function*[3] representing

[2]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.

[3]Assuming that the reward is positive is without any loss of generality

the (normalized) payoff of the adversary, so that for any $(s, \boldsymbol{a}, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$r(s, \boldsymbol{a}, b) + \sum_{k=1}^{n} r_k(s, \boldsymbol{a}, b) = 0, \tag{4.1}$$

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

$$r_k(s, \boldsymbol{a}, b) = r_{\text{team}}(s, \boldsymbol{a}, b). \tag{4.2}$$

- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ 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 player (the adversary) have identical interests (see (4.2)), and the game is globally zero-sum—in the sense of (4.1).

Since the game follows an adversarial team structure, we can argue about the equilibria of the game solely by the adversary's value function:

$$V_s(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}}) \coloneqq \mathbb{E}_{(\boldsymbol{\pi}_{\text{team}}, \boldsymbol{\pi}_{\text{adv}})} \left[\sum_{t=0}^{\infty} \gamma^t r(s^{(t)}, \boldsymbol{a}^{(t)}, b^{(t)}) \big| s_0 = s \right]. \tag{4.3}$$

For this class of games, the NE takes the form,

$$\begin{cases} V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{\text{team}}, \boldsymbol{\pi}^\star_{\text{adv}}) \leq V_{\boldsymbol{\rho}}((\boldsymbol{\pi}'_k, \boldsymbol{\pi}^\star_{-k}), \boldsymbol{\pi}^\star_{\text{adv}}) + \varepsilon, \ \forall k \in [n], \forall \boldsymbol{\pi}'_k \in \Pi_k, \\ V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{\text{team}}, \boldsymbol{\pi}^\star_{\text{adv}}) \geq V_{\boldsymbol{\rho}}(\boldsymbol{\pi}^\star_{\text{team}}, \boldsymbol{\pi}'_{\text{adv}}) - \varepsilon, \qquad \forall \boldsymbol{\pi}'_{\text{adv}} \in \Pi_{\text{adv}}. \end{cases} \tag{4.4}$$

4.4.1 Main Result

Theorem 4.2 (Informal)**.** *There is an algorithm (*IPGMAX*) that, for any $\epsilon > 0$, computes an ϵ-approximate stationary Nash equilibrium policy profile in adversarial team Markov games, and runs in time*

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

In this section, we sketch the main pieces required in the proof of our main result, Theorem 4.2. We begin by describing our algorithm in Section 4.4.2. Next, in Section 4.4.3, we characterize the strategy $\hat{\boldsymbol{x}} \in \mathcal{X}$ for the team returned by IPGMAX, while Section 4.4.4 completes the proof by establishing that $\hat{\boldsymbol{x}}$ can be efficiently extended to an approximate Nash equilibrium. The formal proof of Theorem 4.2 is deferred to the Appendix.

4.4.2 Our Algorithm

In this subsection, we describe in more detail IPGMAX, our algorithm for computing ϵ-approximate Nash equilibria 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 $(\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_n^{(0)}) = \boldsymbol{x}^{(0)} \in \mathcal{X} \coloneqq \bigtimes_{k=1}^{n} \mathcal{X}_k$ (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 η, 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 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 4.4.3, 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 4.4.4.

Algorithm 1 Independent Policy GradientMax (IPGMAX)

1: Precision $\epsilon > 0$
2: Initial Strategy $\boldsymbol{x}^{(0)} \in \mathcal{X}$
3: Learning rate $\eta \coloneqq \frac{\epsilon^2(1-\gamma)^9}{32S^4D^2\left(\sum_{k=1}^n A_k + B\right)^3}$
4: Number of iterations $T \coloneqq \frac{512S^8D^4\left(\sum_{k=1}^n A_k + B\right)^4}{\epsilon^4(1-\gamma)^{12}}$
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: $\hat{\boldsymbol{y}} \leftarrow$ `AdvNashPolicy`$(\hat{\boldsymbol{x}})$ ▷ defined in Algorithm 2
11: **return** $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$

Algorithm 2 Algorithm for computing `AdvNashPolicy`

Input: An ϵ-nearly stationary point $\hat{\mathbf{x}} \in \mathcal{X}$ of $\phi(\mathbf{x}) \equiv \max_{\mathbf{y} \in \mathcal{Y}} V_\rho(\mathbf{x}, \mathbf{y})$
1: Let $\hat{\mathbf{v}}$ be the best-response value vector for the adversary
2: Compute the coefficients of the linear program $\mathrm{LP}_{\mathrm{adv}}$
3: Let λ be any feasible solution of $\mathrm{LP}_{\mathrm{adv}}$
4: Let $\hat{y}_{s,b} = \frac{\lambda(s,b)}{\sum_{b' \in \mathcal{B}} \lambda(s,b')}$ for all $s \in \mathcal{S}, b \in \mathcal{B}$
5: **return** $\hat{\mathbf{y}}$

4.4.3 Analyzing Independent Policy GradientMax

In this subsection, we establish that IPGMAX finds an ϵ-nearly stationary point $\hat{\boldsymbol{x}}$ of $\phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y})$ 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 4.2.

First, we note the by-now standard property that the value function $V_{\boldsymbol{\rho}}$ is L-Lipschitz continuous and ℓ-smooth, where $L \coloneqq \frac{\sqrt{\sum_{k=1}^n A_k+B}}{(1-\gamma)^2}$ and $\ell \coloneqq \frac{2\left(\sum_{k=1}^n A_k+B\right)}{(1-\gamma)^3}$ (Lemma 2.1). 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 ϕ, defined as follows.

Definition 4.4 (Moreau Envelope)**.** *Let $\phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y}\in\mathcal{Y}} V_{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{y})$. For any $0 < \lambda < \frac{1}{\ell}$ the Moreau envelope ϕ_λ of ϕ is defined as*

$$\phi_\lambda(\boldsymbol{x}) \coloneqq \min_{\boldsymbol{x}'\in\mathcal{X}} \left\{ \phi(\boldsymbol{x}') + \frac{1}{2\lambda}\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|^2 \right\}. \tag{4.5}$$

We will let $\lambda \coloneqq \frac{1}{2\ell}$.

Crucially, the Moreau envelope ϕ_λ, as introduced in (4.5), is ℓ-strongly convex; this follows immediately from the fact that $\phi(\boldsymbol{x})$ is *ℓ-weakly convex*, in the sense that $\phi(\boldsymbol{x}) + \frac{\ell}{2}\|\boldsymbol{x}\|^2$ is convex (see Lemma B.1). A related notion that will be useful to measure the progress of IPGMAX is the *proximal mapping* of a function f, defined as $\operatorname{prox}_f : \mathcal{X} \ni \boldsymbol{x} \mapsto \arg\min_{x'\in\mathcal{X}} \left\{ f(\boldsymbol{x}') + \frac{1}{2}\|\boldsymbol{x}'-\boldsymbol{x}\|^2 \right\}$; the proximal of $\phi/(2\ell)$ is well-defined since ϕ is ℓ-weakly convex (Proposition B.1). We are now ready to state the convergence guarantee of IPGMAX.

Proposition 4.2. *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 $\tilde{\boldsymbol{x}}^{(t^\star)} \coloneqq \operatorname{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t^\star)})$.*

The proof relies on the techniques of (Lin et al., 2020), and it is deferred to Appendix D.5. The main takeaway is that $O(1/\epsilon^4)$ iterations suffice in order to reach an ϵ-nearly stationary point of ϕ—in the sense that it is ϵ-far in ℓ_2 distance from its proximal point. A delicate issue here is that Proposition 4.2 only gives a best-iterate guarantee, and identifying that iterate might introduce a substantial computational overhead. To address this, we also show in Corollary D.3 that by randomly selecting $\lceil \log(1/\delta) \rceil$ iterates over the T repetitions of IPGmax, we are guaranteed to recover an ϵ-nearly stationary point with probability at least $1 - \delta$, for any $\delta > 0$.

4.4.4 Efficient Extension to Nash Equilibria

In this subsection, we establish that any ϵ-nearly stationary point $\hat{\boldsymbol{x}}$ of ϕ, 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 and Koller (1997), and it is the crux in the analysis towards establishing our main result, Theorem 4.2.

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

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

(Q-NLP)

Here, we have overloaded notation so that $r(s,\boldsymbol{x},b) \coloneqq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{x}_s}[r(s,\boldsymbol{a},b]$ and $\mathbb{P}(s'|s,\boldsymbol{x},b)) \coloneqq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{x}_s}[\mathbb{P}(s'|s,\boldsymbol{a},b)]$. 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 4.3. *For any fixed $\boldsymbol{x}\in\mathcal{X}$, there is a unique optimal solution $\boldsymbol{v}^\star$ to* ($\mathrm{P_{NE}}$). *Further, if $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$, and $\tilde{\boldsymbol{v}}\in\mathbb{R}^S$ is the corresponding optimal, then $(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{v}})$ is the global optimum of* ($\mathrm{P_{NE}}$).

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 ($\mathrm{P_{NE}}$) is an equivalent formulation of the proximal mapping. These steps are formalized in Appendix D.4.2. 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 ($\mathrm{P_{NE}}$).

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 D.9 we show that any feasible point of ($\mathrm{P_{NE}}$) satisfies that constraint qualification, thereby im-

plying the existence of nonnegative Lagrange multipliers satisfying the KKT conditions for any local optimum (Corollary D.2), and in particular for $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$:

Proposition 4.4. *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 too much error since $\|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| \leq \epsilon$ (Proposition 4.2), and the underlying constraint functions are Lipschitz continuous—with constants that depend favorably on the game $\mathcal{G}$; we formalize that in Lemma D.10. This leads us to our main theorem, summarized below (see Theorem D.7 for a precise statement).

Theorem 4.3. *Let* $\hat{\boldsymbol{x}}$ *be an* ϵ*-nearly stationary point of* ϕ*. 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 4.4. The algorithm

for computing the policy for the adversary is summarized in Algorithm 2 of Appendix D.4. A delicate issue with Theorem 4.3, 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 and 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.

4.5 Reward-Potential Markov Games

We define the class of reward-potential MGs to be the set of MGs whose reward functions in every state are characterized by the existence of a potential function. *I.e.,* when players unilaterally deviate from a given joint policy, changes in the utility of each player are described by the change in the potential function. Formally:

Definition 4.5 (Reward-potential Markov game — RPMG)**.** *We call a Markov game reward-potential when for every state s (and timestep h of the horizon), there exists a function $\phi_h : \mathcal{S} \times \Delta(\mathcal{A}) \to \mathbb{R}$ such that for all players $i \in [n]$, joint policies $\boldsymbol{\pi} \in \Delta(\mathcal{A})$, and unilateral deviations $\boldsymbol{\pi}'_i \in \Delta(\mathcal{A}_i)$,*

$$\phi_h(s, \boldsymbol{\pi}) - \phi_h(s, \boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}) = r_{k,h}(s, \boldsymbol{\pi}) - r_{k,h}(s, \boldsymbol{\pi}'_k, \boldsymbol{\pi}_{-k}).$$

Remark 4.2. *In our opinion, this is a justified and reasonable alternative Markovian extension of the class of potential games. Further, the proposed assumption is rather minimal, a lot more so than the existence of a potential function for the value functions of the players. Moreover, a state based potential game defined in (Marden, 2012) is both an MPG and RMPG. In this class of games, there exists a potential function for the rewards of each state (rendering it an RPMG), while the fact that state transitions are independent of the players' actions satisfy a sufficient condition for it to be an MPG (see Proposition 4.1).*

Theorem 4.4 (PPAD-hardness for perfect equilibria — (Daskalakis et al., 2022, Theorem 3.1))**.** *There exists a constant $\epsilon > 0$ such that the problem of computing an ϵ-approximate perfect NE in 2-player, turn-based stochastic games with $\gamma = 1/2$ is* PPAD*-hard. As such, the problem of computing an ϵ-approximate perfect CCE in 2-player, infinite-horizon stochastic games with $\gamma = 1/2$ is* PPAD*-hard.*

Observation 4.1. *Computing an ϵ-approximate stationary CCE in reward-potential Markov games is* PPAD*-hard.*

Let us make the latter observation clearer. We denote the controller of state $s \in \mathcal{S}_k$, $\text{ctrlr}(s) = k$. From the definition of TBMG, there exist functions r'_j for each player j, such that $r_j(s, \boldsymbol{a}) = r'_j(s, a_{\text{ctrlr}(s)})$. Similarly, there exist $\mathbb{P}'$ such that $\mathbb{P}(s'|s, \boldsymbol{a}) = \mathbb{P}'(s'|s, a_{\text{ctrlr}(s)})$.

Now, we can observe that in a TBMG, the sum of rewards in every state is trivially a potential function for the rewards of that state,

$$\phi(s, \boldsymbol{a}) = \sum_{k \in [n]} r_k(s, \boldsymbol{a}) = \sum_{k \in [n]} r'_k \left(s, a_{\text{ctrlr}(s)}\right).$$

i.e., it holds that,

$$\phi(s, a'_j, \boldsymbol{a}_{-j}) - \phi(s, \boldsymbol{a}) = r_j(s, a'_j, \boldsymbol{a}_{-j}) - r_j(s, \boldsymbol{a}).$$

Hence, TBMGs are in fact a special case of reward-potential Markov games. Next, we show that when transitions assert full generality, even the computation of *nonstationary* approximate NE is PPAD-hard for *finite-horizon* games. Our main complexity contribution is that:

Theorem 4.5. *Computing a nonstationary Markovian ϵ-approximate NE policy in reward-potential Markov games is* PPAD*-hard.*

Proof. Consider a 2-player general-sum game Γ with payoff matrices $(\mathbf{U}, \mathbf{V})$ for player $1, 2$ accordingly. Pure strategies of players 1 and 2 are denoted a_i, b_j, accordingly, with $i \in [m]$ and $j \in [n]$. Hence, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times n}$.

We construct a 2-player reward-potential Markov game Γ' as follows:

- the time horizon of the game is $H = 3$,
- players $1, 2$ have the same set of available actions as players in game Γ;

and $\{a_i\}_{i\in[m]}, \{b_j\}_{j\in[n]}$,

- there is an initial state s_0,
- for every pair of actions a_i, b_j of the initial game there is a state s_{ij};
  i.e., $\mathcal{S} = \{s_{ij},\ ij \in [m] \times [n]\}$
- in state s_{ij} player 1 gets reward U_{ij}, player 2 gets V_{ij}; in s_0, they both get reward 0,
- transitions are deterministic and $\mathbb{P}(s_{ij}|s_0, a_i, b_j) = 1$, while states s_{kj} are absorbing.

In the following figure we offer an illustration of how this simple construction works.

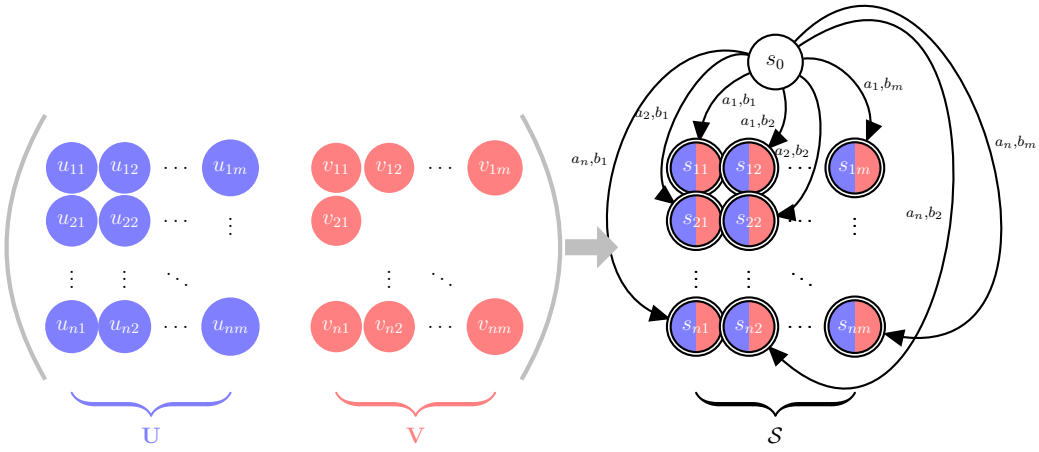

Figure 4.1: Emulating a two-player general-sum normal-form game

The value functions of players $1, 2$ for policies in s_0, where $\boldsymbol{x} := \boldsymbol{\pi}_1(s_0, h = 1)$ and $\boldsymbol{y} :=$

$\boldsymbol{\pi}_2(s_0, h = 1)$, are:

$$\begin{cases} V_1(s_0) & = 0 + \sum_{a,b} \sum_{s_{ij} \in \mathcal{S}} x(a) y(b) \, \mathbb{P}(s_{ij} | s_0, a, b) U_{ij} \\ & = \sum x(a_i) y(b_j) U_{ij} = \boldsymbol{x}^\top \mathbf{U} \boldsymbol{y} \\ V_2(s_0) & = \boldsymbol{x}^\top \mathbf{V} \boldsymbol{y}. \end{cases}$$

Hence, Nash equilibria of game Γ coincide with the $\boldsymbol{x}, \boldsymbol{y}$ policies of Nash equilibria in game Γ' and the complexity of approximating them is known due to (Chen et al., 2009; Daskalakis et al., 2009). □

A Short Remark on Additive Transitions. Before proceeding any further, we would like to make it clear that *additive transitions* is the most general assumption that we can place on the transition function of a tabular MG with finite action-spaces and finite state-spaces. By definition, the transition function is a multilinear function of the individual policies. By our main theorem, Theorem 4.5, we have established that in general, bilinear transition functions can emulate any two-player general-sum normal-form game; in our construction, it is even true that the rewards will be constant in each state and independent of the actions of the players. Additive transitions result in the most general multilinear function that does not lead to intractability of equilibria and consequently the most general assumption on the transitions.

4.5.1 NE Computation in RPMGs with Additive Transitions

Having decisively proven the necessity of assuming a structure on the transitions of the game, we state our main algorithmic result for RPMGs with *additive transitions.*

First, we remark that the NE-Oracle in Algorithm 3 can be implemented in a fully-decentralized

fashion using mirror descent. For completeness, we include the theorem we invoke.

Theorem 4.6 ((Anagnostides et al., 2022)). *Assume a potential game Γ with an L-Lipschitz continuous potential function $\Phi : \prod_{i=1}^{n} \mathcal{A}_i \to \mathbb{R}$. Suppose that each player k employs mirror-descent with a strongly convex and smooth regularizer $\mathcal{R}$. Then, after $O\left(\frac{\Phi_{\max}}{L\epsilon^2}\right)$ iterations, the mirror-descent dynamics converge to an ϵ-approximate Nash equilibrium. Where $\Phi_{\max}$ is the maximum possible value of Φ.*

For the special case of $\mathcal{R}(\cdot) = \frac{1}{2}\|\cdot\|^2$, mirror-descent take the form of projected gradient ascent:

$$\boldsymbol{\pi}_i^{t+1} = \mathrm{Proj}_{\Pi_k}\left\{\boldsymbol{\pi}_k^t + \eta \nabla u_k(\boldsymbol{\pi}^t)\right\}.$$

Where, u_k is player k's utility and Proj_{Π_k} is the projection operator to the set of feasible policies.

As such, we can state our first positive result for RPMGs.

Theorem 4.7 (Informal version of Theorem D.2). *Algorithm 3, with a* NE-Oracle *implemented by every player running mirror descent, computes an ϵ-approximate nonstationary NE for an RPMG with additive transitions in time $O\left(\frac{nH^5|\mathcal{S}|^2 \max_{i\in[n]}|\mathcal{A}_i|^{5/2}}{\epsilon^2}\right)$.*

Algorithm 3 Backwards-Inductive NE Computation in Reward-Potential MGs

input: $n, \mathcal{S}, H$ and accuracy parameter ϵ.
initialization: $\hat{\boldsymbol{V}}_{i,H} = \mathbf{0}$ for all agents $i \in [n]$
for $h = H-1$ to 1 **do**
  `// Approx. NE for subgame` $\Gamma_{s,h}$ `for all` s `with accuracy` ϵ/H
  $\boldsymbol{x}_{s,h} \leftarrow$ NE-Oracle $\left(\frac{\epsilon}{H}, \left\{\boldsymbol{r}_h, \boldsymbol{p}_h, \hat{\boldsymbol{V}}_{h+1}\right\}\right)$ `// for all` $s \in \mathcal{S}$
  `// Update value function`
  $\hat{\boldsymbol{V}}_{i,s,h} \leftarrow \boldsymbol{r}_{i,h}(s, \boldsymbol{x}_h) + \boldsymbol{p}_h(s, \boldsymbol{x}_h)\hat{\boldsymbol{V}}_{i,s,h+1}$
end for
return $\{\boldsymbol{x}_h\}_{h\in[H]}$

Implementing the NE-Oracle. The NE-Oracle takes as input the desired accuracy of an approximate NE and the game. The crucial part regarding Algorithm 3 is the computational cost of implementing the NE-Oracle. In our setting, due to lemma D.1, the oracle can be implemented in a decentralized and distributed manner and its iteration complexity is polynomial in the inverse of the approximation accuracy and the natural parameters of the game. In (Anagnostides et al., 2023) it was proven that the complexity of computing an approximate NE in adversarial potential games matches that of computing a mixed approximate NE in potential games (Rubinstein, 2017).

4.5.2 Properties of RPMGs with Additive Transitions

We conclude this subsection by noting an interesting property of RPMGs. They do inherit the property of asserting pure NEs from their counterpart in nomral and static form. In the case that it was desirable, we could modify the implementation of NE-Oracle in Algorithm 3 in such that could compute pure NE in every state and also retrieve *deterministic* nonstationary NE policies for RPMGs.

Theorem 4.8. *Finite-horizon reward-potential games with additive transitions assert pure Nash equilibria.*

A further note we would like to include is the fact that infinite-horizon RPMGs attain *deterministic* approximate nonstationary equilibria by the standard trick of truncating the horizon of the game. Namely, we set $H = \frac{\log(1/\epsilon)}{1-\gamma}$ and modifying the reward functions such that $r_{i,h}(s,\cdot) = \gamma^{h-1} r_i(s,\cdot)$.

Corollary 4.1. *Infinite-horizon RPMGs with discount parameter γ, attain a deterministic nonstationary approximate NE that can be computed in time* $\mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{1-\gamma}, \sum_{i\in[n+1]} |\mathcal{A}_i|, |\mathcal{S}|\right)$.

4.5.3 An Extension: Adversarial Reward-Potential Markov Games

As an extension, we consider ARPMGs, *i.e.,* MGs whose rewards follow an adversarial potential game structure. It is then straightforward to derive the following corollary of Theorem 4.5,

Corollary 4.2. *Computing a nonstationary Markovian ϵ-approximate NE policy in adversarial reward-potential Markov games is* PPAD*-hard.*

Finally, using the algorithm of (Anagnostides et al., 2023) to implement the NE-Oracle, we see that:

Theorem 4.9. *An ϵ-approximate NE of a finite-horizon ARPMG with additive transitions can be computed in time* $\mathsf{poly}(1/\epsilon, \sum_{i\in[n+1]} |\mathcal{A}_i|, |\mathcal{S}|, H)$.

The proof is defered to the appendix.

4.5.4 Conclusions

We studied Markov games, focusing on the structure of rewards rather than making stronger assumptions about the structure of individual value functions. This setting is often implicitly defined in many modern texts, but its computational aspects have not been thoroughly explored. We addressed the question of the computational complexity of computing equilibria in these games and identified the necessary assumptions for their efficient computation. Additionally, we presented algorithms for this purpose.

4.6 Zero-Sum Polymatrix Markov Games

In this section, we focus on the setting of zero-sum polymatrix switching-control Markov games. This setting encompasses two major assumptions related to the reward functions in every state $\{r_k\}_{k\in[n]}$ and the transition kernel $\mathbb{P}$. The first assumption imposes a zero-sum, polymatrix structure on $\{r_k\}_{k\in[n]}$ for every state and directly generalizes zero-sum polymatrix games for games with multiple states.

Assumption 4.1 (Zero-sum polymatrix games)**.** *The reward functions of every player in any state s are characterized by a* zero-sum*,* polymatrix *structure.*

Polymatrix structure. For every state s there exists an undirected graph $\mathcal{G}_s(\mathcal{V}, \mathcal{E}_s)$ where,

- the set of nodes $\mathcal{V}$ coincides with the set of agents $[n]$; the k-th node is the k-th agent,
- the set of edges $\mathcal{E}_s$ stands for the set of pair-wise interactions; each edge $e = (k, j), k, j \in [n], k \neq j$ stands for a general-sum normal-form game played between players k, j and which we note as $\Big(r_{kj}(s,\cdot,\cdot), r_{jk}(s,\cdot,\cdot)\Big)$ with $r_{kj}, r_{jk} : \mathcal{S}\times\mathcal{A}_k\times\mathcal{A}_j \to [-1, 1]$.

Moreover, we define $\text{adj}(s, k) \coloneqq \{j \in [n] \mid (k, j) \in \mathcal{E}_s\} \subseteq [n]$ to be the set of all neighbors of an arbitrary agent k in state s. The reward of agent k at state s given a joint action $\boldsymbol{a}$ depends solely on interactions with their neighbors,

$$r_{k,h}(s, \boldsymbol{a}) = \sum_{j\in\text{adj}(k)} r_{kj,h}(s, a_k, a_j),\ \forall h \in [H], \forall s \in \mathcal{S}, \forall \boldsymbol{a} \in \mathcal{A}.$$

Further, the *zero-sum* assumption implies that,

$$\sum_k r_{k,h}(s, \boldsymbol{a}) = 0, \quad \forall h \in [H], \forall s \in \mathcal{S}, \forall \boldsymbol{a} \in \mathcal{A}. \tag{4.7}$$

In the infinite-horizon setting, the subscript h can be dropped.

A further assumption (*switching-control*) is necessary in order to ensure the desirable property of equilibrium collapse.

Assumption 4.2 (Switching-control)**.** *In every state $s \in \mathcal{S}$, there exists a single player (not necessarily the same), or controller, whose actions determine the probability of transitioning to a new state.*

Remark 4.3. *It is direct to see that Markov games with a single controller and TBMG, are special case of Markov games with switching controller.*

4.6.1 Main results

In this section we provide the main results of this paper. We shall show the collapsing phenomenon of coarse-correlated equilibria to Nash equilibria in the case of zero-sum, single switching controller polymatrix Markov games. Before we proceed, we provide a formal definition of the notion of collapsing.

Definition 4.6 (CCE collapse to NE)**.** *Let $\boldsymbol{\sigma}$ be any ϵ-CCE policy of a Markov game. Moreover, let the marginal policy $\boldsymbol{\pi}^{\boldsymbol{\sigma}} := (\boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, ..., \boldsymbol{\pi}_n^{\boldsymbol{\sigma}})$ be defined as:*

$$\pi_k^{\boldsymbol{\sigma}}(a|s) = \sum_{\boldsymbol{a}_{-k} \in \mathcal{A}_{-k}} \sigma(a, \boldsymbol{a}_{-k}|s), \ \forall k, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}_k.$$

If $\boldsymbol{\pi}^{\boldsymbol{\sigma}}$ is an $O(\epsilon)$-NE equilibrium for every $\boldsymbol{\sigma}$ then we say the set of approximate CCE's collapses to that of approximate NE's.

4.6.2 Equilibrium collapse in finite-horizon polymatrix Markov games

In this section, we focus on the more challenging case of polymatrix Markov games which is the main focus of this paper. For any finite horizon Markov game, we define ($\mathrm{P_{NE}}$) to be the following nonlinear program with variables $\boldsymbol{\pi}, \boldsymbol{w}$:

$$
\begin{aligned}
&\min \sum_{k\in[n]} \left(w_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) \right) \\
&\text{s.t. } w_{k,h}(s) \geq r_{k,h}(s, a, \boldsymbol{\pi}_{-k,h}) + \mathbb{P}_h(s, a, \boldsymbol{\pi}_{-k,h}) \boldsymbol{w}_{k,h+1}, \\
&\qquad \forall s \in \mathcal{S}, \forall h \in [H], \forall k \in [n], \forall a \in \mathcal{A}_k; \\
&\qquad w_{k,H}(s) = 0, \quad \forall k \in [n], \forall s \in \mathcal{S}; \\
&\qquad \boldsymbol{\pi}_{k,h}(s) \in \Delta(\mathcal{A}_k), \\
&\qquad \forall s \in \mathcal{S}, \forall h \in [H], \forall k \in [n], \forall a \in \mathcal{A}_k.
\end{aligned}
\tag{$\mathrm{P_{NE}}$}
$$

Using the following theorem, we are able to use ($\mathrm{P_{NE}}$) to argue about equilibrium collapse.

Theorem 4.10 (NE and global optima of ($\mathrm{P_{NE}}$))**.** *If $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ yields an ϵ-approximate global minimum of* ($\mathrm{P_{NE}}$)*, then $\boldsymbol{\pi}^\star$ is an $n\epsilon$-approximate NE of the zero-sum polymatrix switching controller MG, Γ. Conversely, if $\boldsymbol{\pi}^\star$ is an ϵ-approximate NE of the MG Γ with corresponding value function vector $\boldsymbol{w}^\star$ such that $w^\star_{k,h}(s) = V^{\boldsymbol{\pi}^\star}_{k,h}(s) \forall (k,h,s) \in [n] \times [H] \times \mathcal{S}$, then $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ attains an ϵ-approximate global minimum of* ($\mathrm{P_{NE}}$)*.*

Following, we are going to use ($\mathrm{P_{NE}}$) in proving the collapse of CCE's to NE's. We observe that the latter program is nonlinear and in general nonconvex. Hence, duality cannot be used in the way it was used in (Cai et al., 2016) to prove equilibrium collapse. Nevertheless, we can prove that given a CCE policy $\boldsymbol{\sigma}$, the marginalized, product policy $\bigtimes_{k\in[n]} \boldsymbol{\pi}^{\boldsymbol{\sigma}}_k$ along with an appropriate vector $\boldsymbol{w}^\sigma$ achieves a global minimum in the nonlinear program ($\mathrm{P_{NE}}$).

More precisely, our main result reads as the following statement.

Theorem 4.11 (CCE collapse to NE in polymatrix MG)**.** *Let a zero-sum polymatrix switching-control Markov game, i.e., a Markov game for which Assumptions 4.1 and 4.2 hold. Further, let an ϵ-approximate CCE of that game $\boldsymbol{\sigma}$. Then, the marginal product policy $\boldsymbol{\pi}^{\sigma}$, with $\boldsymbol{\pi}^{\boldsymbol{\sigma}}_{k,h}(a|s) = \sum_{\boldsymbol{a}_{-k}\in\mathcal{A}_{-k}} \boldsymbol{\sigma}_h(a, \boldsymbol{a}_{-k}),\ \forall k \in [n], \forall h \in [H]$ is an $n\epsilon$-approximate NE.*

Proof. Let an ϵ-approximate CCE policy, $\boldsymbol{\sigma}$, of game Γ. Moreover, let the best-response value-vectors of each agent k to joint policy $\boldsymbol{\sigma}_{-k}$, $\boldsymbol{w}^{\dagger}_k$.

Now, we observe that due to Assumption 4.1,

$$\begin{aligned} w^{\dagger}_{k,h}(s) &\geq r_{k,h}(s, a, \boldsymbol{\sigma}_{-k,h}) + \mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k,h})\boldsymbol{w}^{\dagger}_{k,h+1} \\ &= \sum_{j\in\mathrm{adj}(k)} r_{(k,j),h}(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}}_j) + \mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k,h})\boldsymbol{w}^{\dagger}_{k,h+1}. \end{aligned}$$

Further, due to Assumption 4.2,

$$\mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k,h})\boldsymbol{w}^{\dagger}_{k,h+1} = \mathbb{P}_h(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}}_{\mathrm{argctrlr}(s),h})\boldsymbol{w}^{\dagger}_{k,h+1},$$

or,

$$\mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k,h})\boldsymbol{w}^{\dagger}_{k,h+1} = \mathbb{P}_h(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}})\boldsymbol{w}^{\dagger}_{k,h+1}.$$

Putting these pieces together, we reach the conclusion that $(\boldsymbol{\pi}^{\sigma}, \boldsymbol{w}^{\dagger})$ is feasible for the non-linear program ($\mathrm{P_{NE}}$).

What is left is to prove that it is also an ϵ-approximate global minimum. Indeed, if $\sum_k \boldsymbol{w}^{\dagger}_{k,h}(s_1) \leq \epsilon$ (by assumption of an ϵ-approximate CCE), then the objective function of ($\mathrm{P_{NE}}$) will attain an ϵ-approximate global minimum. In turn, due to Theorem 4.10 the

latter implies that $\boldsymbol{\pi}^{\boldsymbol{\sigma}}$ is an $n\epsilon$-approximate NE. $\square$

We can now conclude that due to the algorithm introduced in (Daskalakis et al., 2022) for CCE computation in general-sum MG's, the next statement holds true.

Corollary 4.3 (Computing a NE—finite-horizon)**.** *Given a finite-horizon switching control zero-sum polymatrix Markov game, we can compute an ϵ-approximate Nash equilibrium policy that is Markovian with probability at least $1-\delta$ in time* $\mathsf{poly}\left(n, H, S, \max_k |\mathcal{A}_k|, \frac{1}{\epsilon}, \log(1/\delta)\right)$.

In the next section, we discuss the necessity of the assumption of switching control using a counter-example of non-collapsing equilibria.

4.6.3 No equilibrium collapse with more than one controllers per-state

Although Assumption 4.1 is sufficient for the collapse of any CCE to a NE in single-state (*i.e.*, normal-form) games, we will prove that Assumption 4.2 is indispensable in guaranteeing such a collapse in zero-sum polymatrix Markov games. That is, if more than one players affect the transition probability from one state to another, a CCE is not guaranteed to collapse to a NE.

Example 4.1. *We consider the following 3-player Markov game that takes place for a time horizon $H = 3$. There exist three states, s_1, s_2, and s_3 and the game starts at state s_1. Player 3 has a single action in every state, while players 1 and 2 have two available actions $\{a_1, a_2\}$ and $\{b_1, b_2\}$ respectively in every state.*

Reward functions. *If player 1 (respectively, player 2) takes action a_1 (resp., b_1), in either of the states s_1 or s_2, they get a reward equal to $\frac{1}{20}$. In state s_3, both players get a reward*

equal to $-\frac{1}{2}$ *regardless of the action they select. Player* 3 *always gets a reward that is equal to the negative sum of the reward of the other two players. This way, the* zero-sum polymatrix property *of the game is ensured (Assumption 4.1).*

Transition probabilities. *If players* 1 *and* 2 *select the joint action* (a_1, b_1) *in state* s_1*, the game will transition to state* s_2*. In any other case, it will transition to state* s_3*. The converse happens if in state* s_2 *they take joint action* (a_1, b_1)*; the game will transition to state* s_3*. For any other joint action, it will transition to state* s_1*. From state* s_3*, the game transitions to state* s_1 *or* s_2 *uniformly at random.*

At this point, it is important to notice that two players control the transition probability from one state to another. In other words, Assumption 4.2 does not hold.

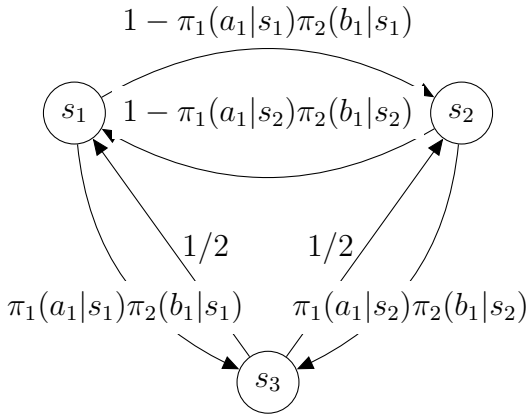

Figure 4.2: A graph of the state space with transition probabilities parametrized with respect to the policy of each player.

Next, we consider the joint policy $\boldsymbol{\sigma}$*,*

$$\boldsymbol{\sigma}(s_1) = \boldsymbol{\sigma}(s_2) = \begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} \end{array}.$$

Claim 4.1. *The joint policy $\boldsymbol{\sigma}$ that assigns probability $\frac{1}{2}$ to the joint actions (a_1, b_2) and (a_2, b_1) in both states s_1, s_2 is a CCE and $V^{\boldsymbol{\sigma}}_{1,1}(s_1) = V^{\boldsymbol{\sigma}}_{2,1}(s_1) = \frac{1}{20}$.*

Yet, the marginalized product policy of $\boldsymbol{\sigma}$ which we note as $\boldsymbol{\pi}^{\boldsymbol{\sigma}}_1 \times \boldsymbol{\pi}^{\boldsymbol{\sigma}}_2$ does not constitute a NE. The components of this policy are,

$$\begin{cases} \boldsymbol{\pi}^{\boldsymbol{\sigma}}_1(s_1) = \boldsymbol{\pi}^{\boldsymbol{\sigma}}_1(s_2) = \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \end{array}, \\ \boldsymbol{\pi}^{\boldsymbol{\sigma}}_2(s_1) = \boldsymbol{\pi}^{\boldsymbol{\sigma}}_2(s_2) = \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \end{array}. \end{cases}$$

I.e., the product policy $\boldsymbol{\pi}^{\boldsymbol{\sigma}}_1 \times \boldsymbol{\pi}^{\boldsymbol{\sigma}}_2$ selects any of the two actions of each player in states s_1, s_2 independently and uniformly at random. With the following claim, it can be concluded that in general when more than one player control the transition the set of equilibria do not collapse.

Claim 4.2. *The product policy $\boldsymbol{\pi}^{\boldsymbol{\sigma}}_1 \times \boldsymbol{\pi}^{\boldsymbol{\sigma}}_2$ is not a NE.*

In conclusion, Assumption 4.1 does not suffice to ensure equilibrium collapse.

Theorem 4.12. *There exists a zero-sum polymatrix Markov game (Assumption 4.2 is not satisfied) that has a CCE which does not collapse to a NE.*

4.6.4 Equilibrium collapse in infinite-horizon polymatrix Markov games

In proving equilibrium collapse for infinite-horizon polymatrix Markov games, we use similar arguments and the following nonlinear program with variables $\boldsymbol{\pi}, \boldsymbol{w}$,

$$
\begin{aligned}
(\mathrm{P}'_{\mathrm{NE}}) \qquad & \min \sum_{k\in[n]} \boldsymbol{\rho}^\top \left(\boldsymbol{w}_k - (\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}))^{-1} \boldsymbol{r}_k(\boldsymbol{\pi})\right) \\
& \text{s.t. } w_k(s) \geq r_k(s, a, \boldsymbol{\pi}_{-k}) + \gamma\, \mathbb{P}(s, a, \boldsymbol{\pi}_{-k}) \boldsymbol{w}_k, \\
& \qquad \forall s \in \mathcal{S}, \forall k \in [n], \forall a \in \mathcal{A}_k; \\
& \boldsymbol{\pi}_k(s) \in \Delta(\mathcal{A}_k), \\
& \qquad \forall s \in \mathcal{S}, \forall k \in [n], \forall a \in \mathcal{A}_k.
\end{aligned}
$$

We note that Example 4.1 can be properly adjusted to show that the switching-control assumption is necessary for equilibrium collapse in infinite-horizon games as well. Compared to finite-horizon games, infinite-horizon games cannot be possibly solved using backward induction. They pose a genuine computational challenge and, in that sense, the importance of the property of equilibrium collapse gets highlighted.

Computational implications. Equilibrium collapse in infinite-horizon MG's allows us to use the CCE computation technique found in (Daskalakis et al., 2022) in order to compute an ϵ-approximate NE. Namely, given an accuracy threshold ϵ, we truncate the infinite-horizon game to its *effective horizon* $H \coloneqq \frac{\log(1/\epsilon)}{1-\gamma}$. Then, we define reward functions that depend on the time-step h, *i.e.*, $r_{k,h} = \gamma^{h-1} r_k$. Finally,

Corollary 4.4. *(Computing a NE—infinite-horizon) Given an infinite-horizon switching control zero-sum polymatrix game Γ, it is possible to compute a Nash equilibrium policy that is Markovian and nonstationary with probability at least $1-\delta$ in time*

$$
\mathsf{poly}\left(n, \frac{1}{1-\gamma}, S, \max_k |\mathcal{A}_k|, \frac{1}{\epsilon}, \log(1/\delta)\right).
$$

4.6.5 Hardness without assumptions on transitions

Theorem 4.13. *Finite-horizon zero-sum polymatrix Markov games with more than one controller are* PPAD*-hard.*

Proof. TL;DR:*For any 2-player general-sum game Γ we can construct (in polynomial time) a 3-player zero-sum polymatrix Markov game Γ' with two controllers such that a NE in Γ' can be used to retrieve in polynomial time NE in Γ.*

Consider a 2-player general-sum game Γ with payoff matrices $(\mathbf{U}, \mathbf{V})$ for player $1, 2$ accordingly. Pure strategies of players 1 and 2 are denoted a_i, b_j, accordingly, with $i \in [m]$ and $j \in [n]$. Hence, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times n}$.

We construct a 3-player polymatrix zero-sum Markov game Γ' as follows:

- the time horizon of the game is $H = 3$,
- players $1, 2$ have the same set of available actions as players in game Γ; $\{a_i\}_{i \in [m]}, \{b_j\}_{j \in [n]}$; tha action-set of player 3 is a singleton (dummy player),
- there is an initial state s_0,
- for every pair of actions a_i, b_j of the initial game there is a state s_{ij}; *i.e.,* $\mathcal{S} = \{s_{ij},\ ij \in [m] \times [n]\}$
- in state s_{ij} player 1 gets reward U_{ij}, player 2 gets V_{ij} and player three gets $-(U_{ij} + V_{ij})$; in s_0, they all get reward 0,
- transitions are deterministic and $\mathbb{P}(s_{ij} | s_0, a_i, b_j) = 1$, while states s_{ij} are absorbing.

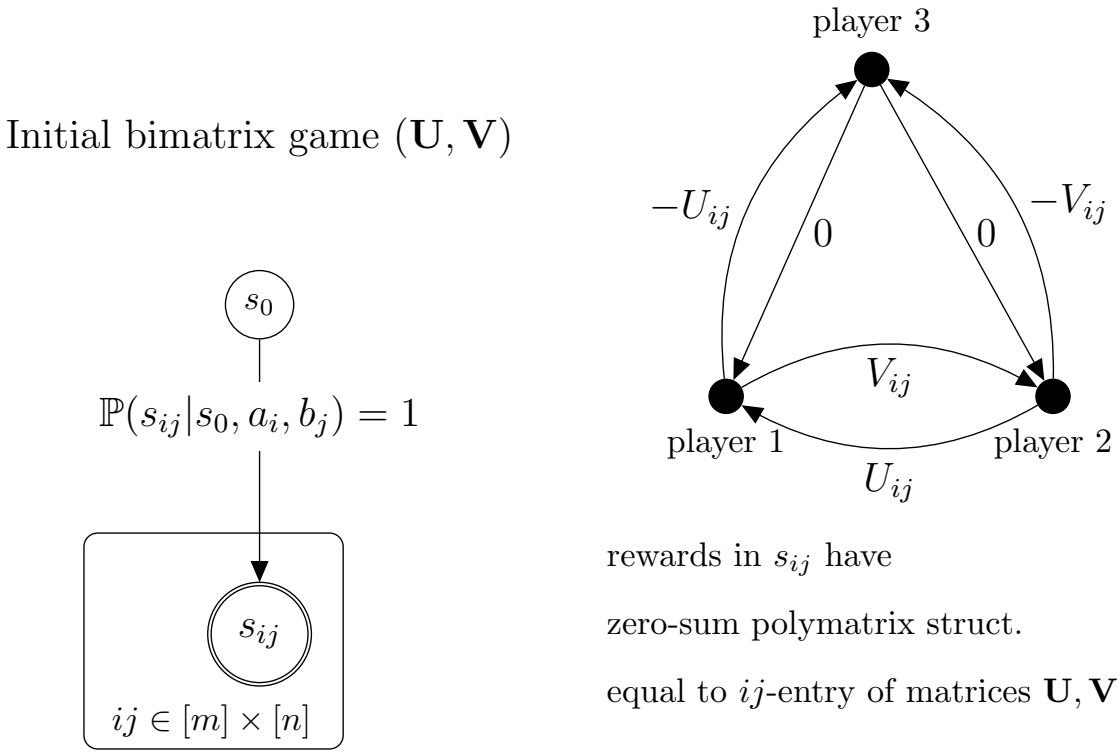

Figure 4.3: PPAD-hardness of **nonstationary** NE proof constuction.

The value functions of players $1, 2$ for policies in s_0 $\boldsymbol{x} := \boldsymbol{\pi}_1(s_0, h = 1), \boldsymbol{y} =: \boldsymbol{\pi}_2(s_0, h = 1)$ are:

$$\begin{cases} V_1(s_0) & = 0 + \sum_{a,b} \sum_{s_{ij} \in \mathcal{S}} x(a) y(b) \, \mathbb{P}(s_{ij}|s_0, a, b) U_{ij} \\ & = \sum x(a_j) y(b_j) U_{ij} = \boldsymbol{x}^\top U \boldsymbol{y} \\ V_2(s_0) & = \boldsymbol{x}^\top V \boldsymbol{y}. \end{cases}$$

Hence, Nash equilibria of game Γ coincide with the $\boldsymbol{x}, \boldsymbol{y}$ policies of Nash equilibria in game Γ'.

□

4.6.6 Conclusion

In this section, we unified switching-control Markov games and zero-sum polymatrix normal-form games. We highlighted how numerous applications can be modeled using this framework and we focused on the phenomenon of equilibrium collapse from the set of coarse-correlated equilibria to that of Nash equilibria. This property holds implications for computing approximate Nash equilibria in switching control zero-sum polymatrix Markov games; it ensures that it can be done efficiently.

Chapter 5

It's all about Transitions

In this chapter we will briefly demonstrate how any two-player general-sum Markov game can be transformed into a strategically-equivalent Markov game with a polynomially larger state space where the rewards of the two players are constant. Doing so, we highlight our main message that, in general, a Markov game is a game of state transitions.

5.1 A Simple Insightful Construction

Observation 5.1. *The computation of a Markov-perfect equilibrium in a Markov game* Γ*, with* $\Gamma(H, \mathcal{S}, \{\mathcal{A}, \mathcal{B}\}, \mathbb{P}, \{r_1, r_2\}, \gamma, \boldsymbol{\rho})$*, can be reduced to the problem of computing an equilibrium in a constant-reward Markov game* $\Gamma'(H', \mathcal{S}', \{\mathcal{A}, \mathcal{B}\}, \mathbb{P}', \{r_1', r_2'\}, \gamma', \boldsymbol{\rho}')$ *whose size is polynomial in the parameters of* Γ*. I.e., a Markov game* Γ' *where*

$$r_k(s, a, b) = r_k(s, a', b'),\ \forall a, a' \in \mathcal{A}, b, b' \in \mathcal{B}, \forall k \in \{1, 2\}, \forall s \in \mathcal{S}'.$$

We use the construction used in proving Theorem 4.5 for the reward functions of every state

of the game Γ. In short, using the previous construct, we add a small number of states to emulate the reward function of each state of the original game.

In particular, in place of every state $s^\kappa \in \mathcal{S}$ of the initial game we put a new set of states $\mathcal{S}^\kappa$ with size $|\mathcal{S}^\kappa| = 1 + mn$. (Reminder: $|\mathcal{A}| = m, |\mathcal{B}| = n$).

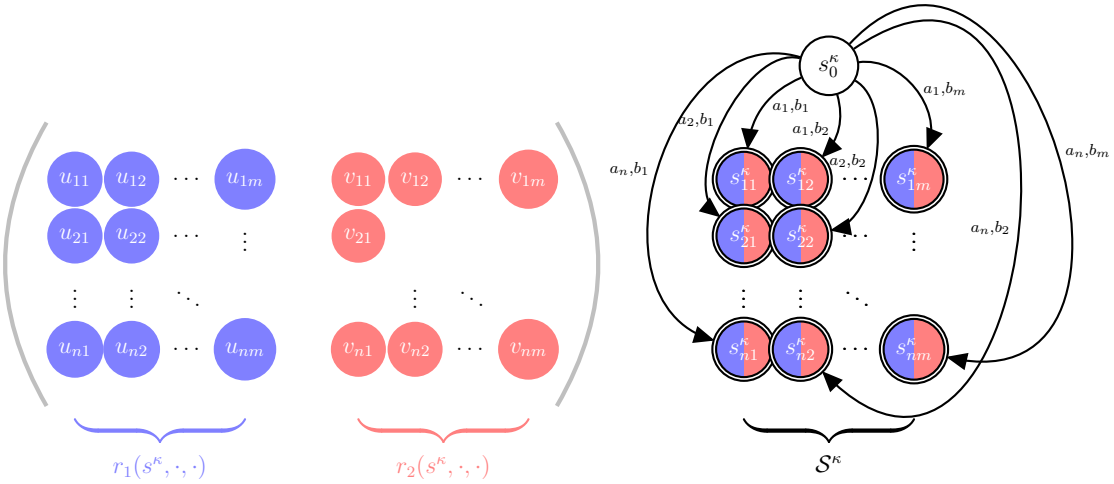

Figure 5.1: Emulating the reward function

In detail, we let,

- $\mathcal{S}' := \{s_0^\kappa\} \cup \{s_{ij}^\kappa\}_{i,j \in [m] \times [n]}$, with $\mathcal{S}'_0 := \{s_0^\kappa\}$;
- $r_1(s_0^\kappa, \cdot, \cdot) = r_2(s_0^\kappa, \cdot, \cdot) = 0$;
- $r'_1(s_{ij}^\kappa, \cdot, \cdot) := \frac{1}{\sqrt{\gamma}} r_1(s^\kappa, a_i, b_j)$ and $r'_2(s_{ij}^\kappa, \cdot, \cdot) := \frac{1}{\sqrt{\gamma}} r_2(s^\kappa, a_i, b_j)$;
- $\mathbb{P}'(s^{\kappa'_0} | s_{ij}^\kappa, \cdot, \cdot) := \mathbb{P}(s^{\kappa'} | s^\kappa, a_i, a_j)$;
- for $\boldsymbol{\rho}'$ it suffices that $\sum_{s_{ij}^{\kappa'} \in \mathcal{S}'} \mathbb{P}(s_0^\kappa | s_{ij}^{\kappa'}) \rho'(s_{ij}^{\kappa'}) = \frac{1}{2} \rho(s^\kappa)$, and $\rho'(s_0^\kappa) = \frac{1}{2} \rho(s^\kappa)$;

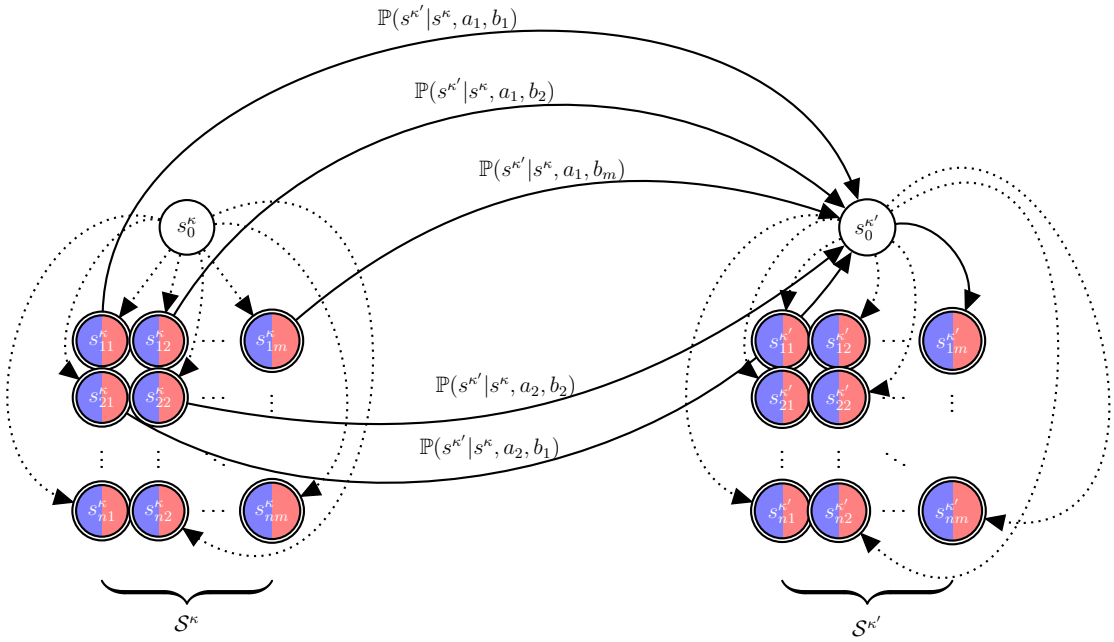

Figure 5.2: Emulating the transition function

- $\gamma' = \sqrt{\gamma}$;
- $H' = 2H$.

We do not define $\boldsymbol{\rho}' \in \Delta(\mathcal{S}')$ in the seemingly more intuitive way of $\rho'(s_0^\kappa) = \rho(s^\kappa),\ \forall s^\kappa \in \mathcal{S}, \forall s_0^\kappa \in \mathcal{S}'$. We want to circumvent having to deal with the technicality of having an initial state distribution that is not of full support. For this reason an additive dummy term will appear in the value of each player which does not affect the equilibria.

Claim 5.1. *The value functions of the states $\mathcal{S}$ of game Γ and that of the subset $\mathcal{S}_0$ of the game Γ' are equal when $\boldsymbol{\pi}(s_0^\kappa) = \boldsymbol{\pi}(s^\kappa)$.*

Proof.

$$
\begin{aligned}
\tilde{V}_1^{\boldsymbol{\pi}}(s_0^\kappa) &= 0 + \gamma' \sum_{s_{ij}^\kappa, \forall(i,j)} \pi_1(a_i|s_0^\kappa)\pi_2(b_j|s_0^\kappa)\tilde{V}^{\boldsymbol{\pi}}(s_{ij}^{\kappa'}) \\
&= \gamma' \sum_{s_{ij}^\kappa, \forall(i,j)} \pi_1(a_i|s_0^\kappa)\pi_2(b_j|s_0^\kappa) \left(r_1'(s_{ij}^{\kappa'}) + \gamma' \sum_{s_0^{\kappa'} \in \mathcal{S}_0} \mathbb{P}'(s_0^{\kappa'}|s_{ij}^\kappa)\tilde{V}_1^{\boldsymbol{\pi}}(s_0^{\kappa'}) \right) \\
&= \sqrt{\gamma} \sum_{s_{ij}^\kappa, \forall(i,j)} \pi_1(a_i|s^\kappa)\pi_2(b_j|s^\kappa) \left(\frac{1}{\sqrt{\gamma}} r_1(s^\kappa, a_i, b_j) + \sqrt{\gamma} \sum_{s_0^{\kappa'} \in \mathcal{S}_0} \mathbb{P}'(s_0^{\kappa'}|s^\kappa, a_i, b_j)\tilde{V}_1^{\boldsymbol{\pi}}(s_0^{\kappa'}) \right) \\
&= r_1(s^\kappa, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) + \gamma \sum_{s_0^{\kappa'} \in \mathcal{S}_0} \mathbb{P}'(s_0^{\kappa'}|s^\kappa, a_i, b_j)\tilde{V}_1^{\boldsymbol{\pi}}(s_0^{\kappa'})
\end{aligned}
$$

Hence, the entries of $\mathcal{S}_0$ of the value vector $\tilde{\boldsymbol{V}}_1^{\boldsymbol{\pi}}$ satisfy the Bellman equations of the original game Γ. □

It is rather direct to observe that for a Markov perfect ϵ-approximate Nash equilibrium of

the game Γ', $\hat{\boldsymbol{\pi}}$, *i.e.,*

$$\tilde{V_1}^{\dagger,\hat{\boldsymbol{\pi}}_2}(s) - \tilde{V_1}^{\hat{\boldsymbol{\pi}}}(s) \leq \epsilon, \quad \forall s \in \mathcal{S}';$$
$$\tilde{V_2}^{\hat{\boldsymbol{\pi}}_1,\dagger}(s) - \tilde{V_2}^{\hat{\boldsymbol{\pi}}}(s) \leq \epsilon, \quad \forall s \in \mathcal{S}'.$$

it is the case that,

$$V_1{}^{\dagger,\hat{\boldsymbol{\pi}}_2}(s) - V_1{}^{\hat{\boldsymbol{\pi}}}(s) \leq \epsilon, \quad \forall s \in \mathcal{S};$$
$$V_2{}^{\hat{\boldsymbol{\pi}}_1,\dagger}(s) - V_2{}^{\hat{\boldsymbol{\pi}}}(s) \leq \epsilon, \quad \forall s \in \mathcal{S}.$$

For $s \in \mathcal{S}'_0$, it follows from the previous claim. For $s \in \mathcal{S}' \setminus \mathcal{S}'_0$, we observe that varying policies do not alter the value vector $\tilde{V}$ of any given player.

5.2 Conclusion

After considering a number of assumptions on the reward function structure of Markov games, we were able to retrieve an array of positive as well as negative results. After experimenting with monotone and potential structures, we are able to retrieve the hardness results of (Deng et al., 2021) even for the case of Markov games where the rewards are independent of the actions, and only varying from state to state. We conclude that, in general, assumptions on the structure of the rewards are only *necessary* for tractability of equilibria. Further, even a quite strong such assumption, constant-per-state rewards, is ineffective in ameliorating the computational hardness of equilibrium computation. In our opinion, one lesson to be learned is that one needs to dive deeper into the exciting world of Markov games viewed from a perspective of distributed control of state visitation. What are the natural assumptions that we can distill from observing real-world strategic interactions in dynamic environments?

Bibliography

Ilan Adler, Constantinos Daskalakis, and Christos H Papadimitriou. A note on strictly competitive games. In *Internet and Network Economics: 5th International Workshop, WINE 2009, Rome, Italy, December 14-18, 2009. Proceedings 5*, pages 471–474. Springer, 2009.

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*, pages 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.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, 2022. URL `https://api.semanticscholar.org/CorpusID:247619158`.

Ioannis Anagnostides, Fivos Kalogiannis, Ioannis Panageas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Stephen McAleer. Algorithms and complexity for computing nash equilibria in adversarial team games. *arXiv preprint arXiv:2301.02129*, 2023.

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

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

Shant Boodaghians, Joshua Brakensiek, Samuel B Hopkins, and Aviad Rubinstein. Smoothed complexity of 2-player nash equilibria. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 271–282. IEEE, 2020.

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

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

LM Bregman and IN Fokin. Methods of determining equilibrium situations in zero-sum polymatrix games. *Optimizatsia*, 40(57):70–82, 1987.

Yang Cai and Constantinos Daskalakis. On minmax theorems for multiplayer games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, pages 217–234. SIAM, 2011.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41 (2):648–655, 2016.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. *Advances in Neural Information Processing Systems*, 35: 33904–33919, 2022.

Xi Chen and Xiaotie Deng. Settling the complexity of two-player nash equilibrium. In *FOCS*, volume 6, pages 261–272, 2006.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Constantinos Daskalakis and Christos H Papadimitriou. On a network generalization of the minmax theorem. In *International Colloquium on Automata, Languages, and Programming*, pages 423–434. Springer, 2009.

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.

Xiaotie Deng, Yuhao Li, David Henry Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *arXiv preprint arXiv:2109.01795*, 2021.

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

Eyal Even-Dar, Yishay Mansour, and Uri Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 523–532, 2009.

Alex Fabrikant, Christos Papadimitriou, and Kunal Talwar. The complexity of pure nash equilibria. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 604–612, 2004.

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.

János Flesch, Frank Thuijsman, and Okko Jan Vrieze. Stochastic games with additive transitions. *European Journal of Operational Research*, 179(2):483–497, 2007.

János Flesch, Gijs Schoenmakers, and Koos Vrieze. Stochastic games on a product state space. *Mathematics of Operations Research*, 33(2):403–420, 2008.

Drew Fudenberg and David K Levine. Open-loop and closed-loop equilibria in dynamic games with many players. *Journal of Economic Theory*, 44(1):1–18, 1988.

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.

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

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*, pages 4880–4889. PMLR, 2020.

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

Fivos Kalogiannis and Ioannis Panageas. Zero-sum polymatrix markov games: Equilibrium collapse and efficient computation of nash equilibria. *arXiv preprint arXiv:2305.14329*, 2023.

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

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.

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

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

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

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

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

David H Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*, pages 7688–7699. PMLR, 2021.

SR Mohan and TES Raghavan. An algorithm for discounted switching control stochastic games. *Operations-Research-Spektrum*, 9(1):41–45, 1987.

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

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*, pages 286–295, 1951.

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

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994.

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

Christos H Papadimitriou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Manolis Zampetakis. The computational complexity of multi-player concave games and kakutani fixed points. *arXiv e-prints*, pages arXiv–2207, 2022.

Chanwoo Park, Kaiqing Zhang, and Asuman Ozdaglar. Multi-player zero-sum markov games with networked separable interactions. *arXiv preprint arXiv:2307.09470*, 2023.

Thiruvenkatachari Parthasarathy and TES Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.

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

Tiancheng Qin and S Rasoul Etesami. Scalable and independent learning of nash equilibrium policies in n-player stochastic games with unknown independent chains. *arXiv preprint arXiv:2312.01587*, 2023.

Tirukkannamangai ES Raghavan, SH Tijs, and OJ Vrieze. On stochastic games with additive reward and transition structure. *Journal of Optimization Theory and Applications*, 47: 451–464, 1985.

Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

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

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

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 O Sayin. Decentralized learning for stochastic games: Beyond zero sum and identical interest. *arXiv preprint arXiv:2310.07256*, 2023.

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

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

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

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(1-2):309–321, 1997.

OJ Vrieze, SH Tijs, TES Raghavan, and JA Filar. A finite algorithm for the switching control stochastic game. *Or Spektrum*, 5(1):15–24, 1983.

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

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.

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

Runyu Zhang, Yuyang Zhang, Rohit Konda, Bryce L. Ferguson, Jason R. Marden, and Na Li. Markov games with decoupled dynamics: Price of anarchy and sample complexity. *CoRR*, abs/2304.03840, 2023. doi: 10.48550/arXiv.2304.03840. URL `https://doi.org/10.48550/arXiv.2304.03840`.

Appendix A

Background on Nonlinear Programming

In this chapter, we provide additional background on the theory of nonlinear programming (Mangasarian, 1994).

When dealing with a constrained minimization problem, we aim to identify conditions that confirm the presence of (nonnegative) Lagrange multipliers that adhere to the *Karush-Kuhn-Tucker (KKT)* conditions. In the case where there are no constraints, this criterion aligns with the gradient being zero at a local optimum (Fermat's Theorem). However, in cases with constraints, additional regularity conditions regarding the feasible set must be satisfied. This requirement is formalized through what are known as *constraint qualifications* (Bazaraa et al., 1972; Giorgi et al., 2018). For our needs, 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 D.9).

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 in a Euclidean space $\mathbb{R}^d$, where d is a natural number, the objective is to optimize a given function $f : \mathbb{R}^d \to \mathbb{R}$ over a nonempty set $\mathcal{D} \subseteq \mathbb{R}^d$. The function f is referred to as the objective function, and $\mathcal{D}$ is the constraint or feasibility set. Such problems are denoted as follows:

$$\text{“Minimize } f(\mathbf{z}) \text{ subject to } \mathbf{z} \in \mathcal{D}\text{”},$$

or more concisely as

$$\text{“min}\{f(\mathbf{z}) \mid \mathbf{z} \in \mathcal{D}\}.\text{”}$$

A global solution to such a problem is a point $\mathbf{z}^*$ in $\mathcal{D}$ such that $f(\mathbf{z}^*) \leq f(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{D}$; the existence of such a solution is typically guaranteed by Weierstrass' theorem.

Relaxing the requirement of global optimality, we define a local minimum as follows:

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

$$f(\mathbf{z}_0) \leq f(\mathbf{z}), \quad \forall \mathbf{z} \in \{\mathbf{z} \mid \mathbf{z} \in B(\mathbf{z}_0, \delta) \cap \mathcal{D}\},$$

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

We now turn to study constrained optimization problems with feasible sets defined by in-

equality constraints. Namely, the constraint set will have the form

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

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.

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

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*

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

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* (A.1)*. 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*

$$\begin{cases} \boldsymbol{w}^\top \nabla_{\boldsymbol{z}} g_i(\boldsymbol{z}_0) \geq 0, \quad \forall i \in A'; \text{ and} \\ \boldsymbol{w}^\top \nabla_{\boldsymbol{z}} g_i(\boldsymbol{z}_0) > 0, \quad \forall i \in A'', \end{cases} \tag{A.2}$$

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 (A.2)—does *not* involve the objective function; this is the case more broadly for constraint qualifications.

Appendix B

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 and 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 dist the distance between a vector $\boldsymbol{x}$ and $\mathcal{Y}$, defined as follows.

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

Definition B.1 (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 λ-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 λ-strongly convex.

Corollary B.1. *Let $f : \mathcal{X} \ni \boldsymbol{x} \mapsto \mathbb{R}$ be a λ-weakly convex function. Then, the function $f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|^2$ is λ-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 ℓ being the corresponding parameter of weak convex, the Moreau envelope f_λ is C^1-smooth, and its gradient given by $\nabla f_\lambda = \lambda^{-1}(\boldsymbol{x} - \operatorname{prox}_{\lambda f}(\boldsymbol{x}))$ (Rockafellar, 1970, Theorem 31.5), where $\operatorname{prox}_{\lambda f}(\cdot)$ is the *proximal mapping*. Namely, for a convex and continuous function $f : \mathcal{X} \to \mathbb{R}$ we define its proximal operator $\operatorname{prox}_f : \mathbb{R}^d \to \mathbb{R}^d$ as follows.

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

The point $\tilde{\boldsymbol{x}} \coloneqq \operatorname{prox}_f(\boldsymbol{x})$ that results by applying the proximal operator (B.1) on $\boldsymbol{x}$ is called the *proximal point* of $\boldsymbol{x}$. The proximal point of the scaled function λf 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 ℓ-weakly convex function is well-defined, as as long as λ is sufficiently small:

Proposition B.1 (Lin et al. (2020))**.** *Let ϕ be a ℓ-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 and Kabadi, 1985). Instead, one is often interested in computing an approximate stationary point of the objective function.

More precisely, an ϵ-approximate stationary point $\boldsymbol{x}_0$ of a nondifferentiable function is a point such that $\text{dist}(\mathbf{0}; \partial f(\boldsymbol{x}_0)) \leq \epsilon$ where $\partial f(\boldsymbol{x}_0)$ is the *subdifferential* of f at $\boldsymbol{x}_0$ (see (Davis and 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 ϵ-approximate stationary point for $\epsilon \in [0, 1)$.

The alternative notion of *near stationarity* for a nonsmooth function $f(\boldsymbol{x})$, contributed by Davis and 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 and Paquette, 2019, Section 4.1).) More precisely, we measure stationarity by means of the proximal operator:

Definition B.2 (ϵ-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 ϵ-nearly stationary if*

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

where $\tilde{\boldsymbol{x}}_0 \coloneqq \text{prox}_{\lambda f}(\boldsymbol{x}_0)$ 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 B.1 ((Davis and Drusvyatskiy, 2019))**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be an ℓ-weakly convex function*

and $\lambda < \frac{1}{\ell}$. *Further, let* $\boldsymbol{x} \in \mathcal{X}$ *and* $\tilde{\boldsymbol{x}} \coloneqq \operatorname{prox}_{\lambda f}(\boldsymbol{x})$ *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 B.1. *An* $\frac{\epsilon}{\lambda}$*-approximate first-order order stationary point of* f_λ *is an* ϵ*-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 ϕ is fairly common in the context of min-max optimization. The following lemma points out some useful properties of ϕ.

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

- L*-Lipschitz continuous; and*
- ℓ*-weakly convex.*

Appendix C

Auxiliary Lemmata for Markov Games

Boundedness of value.

Fact C.1. *Let the reward functions be bounded in* $[0,1]$*, i.e.,* $0 \leq r_h(s, \boldsymbol{a}) \leq 1,\ \forall s \in \mathcal{S}, \forall \boldsymbol{a}\mathcal{A}$*, it holds that,*

- $V_{i,h}(s) \leq H - h,\ \forall i \in [n], \forall h \in [H]$*;*
- $Q_{i,h}(s,a) \leq h,\ \forall i \in [n], \forall H - h \in [H], \forall a \in \mathcal{A}_i$*.*

Lipschitz continuity of rewards and transitions.

Claim C.1. *In a MG* $\Gamma\left(n, H, \mathcal{S}, \mathcal{A}, \mathbb{P}, \{r_i\}_{i\in[n]}, \gamma, \boldsymbol{\rho}\right)$ *with additive transitions, the following inequalities hold true for any* $\boldsymbol{\pi}_{s,h}, \boldsymbol{\pi}'_{s,h}$ *and any* $s \in \mathcal{S}$*:*

- $r_{i,h}(s, \boldsymbol{\pi}_{s,h}) - r_{i,h}(s, \boldsymbol{\pi}'_{s,h}) \leq \sqrt{\sum_{i\in[n]} |\mathcal{A}_i|}\, \left\|\boldsymbol{\pi}_{s,h} - \boldsymbol{\pi}'_{s,h}\right\|$*;*

- $\left|\sum_{s' \in \mathcal{S}} \left(\mathbb{P}_h(s'|s, \boldsymbol{\pi}_h) - \mathbb{P}_h(s'|s, \boldsymbol{\pi}'_h)\right) V_{i,h+1}(s')\right| \leq H|\mathcal{S}| \max_{i \in [n]} \sqrt{|\mathcal{A}_i|}$.

Proof. We use standard inequalities:

- Fixing any $i, s, h \in [n] \times \mathcal{S} \times [H]$, we have

$$r_{i,h}(s, \boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}}[r_{i,h}(s, \boldsymbol{a})] = \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r_{i,h}(s, \boldsymbol{a}) \prod_{i=1}^{n} \pi_{i,s,h}(a_i).$$

  As a result,

$$\begin{aligned}
&|r_{i,h}(s, \boldsymbol{\pi}) - r_{i,h}(s, \boldsymbol{\pi}')| \\
&= \left| \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r_{i,h}(s, \boldsymbol{a}) \prod_{i=1}^{n} \boldsymbol{\pi}_{i,s,h}(a_i) - \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r_{i,h}(s, \boldsymbol{a}) \prod_{i=1}^{n} \boldsymbol{\pi}'_{i,s,h}(a_i) \right| \\
&= \left| \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} r_{i,h}(s, \boldsymbol{a}) \left(\prod_{i=1}^{n} \boldsymbol{\pi}_{i,s,h}(a_i) - \prod_{i=1}^{n} \boldsymbol{\pi}'_{i,s,h}(a_i) \right) \right| \\
&\leq \sum_{(a_1, \ldots, a_n) \in \mathcal{A}} \left| \prod_{i=1}^{n} \boldsymbol{\pi}_{i,s,h}(a_i) - \prod_{i=1}^{n} \boldsymbol{\pi}'_{i,s,h}(a_i) \right| \qquad \text{(C.1)} \\
&\leq \sum_{k=1}^{n} \|\boldsymbol{\pi}_{i,s,h} - \boldsymbol{\pi}'_{i,s,h}\|_1 = \|\boldsymbol{\pi}_{s,h} - \boldsymbol{\pi}'_{s,h}\|_1 \\
&\leq \left(\sqrt{\sum_{i=1}^{n} A_i} \right) \|\boldsymbol{\pi}_{s,h} - \boldsymbol{\pi}'_{s,h}\|_2, \qquad \text{(C.2)}
\end{aligned}$$

  where (C.1) follows from the fact that $|r_{i,h}(s, \cdot)| \leq 1$ and the triangle inequality. (C.2) follows from the fact that the total variation distance between two distributions is bounded by the sum of total variation distances between their respective marginal distributions (Hoeffding and Wolfowitz, 1958), and the equivalence between ℓ_1-norm and ℓ_2-norm — *i.e.*, $\|\boldsymbol{x}\|_1 \leq \sqrt{m}\|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{R}^m$).

- the second item is proved using the same line of arguments along with the assumption

of additive transitions and the fact that $\left|V_{i,h}^{\boldsymbol{\pi}}(s)\right| \leq H - h$.

□

Appendix D

Missing Proofs and Statements

D.1 Statements for Section 4.2

Claim D.1. *Let a two-player Markov game where both players affect the transition. Further, consider a correlated policy* $\boldsymbol{\sigma}$ *and its corresponding marginalized product policy* $\boldsymbol{\pi}^{\boldsymbol{\sigma}} = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$. *Then, for any* $\boldsymbol{\pi}_1', \boldsymbol{\pi}_2'$,

$$V_{k,1}^{\boldsymbol{\pi}_1', \boldsymbol{\sigma}_{-1}}(s_1) = V_{k,1}^{\boldsymbol{\pi}_1', \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}}(s_1),$$
$$V_{k,2}^{\boldsymbol{\sigma}_{-2}, \boldsymbol{\pi}_2'}(s_1) = V_{k,2}^{\boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, \boldsymbol{\pi}_2'}(s_1).$$

Proof. We will effectively show that the problem of best-responding to a correlated policy $\boldsymbol{\sigma}$ is equivalent to best-responding to the marginal policy of $\boldsymbol{\sigma}$ for the opponent. The proof follows from the equivalence of the two MDPs.

As a reminder,

$$\pi^{\sigma}_{1,h}(a|s) = \sum_{b \in \mathcal{A}_2} \boldsymbol{\sigma}_h(a, b|s)$$
$$\pi^{\sigma}_{2,h}(b|s) = \sum_{a \in \mathcal{A}_1} \boldsymbol{\sigma}_h(a, b|s)$$

As we have seen in Section 2.2, in the case of unilateral deviation from joint policy $\boldsymbol{\sigma}$, an agent faces a single agent MDP. More specifically, agent 2, best-responds by optimizing a reward function $\overline{r}_{2,h}(s, b)$ under a transition kernel $\overline{\mathbb{P}}_2$ for which,

$$\overline{r}_{2,h}(s, b) = \mathbb{E}_{b \sim \boldsymbol{\sigma}} \left[r_{2,h}(s, a, b) \right] = \mathbb{E}_{b \sim \boldsymbol{\pi}^{\sigma}_1} \left[r_{2,h}(s, a, b) \right] = r_{2,h}(s, \boldsymbol{\pi}^{\sigma}_1, b).$$

Similarly,

$$\overline{r}_{1,h}(s, b) = r_{1,h}(s, a, \boldsymbol{\pi}^{\sigma}_2).$$

Analogously, for each of the transition kernels,

$$\overline{\mathbb{P}}_{2,h}(s'|s, b) = \mathbb{E}_{a \sim \boldsymbol{\sigma}} \left[\mathbb{P}_{2,h}(s'|s, a, b) \right] = \mathbb{E}_{a \sim \boldsymbol{\pi}^{\sigma}_2} \left[\mathbb{P}_{2,h}(s'|s, a, b) \right] = \mathbb{P}_{2,h}(s'|s, \boldsymbol{\pi}^{\sigma}_1, b),$$

as for agent 1,

$$\overline{\mathbb{P}}_{1,h}(s'|s, a) = \mathbb{P}_{1,h}(s'|s, a, \boldsymbol{\pi}^{\sigma}_2).$$

Hence, it follows that, $V^{\boldsymbol{\sigma}^{-2} \times \boldsymbol{\pi}'_2}_{2,1}(s_1) = V^{\boldsymbol{\pi}^{\sigma}_1 \times \boldsymbol{\pi}'_2}_{2,1}(s_1),\ \forall \boldsymbol{\pi}'_2$ and $V^{\boldsymbol{\pi}'_1 \times \boldsymbol{\sigma}^{-1}}_{1,1}(s_1) = V^{\boldsymbol{\pi}'_1 \times \boldsymbol{\pi}^{\sigma}_2}_{1,1}(s_1),\ \forall \boldsymbol{\pi}'_2$.

□

D.2 Missing statements and proofs for Section 4.5

D.2.1 Proof of Theorem 4.7: NE computation in RPMGs

Auxiliary lemmata. There are two key lemmata in the proof of Theorem 4.7; one of them tells us that the game with individual utilities $\left\{r_{i,h}(s,\cdot) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,\cdot)V_{i,h+1}(s')\right\}_{i\in[n]}$ is a potential game —w.r.t. policies $\boldsymbol{\pi}_h$ of the corresponding timestep h— no matter the (fixed) value vector, $\boldsymbol{V}_{i,h+1}$, of the future states. The second lemma parametrizes the latter games with vectors $\boldsymbol{V}_{i,h+1}$ that correspond to δ-approximate NE for the $\Gamma_{s,h+1}$ subgames; then, it is demonstrated that an ϵ-approximate NE in this game is also a $(\delta+\epsilon)$-approximate NE of the $\Gamma_{s,h}$ subgames.

Lemma D.1 (Potential game when future values fixed)**.** *Fix a timestep $h \in [H]$ and let arbitrary vectors $\{\boldsymbol{v}_i \in \mathbb{R}^{|\mathcal{S}|}\}_{i\in[n]}$. Moreover, for every $s \in \mathcal{S}$ assume game with individual utilities $\{r_{i,h}(s,\cdot) + \sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,\cdot)v_i(s)\}$. Each such game is a potential game.*

Proof. Indeed, let function $\psi_h(s,\cdot) = \phi_h(s,\cdot) + \sum_{i\in[n]} \sum_{s'\in\mathcal{S}} \omega_{i,s,h}\, \mathbb{P}_{i,h}(s,\cdot)v_i(s')$. We remind the reader that $\mathbb{P}_h(s'|s,\boldsymbol{\pi}) = \sum_{i\in[n]} \omega_{i,s,h}\, \mathbb{P}(s'|s,\boldsymbol{\pi}_i)$ due to the additive transitions assumption. It holds for function $\psi_h(s,\cdot)$, that,

$$
\begin{aligned}
&\psi_h(s,\boldsymbol{\pi}_h) - \psi_h(s,\boldsymbol{\pi}'_{i,h},\boldsymbol{\pi}_{-i,h}) \\
&\quad= \phi_h(s,\boldsymbol{\pi}_h) - \phi_h(s,\boldsymbol{\pi}'_{i,h},\boldsymbol{\pi}_{-i,h}) + \omega_{i,s,h} \sum_{s'\in\mathcal{S}} \left(\mathbb{P}_{i,h}(s'|s,\boldsymbol{\pi}_{i,h})v(s') - \mathbb{P}_{i,h}(s'|s,\boldsymbol{\pi}'_{i.h})v(s')\right) \\
&\quad= r_{i,h}(s,\boldsymbol{\pi}_h) - r_{i,h}(s,\boldsymbol{\pi}'_{i,h},\boldsymbol{\pi}_{-i,h}) + \omega_{i,s,h} \sum_{s'\in\mathcal{S}} \left(\mathbb{P}_{i,h}(s'|s,\boldsymbol{\pi}_{i,h})v(s') - \mathbb{P}_{i,h}(s'|s,\boldsymbol{\pi}'_{i.h})v(s')\right)
\end{aligned}
$$

The last inequality follows from the reward-potential assumption and completes the proof. □

For brevity, we simplify the notation for the following claim that we need for the promised

second lemma.

Claim D.2 (Approximate best reponses)**.** *Let* $\hat{\boldsymbol{v}}, \boldsymbol{v}^\dagger \in \mathbb{R}^{\mathcal{S}}$ *such that* $\|\hat{\boldsymbol{v}} - \boldsymbol{v}^\dagger\|_\infty \leq \delta$*. Further, let function* $r : \mathcal{A} \to \mathbb{R}$ *and transition kernel* $\boldsymbol{p} : \mathcal{A} \to \Delta(\mathcal{S})$*, it holds that,*

$$\left| \max_{\boldsymbol{x}' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}') \hat{v}(s') \right\} - \max_{\boldsymbol{x}'' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}'') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}'') v^\dagger(s') \right\} \right| \leq \delta.$$

Proof. It follows that for every $a \in \mathcal{A}$,

$$r(a) + \sum_{s' \in \mathcal{S}} p(s'|a) \hat{v}(s') - \left(r(a) + \sum_{s' \in \mathcal{S}} p(s'|a) v^\dagger(s') \right) = \sum_{s' \in \mathcal{S}} p(s'|a) \left(\hat{v}(s') - v^\dagger(s') \right) \leq \delta.$$

Since the difference,

$$\left| \max_{\boldsymbol{x}' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}') \hat{v}(s') \right\} - \max_{\boldsymbol{x}'' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}'') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}'') v^\dagger(s') \right\} \right|. \tag{D.1}$$

From linearity, it holds that,

$$\max_{\boldsymbol{x}' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}') \hat{v}(s') \right\} = \max_{a \in \mathcal{A}} \left\{ r(a) + \sum_{s' \in \mathcal{S}} p(s'|a) \hat{v}(s') \right\}$$

and

$$\max_{\boldsymbol{x}'' \in \Delta(\mathcal{A})} \left\{ r(\boldsymbol{x}'') + \sum_{s' \in \mathcal{S}} p(s'|\boldsymbol{x}'') v^\dagger(s') \right\} = \max_{a \in \mathcal{A}} \left\{ r(a) + \sum_{s' \in \mathcal{S}} p(s'|a) v^\dagger(s') \right\}$$

.

The last two displays in combination with (D.1) which holds for all $a \in \mathcal{A}$ complete the proof of the claim. □

The last claim proves the following lemma,

Lemma D.2. *Let* $\{\hat{\boldsymbol{V}}_{i,h+1}\}_{i\in[n]}$ *be a collection of value vectors that corresponds to a* δ*-approximate NE,* $\{\boldsymbol{\pi}_\tau\}_{\tau\in\{h+1,\ldots,H\}}$*, for the subgames* $\{\Gamma_{s,h+1}\}_{s\in\mathcal{S}}$*. Further, let an* ϵ*-approximate NE,* $\hat{\boldsymbol{\pi}}_h$ *of the games with individual utilities* $\left\{r_{i,h}(s,\cdot)+\sum_{s'\in\mathcal{S}}\mathbb{P}_h(s'|s,\cdot)\hat{V}_{i,h+1}(s')\right\}_{i\in[n]}$*. Then* $\{\boldsymbol{\pi}_\tau\}_{\tau=\{h,\ldots,H\}}$ *is a* $(\delta+\epsilon)$*-approximate NE for subgames* $\{\Gamma_{s,h}\}_{s\in\mathcal{S}}$*.*

The complexity of implementing the NE-Oracle**.** Now, we invoke a theorem that bounds the number of iterations needed to compute an ϵ-approximate NE in a potential game when every player employs the mirror-descent algorithm with a fixed stepsize.

Theorem D.1 (Theorem B.6 in (Anagnostides et al., 2022))**.** *Assume a potential game* $\Gamma\left(n,\{\mathcal{A}_i\}_{i\in[n]},\{u_{i\in[n]}\}\right)$ *with potential function* $\Phi:\prod_{i=1}^n\mathcal{A}_i\to\mathbb{R}$*.* Φ *is* L*-Lipschitz continuous. Suppose that each player* i *employs mirror-descent*

- *with stepsize* $\eta=\frac{1}{2L}$*,*
- *with regularizer* $\mathcal{R}_i(\boldsymbol{x})$*, and* $\nabla\mathcal{R}_i(\boldsymbol{x})$ G*-Lipschitz continuous,*
- *and* Diam *is the maximum diameter of the a player's probability simplex due to their use of regularizer* $\mathcal{R}_i$*.*

Further, let $T=\left\lceil\frac{\eta\Phi_{\max}}{\epsilon^2}\right\rceil+2$*, then it holds that,* $\exists t^\star\in[T]$*, such that,* $\boldsymbol{x}^{t^\star}$ *is an* $\epsilon\left(\frac{G\text{Diam}}{\eta}+\max_{i\in[n]}\sqrt{|\mathcal{A}_i|}\right)$*-approximate Nash equilibrium.*

Bounding the total iteration complexity. Equipped with the latter bound, we are ready to state our bound on the iteration complexity of computing an approximate NE in RPMGs.

Theorem D.2 (Full version of Theorem 4.7)**.** *Algorithm 3 with* NE-Oracle *implemented using projected gradient descent with stepsize* $\eta=\frac{1}{2L}$ *for every agent* $i\in[n]$*, input accuracy*

ϵ/H for every h, computes an ϵ-approximate nonstationary NE for an RPMG with additive transitions converges with a total number of iterations

$$\frac{128nH^5|\mathcal{S}|^2 \max_{i\in[n]} |\mathcal{A}_i|^{5/2}}{\epsilon^2}.$$

Proof. We remind the reader that the projected gradient descent algorithm is equivalent to mirror descent with $\mathcal{R}_i(\cdot) = \frac{1}{2}\|\cdot\|^2$. Hence, order to achieve accuracy ϵ/H, every projected gradient descent subroutine needs $T = \left\lceil \frac{8L\Phi_{\max}G^2\mathrm{Diam}^2 \max_{i\in[n]} |\mathcal{A}_i|}{\epsilon^2} \right\rceil + 2$ iterations. In our context, this translates to:

$$T = \left\lceil \frac{128nH^2|\mathcal{S}| \max_{i\in[n]} |\mathcal{A}_i|^{5/2}}{\epsilon^2} \right\rceil + 2.$$

Where we have taken $\mathrm{Diam} = 2\max_{i\in[n]} \sqrt{|\mathcal{A}_i|}$, $G = 1$, $\Phi_{\max} = H$. and we have bounded the Lispchitz-continuity parameter of each $\Gamma_{s,h}$ subgame by $L = 4nH|\mathcal{S}| \max_{i\in n} \sqrt{\mathcal{A}}$ due to Claim C.1. Then, we inductively invoke lemma D.2 to conclude that after H (backwards) inductive steps, we accumulate an approximation error at most $H\frac{\epsilon}{H} = \epsilon$.

Concluding, we need $|\mathcal{S}|H$ calls to the NE-Oracle with accuracy ϵ/H, raising the total iteration complexity to the stated number.

□

D.2.2 Proofs for Section 4.5.1

Theorem D.3. *Finite-horizon reward-potential games with additive transitions assert pure Nash equilibria.*

Proof. By convention $V_{i,H}(s) = 0, \forall i \in [n], \forall s \in \mathcal{S}$. Further, for $h = H - 1$, the game played in every state s asserts at least one pure Nash equilibrium (Monderer and Shapley, 1996). Then, by Lemma D.1 and Lemma D.2 the claim holds. □

Following using a standard trick we prove the following:

Corollary D.1. *Infinite-horizon RPMGs with discount parameter γ, attain a deterministic nonstationary approximate NE that can be computed in time* $\mathsf{poly}\left(\frac{1}{\epsilon}, \frac{1}{1-\gamma}, \sum_{i\in[n+1]} |\mathcal{A}_i|, |\mathcal{S}|\right)$.

Proof. As proposed in (Daskalakis et al., 2022, Theorem 4.2), the infinite-horizon game can be converted into a finite-horizon one in order to compute nonstationary policies of the initial game. These nonstationary policies of course cannot span the whole horizon of the game; it suffices that they only consider the first $H := \frac{\log(1/\epsilon)}{1-\gamma}$ steps of the game where ϵ is the desired accuracy of the equilibrium that is sought after.

After truncating the horizon into a finite one, every reward function is scaled according to the initial discounting factor, *i.e.,* $r_{i,h}(s, \cdot) = \gamma^{h-1} r_i(s, \cdot)$, where $r_i(s, \cdot)$ are the reward functions of the infinite-horizon game.

The complexity of computation follows from known results about the computational complexity of pure approximate NE in potential games (Fabrikant et al., 2004) and the use of backwards induction. □

D.2.3 Proofs for Section 4.5.3: ARPMGs

First, we prove that although the subgames defined are not adversarial potential games *per se*, the variational inequalities corresponding to their approximate NE coincide with the variational inequalities of a certain adversarial team game.

Proposition D.1. *Let an ARPMG with additive transitions,* $\Gamma(n+1, H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \boldsymbol{\rho})$*, and* $\hat{\boldsymbol{V}}_{i,h+1}$ *be the value vector for the* δ*-approximate NE of the subgames* $\Gamma_{s,h+1}$*. Let the adversarial team normal-form games* $\Gamma'_s, \forall s \in \mathcal{S}$*, each with* n *players in the team and one adversary. Define the utility function of the team to be,*

$$\begin{aligned} u(s, \boldsymbol{\pi}) \coloneqq & \phi_h(s, \boldsymbol{\pi}) + \sum_{s' \in \mathcal{S}} \sum_{j \in [n]} \omega_{j,s,h} \, \mathbb{P}_{j,h}(\boldsymbol{\pi}_j) \hat{V}_{j,h+1}(s') \\ & - \sum_{s' \in \mathcal{S}} \omega_{adv,s,h} \, \mathbb{P}_{adv,h}(\boldsymbol{\pi}_{adv}) \hat{V}_{adv,h+1}(s'). \end{aligned}$$

An ϵ*-approximate NE of each subgame* Γ'_s *is also an* $(\epsilon + \delta)$*-approximate NE of the* $\Gamma_{s,h}$ *subgame.*

Proof. For brevity, let $\boldsymbol{x}_i \coloneqq \boldsymbol{\pi}_{i,h}, \forall i \in [n]$, with $\boldsymbol{x} \coloneqq (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and $\boldsymbol{y} \coloneqq \boldsymbol{\pi}_{\text{adv},h}$. Further, $\mathcal{X} \coloneqq \prod_{i \in [n]} \Delta(\mathcal{A}_i)$ and $\mathcal{Y} \coloneqq \Delta(\mathcal{A}_{n+1})$. Then, we write $u^s(\boldsymbol{\pi}) = u^s(\boldsymbol{x}, \boldsymbol{y})$. An ϵ-approximate NE to the game is computed by solving the following variational inequality problem,

$$\nabla_{\boldsymbol{x}} u(s, \boldsymbol{x}^\star, \boldsymbol{y}^\star)^\top (\boldsymbol{x}^\star - \boldsymbol{x}) \leq \epsilon, \forall \boldsymbol{x} \in \mathcal{X}$$

and

$$\nabla_{\boldsymbol{y}} u(s, \boldsymbol{x}^\star, \boldsymbol{y}^\star)^\top (\boldsymbol{y}^\star - \boldsymbol{y}) \geq -\epsilon, \forall \boldsymbol{y} \in \mathcal{Y}.$$

By computing such a point $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, it is also the case that,

$$\begin{aligned} & \nabla_{\boldsymbol{y}} \left(r_{\text{adv},h}(s, \boldsymbol{x}^\star, \boldsymbol{y}^\star) + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s' | s, \boldsymbol{x}^\star, \boldsymbol{y}^\star) \hat{V}_{\text{adv},h+1}(s') \right) \\ & = \nabla_{\boldsymbol{y}} \left(-u(s, \boldsymbol{x}^\star, \boldsymbol{y}^\star) \right) \end{aligned}$$

We observe that,

$$\begin{aligned}
&\nabla_{\boldsymbol{y}} \left(r_{\text{adv},h}(s, \boldsymbol{x}, \boldsymbol{y}) + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, \boldsymbol{x}, \boldsymbol{y}) \hat{V}_{\text{adv},h+1}(s') \right) \\
&\quad = \nabla_{\boldsymbol{y}} \left(-\phi_{s,h}(\boldsymbol{x}, \boldsymbol{y}) + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, \boldsymbol{x}, \boldsymbol{y}) \hat{V}_{\text{adv},h+1}(s') \right) \\
&\quad = -\nabla_{\boldsymbol{y}} u(s, \boldsymbol{x}, \boldsymbol{y}).
\end{aligned}$$

By computing such a point $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, it is also the case that,

$$\begin{aligned}
&\nabla_{\boldsymbol{x}} \left(\phi_h(s, \boldsymbol{x}, \boldsymbol{y}) + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, \boldsymbol{x}, \boldsymbol{y}) \hat{V}_{\text{adv},h+1}(s') \right)^\top (\boldsymbol{y}^\star - \boldsymbol{y}) \leq \epsilon, \forall \boldsymbol{y} \in \mathcal{Y}, \\
&\nabla_{\boldsymbol{y}} \left(r_{\text{adv},h}(s, \boldsymbol{x}, \boldsymbol{y}) + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s, \boldsymbol{x}, \boldsymbol{y}) \hat{V}_{\text{adv},h+1}(s') \right)^\top (\boldsymbol{y}^\star - \boldsymbol{y}) \geq -\epsilon, \forall \boldsymbol{y} \in \mathcal{Y}.
\end{aligned}$$

Concluding, such a strategy $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is also a $(\delta + \epsilon)$-approximate NE for the subgame $\Gamma_{s,h}$. □

This translates to the fact that the template algorithm, Algorithm 3, can be modified in order to compute apprximate NEs for ARPMG using the algorithm proposed in (Anagnostides et al., 2023).

D.3 Missing statements and proofs for Section 4.6

D.3.1 Proof of Theorem 4.10

The best-response program. First, we state the following lemma that will prove useful for several of our arguments,

Lemma D.3 (Best-response LP)**.** *Let a (possibly correlated) joint policy $\hat{\boldsymbol{\sigma}}$. Consider the following linear program with variables $\boldsymbol{w} \in \mathbb{R}^{n \times H \times S}$,*

$$
\begin{aligned}
\min \quad & \sum_{k \in [n]} w_{k,s}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\hat{\boldsymbol{\sigma}}_\tau) \right) \boldsymbol{r}_{k,h}(\hat{\boldsymbol{\sigma}}_h) \\
\text{s.t. } & w_{k,h}(s) \geq r_{k,h}(s, a, \hat{\boldsymbol{\sigma}}_{-k,h}) + \mathbb{P}_h(s, a, \hat{\boldsymbol{\sigma}}_{-k,h}) \boldsymbol{w}_{k,h+1}, \\
& \qquad \forall s \in \mathcal{S}, \forall h \in [H], \forall k \in [n], \forall a \in \mathcal{A}_k; \\
& \qquad w_{k,H}(s) = 0, \ \forall k \in [n], \forall s \in \mathcal{S}.
\end{aligned} \tag{$\mathrm{P_{BR}}$}
$$

The optimal solution $\boldsymbol{w}^\dagger$ of the program is unique and corresponds to the value function of each player $k \in [n]$ when player k best-responds to $\hat{\boldsymbol{\sigma}}$.

Proof. We observe that the program is separable to n independent linear programs, each with variables $\boldsymbol{w}_k \in \mathbb{R}^{n \times H}$,

$$
\begin{aligned}
\min \ & w_{k,1}(s_1) \\
\text{s.t. } & w_{k,h}(s) \geq r_{k,h}(s, a, \hat{\boldsymbol{\sigma}}_{-k,h}) + \mathbb{P}_h(s, a, \hat{\boldsymbol{\sigma}}_{-k,h}) \boldsymbol{w}_{k,h+1}, \\
& \qquad \forall s \in \mathcal{S}, \forall h \in [H], \forall a \in \mathcal{A}_k; \\
& \qquad w_{k,H}(s) = 0, \ \forall k \in [n], \forall s \in \mathcal{S}.
\end{aligned}
$$

Each of these linear programs describes the problem of a single agent MDP (Neu and Pike-Burke, 2020, Section 2) —that agent being k— which, as we have seen in Best-response policies, is equivalent to the problem of finding a best-response to $\hat{\boldsymbol{\sigma}}_{-k}$. It follows that the optimal $\boldsymbol{w}_k^\dagger$ for every program is unique (each program corresponds to a set of Bellman optimality equations). □

Properties of the NE program. Second, we need to prove that the minimum value of the objective function of the program is nonnegative.

Lemma D.4 (Feasibility of ($\mathrm{P}'_{\mathrm{NE}}$) and global optimum)**.** *The nonlinear program* ($\mathrm{P}'_{\mathrm{NE}}$) *is feasible, has a nonnegative objective value, and its global minimum is equal to* 0*.*

Proof. Analogously to the finite-horizon case, for the feasibility of the nonlinear program, we invoke the theorem of the existence of a Nash equilibrium. We let a NE product policy, $\boldsymbol{\pi}^\star$, and a vector $\boldsymbol{w}^\star \in \mathbb{R}^{n\times S}$ such that $w_k^\star(s) = V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s),\ \forall k \in [n]\times\mathcal{S}$.

By Lemma D.3, we know that $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ satisfies all the constraints of (P_{NE}). Additionaly, because $\boldsymbol{\pi}^\star$ is a NE, $V_{k,h}^{\boldsymbol{\pi}^\star}(s_1) = V_{k,h}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1)$ for all $k \in [n]$. Observing that,

$$w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H}\left(\prod_{\tau=1}^{h}\mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star)\right)\boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) = V_{k,h}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1) - V_{k,h}^{\boldsymbol{\pi}^\star}(s_1) = 0,$$

concludes the argument that a NE attains an objective value equal to 0.

Continuing, we observe that due to (4.7) the objective function can be equivalently rewritten as,

$$\begin{aligned}
&\sum_{k\in[n]}\left(w_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H}\left(\prod_{\tau=1}^{h}\mathbb{P}_\tau(\boldsymbol{\pi}_\tau)\right)\boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h)\right)\\
&\quad= \sum_{k\in[n]} w_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H}\left(\prod_{\tau=1}^{h}\mathbb{P}_\tau(\boldsymbol{\pi}_\tau)\right)\sum_{k\in[n]}\boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h)\\
&\quad= \sum_{k\in[n]} w_{k,1}(s_1).
\end{aligned}$$

Next, we focus on the inequality constraint

$$w_{k,h}(s) \geq r_{k,h}(s, a, \boldsymbol{\pi}_{-k,h}) + \mathbb{P}_h(s, a, \boldsymbol{\pi}_{-k,h})\boldsymbol{w}_{k,h+1}$$

which holds for all $s \in \mathcal{S}$, all players $k \in [n]$, all $a \in \mathcal{A}_k$, and all timesteps $h \in [H-1]$.

By summing over $a \in \mathcal{A}_k$ while multiplying each term with a corresponding coefficient $\pi_{k,h}(a|s)$, the display written in an equivalent element-wise vector inequality reads:

$$\boldsymbol{w}_{k,h} \geq \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) + \mathbb{P}_h(\boldsymbol{\pi}_h)\boldsymbol{w}_{k,h+1}.$$

Finally, after consecutively substituting $\boldsymbol{w}_{k,h+1}$ with the element-wise lesser term $\boldsymbol{r}_{k,h+1}(\boldsymbol{\pi}_{h+1}) + \mathbb{P}_{h+1}(\boldsymbol{\pi}_{h+1})\boldsymbol{w}_{k,h+2}$, we end up with the inequality:

$$\boldsymbol{w}_{k,1} \geq \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h). \tag{D.3}$$

Summing over k, it holds for the s_1-th entry of the inequality,

$$\sum_{k\in[n]} w_{k,1} \geq \sum_{k\in[n]} \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) = 0.$$

Where the equality holds due to the zero-sum property, (4.7). □

An approximate NE is an approximate global minimum. We show that an ϵ-approximate NE, $\boldsymbol{\pi}^\star$, achieves an $n\epsilon$-approximate global minimum of the program. Utilizing Lemma D.3, setting $w_k^\star(s_1) = V_{k,1}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1)$, and the definition of an ϵ-approximate NE we see that,

$$\begin{aligned}\sum_{k\in[n]} \left(w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) \right) &= \sum_{k\in[n]} \left(w_{k,1}^\star(s_1) - V_{k,1}^{\boldsymbol{\pi}^\star}(s_1) \right) \\ &\leq \sum_{k\in[n]} \epsilon = n\epsilon.\end{aligned}$$

Indeed, this means that $\boldsymbol{\pi}^\star, \boldsymbol{w}^\star$ is an $n\epsilon$-approximate global minimizer of ($\mathrm{P_{NE}}$).

An approximate global minimum is an approximate NE. For the opposite direction, we let a feasible ϵ-approximate global minimizer of the program ($\mathrm{P_{NE}}$), $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$. Because a global minimum of the program is equal to 0, an ϵ-approximate global optimum must be at most $\epsilon > 0$. We observe that for every $k \in [n]$,

$$w^\star_{k,1}(s_1) \geq \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}^\star_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}^\star_h), \tag{D.4}$$

which follows from induction on the inequality constraint over all h similar to (D.3).

Consequently, the assumption that

$$\epsilon \geq \sum_{k\in[n]} \left(w^\star_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}^\star_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}^\star_h) \right),$$

and Equation (D.4), yields the fact that

$$\begin{aligned}
\epsilon &\geq w^\star_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}^\star_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}^\star_h) \\
&\geq V_{k,1}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1) - V_{k,1}^{\boldsymbol{\pi}^\star}(s_1),
\end{aligned}$$

where the second inequality holds from the fact that $\boldsymbol{w}^\star$ is feasible for ($\mathrm{P_{BR}}$). The latter concludes the proof, as the display coincides with the definition of an ϵ-approximate NE.

D.3.2 Proof of Claim 4.1

Proof. The value function of s_1 for $h = 1$ of players 1 and 2 read:

$$\begin{aligned}
V_{1,1}^{\boldsymbol{\sigma}}(s_1) &= \boldsymbol{e}_{s_1}^\top \left(\boldsymbol{r}_1(\boldsymbol{\sigma}) + \mathbb{P}(\boldsymbol{\sigma}) \boldsymbol{r}_1(\boldsymbol{\sigma}) \right) \\
&= -\frac{9\sigma(a_1, b_1|s_1)}{20} + \frac{\sigma(a_1, b_2|s_1)}{20} + \frac{(1 - \sigma(a_1, b_1|s_1))\left(\sigma(a_1, b_1|s_2) + \sigma(a_1, b_2|s_2)\right)}{20},
\end{aligned}$$

and,

$$V_{2,1}^{\boldsymbol{\sigma}}(s_1) = \boldsymbol{e}_{s_1}^{\top}\left(\boldsymbol{r}_2(\boldsymbol{\sigma}) + \mathbb{P}(\boldsymbol{\sigma})\boldsymbol{r}_2(\boldsymbol{\sigma})\right)$$
$$= -\frac{9\sigma(a_1, b_1|s_1)}{20} + \frac{\sigma(a_2, b_2|s_1)}{20} + \frac{\left(1 - \sigma(a_1, b_1|s_1)\right)\left(\sigma(a_1, b_1|s_2) + \sigma(a_2, b_1|s_2)\right)}{20}.$$

We are indifferent to the corresponding value function of player 3 as they only have one available action per state and hence, cannot affect their rewards. For the joint policy $\boldsymbol{\sigma}$, the corresponding value functions of both players 1 and 2 are $V_{1,1}^{\boldsymbol{\sigma}}(s_1) = V_{2,1}^{\boldsymbol{\sigma}}(s_1) = \frac{1}{20}$.

Deviations. We will now prove that no deviation of player 1 manages to accumulate a reward greater than $\frac{1}{20}$. The same follows for player 2 due to symmetry.

When a player deviates unilaterally from a joint policy, they experience a single agent Markov decision process (MDP). It is well-known that MDPs always have a deterministic optimal policy. As such, it suffices to check whether $V_{1,1}^{\boldsymbol{\pi}_1, \boldsymbol{\sigma}_{-1}}(s_1)$ is greater than $\frac{1}{20}$ for any of the four possible deterministic policies:

- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}$,
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 1 & 0 \end{pmatrix}$, $\boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}$,
- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}$,
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 0 & 1 \end{pmatrix}$, $\boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

Finally, the value function of any deviation $\boldsymbol{\pi}_1'$ writes,

$$V_{1,1}^{\boldsymbol{\pi}_1' \times \boldsymbol{\sigma}_{-1}}(s_1) = -\frac{\pi_1'(a_1|s_1)}{5} - \frac{\pi_1'(a_1|s_2)\left(\pi_1'(a_1|s_1) - 2\right)}{40}.$$

We can now check that for all deterministic policies $V_{1,1}^{\boldsymbol{\pi}_1' \times \boldsymbol{\sigma}_{-1}}(s_1) \leq \frac{1}{20}$. By symmetry, it follows that $V_{2,1}^{\boldsymbol{\pi}_2' \times \boldsymbol{\sigma}_{-2}}(s_1) \leq \frac{1}{20}$ and as such $\boldsymbol{\sigma}$ is indeed a CCE. □

D.3.3 Proof of Claim 4.2

Proof. In general, the value functions of each player 1 and 2 are:

$$V_{1,1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2}(s_1) = -\frac{\pi_1(a_1|s_1)\pi_2(b_1|s_1)}{2} + \frac{\pi_1(a_1|s_1)}{20} - \frac{\pi_1(a_1|s_2)\left(\pi_1(a_1|s_1)\pi_2(b_1|s_1)-1\right)}{20},$$

and

$$V_{2,1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2}(s_1) = -\frac{\pi_1(a_1|s_1)\pi_2(b_1|s_1)}{2} + \frac{\pi_1(b_1|s_1)}{20} - \frac{\pi_1(b_1|s_2)\left(\pi_1(a_1|s_1)\pi_2(b_1|s_1)-1\right)}{20}.$$

Plugging in $\boldsymbol{\pi}_1^\sigma, \boldsymbol{\pi}_2^\sigma$ yields $V_{1,1}^{\boldsymbol{\pi}_1^\sigma\times\boldsymbol{\pi}_2^\sigma}(s_1) = V_{2,1}^{\boldsymbol{\pi}_1^\sigma\times\boldsymbol{\pi}_2^\sigma}(s_1) = -\frac{13}{160}$. But, if player 1 deviates to say $\pi_1'(s_1) = \pi_1'(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}$, they get a value equal to 0 which is clearly greater than $-\frac{13}{160}$. Hence, $\boldsymbol{\pi}_1^\sigma \times \boldsymbol{\pi}_2^\sigma$ is not a NE. □

D.3.4 Proof of Theorem 4.12

Proof. The proof follows from the game of Example 4.1, and Claims 4.1 and 4.2. □

D.3.5 Proofs for Infinite-Horizon Zero-Sum Polymatrix Markov Games

In this section we will explicitly state definitions, theorems and proofs relating to the infinite-horizon discounted zero-sum polymatrix Markov games.

D.3.5.1 Definitions of equilibria for the infinite-horizon

Let us restate the definition specifically for infinite-horizon Markov games. They are defined as a tuple $\Gamma(H, \mathcal{S}, \{\mathcal{A}_k\}_{k\in[n]}, \mathbb{P}, \{r_k\}_{k\in[n]}, \gamma, \boldsymbol{\rho})$.

- $H = \infty$ denotes the *time horizon*
- $\mathcal{S}$, with cardinality $S \coloneqq |\mathcal{S}|$, stands for the state space,
- $\{\mathcal{A}_k\}_{k\in[n]}$ is the collection of every player's action space, while $\mathcal{A} \coloneqq \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the *joint action space*; further, an element of that set —a joint action— is generally noted as $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathcal{A}$,
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function,
- $r_k : \mathcal{S}, \mathcal{A} \to [-1, 1]$ yields the reward of player k at a given state and joint action,
- a discount factor $0 < \gamma < 1$,
- an initial state distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$.

Policies and value functions. In infinite-horizon Markov games policies can still be distinguished in two main ways, *Markovian/non-Markovian* and *stationary/nonstationary*. Moreover, a joint policy can be a *correlated* policy or a *product* policy.

Markovian policies attribute a probability over the simplex of actions solely depending on the running state s of the game. On the other hand, *non-Markovian* policies attribute a probability over the simplex of actions that depends on any subset of the history of the game. *I.e.*, they can depend on any sub-sequence of actions and states up until the running timestep of the horizon.

Stationary policies are those that will attribute the same probability distribution over the simplex of actions for every timestep of the horizon. *Nonstationary* policies, on the contrary can change depending on the timestep of the horizon.

A joint Markovian stationary policy $\boldsymbol{\sigma}$ is said to be *correlated* when for every state $s \in \mathcal{S}$, attributes a probability distribution over the simplex of joint actions $\mathcal{A}$ for all players, *i.e.*, $\boldsymbol{\sigma}(s) \in \Delta(\mathcal{A})$. A Markovian stationary policy $\boldsymbol{\pi}$ is said to be a *product* policy when for every $s \in \mathcal{S}$, $\boldsymbol{\pi}(s) \in \prod_{k=1}^{n} \Delta(\mathcal{A}_k)$. It is rather easy to define *correlated/product* policies for the case of non-Markovian and nonstationary policies.

Given a Markovian stationary policy $\boldsymbol{\pi}$, the value function for an infinite-horizon discounted game is defined as,

$$V_k^{\boldsymbol{\pi}}(s_1) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{h=1}^{H} \gamma^{h-1} r_{k,h}(s_h, \boldsymbol{a}_h) \big| s_1\right] = \boldsymbol{e}_{s_1}^{\top} \sum_{h=1}^{H} \left(\gamma^{h-1} \prod_{\tau=1}^{h} \mathbb{P}_{\tau}(\boldsymbol{\pi}_{\tau})\right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h).$$

It is possible to express the value function of each player k in the following way,

$$V_k^{\boldsymbol{\pi}}(s_1) = \boldsymbol{e}_{s_1}^{\top} \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}(\boldsymbol{\pi}).$$

Where $\mathbf{I}$ is the identity matrix of appropriate dimensions. Also, when the initial state is drawn from the initial state distribution, we denote, the value function reads $V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) = \boldsymbol{\rho}^{\top} \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}(\boldsymbol{\pi})$.

Best-response policies. Given an arbitrary joint policy $\boldsymbol{\sigma}$ (which can be either a correlated or product policy), a best-response policy of a player k is defined to be $\boldsymbol{\pi}_k^{\dagger} \in \Delta(\mathcal{A}_k)^S$ such that $\boldsymbol{\pi}_k^{\dagger} \in \arg\max_{\boldsymbol{\pi}_k'} V_k^{\boldsymbol{\pi}_k' \times \boldsymbol{\sigma}_{-k}}(s)$. Also, we will denote $V_k^{\dagger, \boldsymbol{\sigma}_{-k}}(s) = \max_{\boldsymbol{\pi}_k'} V_k^{\boldsymbol{\pi}_k', \boldsymbol{\sigma}_{-k}}(s)$. It is rather straightforward to see that the problem of computing a best-response to a given policy is equivalent to solving a single-agent MDP problem.

Notions of equilibria. Now that best-response policies have been defined, it is straightforward to define the different notions of equilibria. First, we define the notion of a coarse-correlated equilibrium.

Definition D.1 (CCE—infinite-horizon)**.** *A joint (potentially correlated) policy $\boldsymbol{\sigma} \in \Delta(\mathcal{A})^S$ is an ϵ-approximate coarse-correlated equilibrium if it holds that for an ϵ,*

$$V_k^{\dagger,\boldsymbol{\sigma}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

Second, we define the notion of a Nash equilibrium. The main difference of the definition of the coarse-correlated equilibrium, is the fact that a NE Markovian stationary policy is a *product policy.*

Definition D.2 (NE—infinite-horizon)**.** *A joint (potentially correlated) policy $\boldsymbol{\pi} \in \prod_{k\in[n]} \Delta(\mathcal{A}_k)^S$ is an ϵ-approximate coarse-correlated equilibrium if it holds that for an ϵ,*

$$V_k^{\dagger,\boldsymbol{\pi}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

As it is folklore by now, infinite-horizon discounted Markov games have a stationary Markovian Nash equilibrium.

D.3.6 Main results for Infinite-Horizon MGs

The workhorse of our arguments in the following results is still the following nonlinear program with variables $\boldsymbol{\pi}, \boldsymbol{w}$,

$$
\begin{aligned}
&\min \sum_{k\in[n]} \boldsymbol{\rho}^\top \left(\boldsymbol{w}_k - (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\boldsymbol{r}_k(\boldsymbol{\pi})\right)\\
&\text{s.t. } w_k(s) \geq r_k(s,a,\boldsymbol{\pi}_{-k}) + \gamma\,\mathbb{P}(s,a,\boldsymbol{\pi}_{-k})\boldsymbol{w}_k,\\
&\qquad \forall s\in\mathcal{S}, \forall k\in[n], \forall a\in\mathcal{A}_k;\\
&\qquad \boldsymbol{\pi}_k(s)\in\Delta(\mathcal{A}_k),\\
&\qquad \forall s\in\mathcal{S}, \forall k\in[n], \forall a\in\mathcal{A}_k.
\end{aligned}
\tag{$\mathrm{P}'_{\mathrm{NE}}$}
$$

As we will prove, approximate NE's correspond to approximate global minima of ($\mathrm{P}'_{\mathrm{NE}}$) and vice-versa. Before that, we need some intermediate lemmas. The first lemma we prove is about the best-response program.

The best-response program. Even for the infinite-horizon, we can define a linear program for the best-responses of all players. That program is the following, with variables $\boldsymbol{w}$,

$$
\begin{aligned}
&\min \sum_{k\in[n]} \boldsymbol{\rho}^\top \left(\boldsymbol{w}_k - (\mathbf{I} - \gamma\,\mathbb{P}(\hat{\boldsymbol{\sigma}}))^{-1}\boldsymbol{r}_k(\hat{\boldsymbol{\sigma}})\right)\\
&\text{s.t. } w_k(s) \geq r_k(s,a,\hat{\boldsymbol{\sigma}}_{-k}) + \mathbb{P}(s,a,\hat{\boldsymbol{\sigma}}_{-k})\boldsymbol{w}_k,\\
&\qquad \forall s\in\mathcal{S}, \forall k\in[n], \forall a\in\mathcal{A}_k.
\end{aligned}
\tag{$\mathrm{P}'_{\mathrm{BR}}$}
$$

Lemma D.5 (Best-response LP—infinite-horizon)**.** *Let a (possibly correlated) joint policy $\hat{\boldsymbol{\sigma}}$. Consider the linear program ($\mathrm{P}'_{\mathrm{BR}}$). The optimal solution $\boldsymbol{w}^\dagger$ of the program is unique and corresponds to the value function of each player $k\in[n]$ when player k best-responds to $\hat{\boldsymbol{\sigma}}$.*

Proof. We observe that the program is separable to n independent linear programs, each with variables $\boldsymbol{w}_k \in \mathbb{R}^n$,

$$\begin{aligned}
&\min\ \boldsymbol{\rho}^\top \boldsymbol{w}_k \\
&\text{s.t. } w_k(s) \geq r_k(s, a, \hat{\boldsymbol{\sigma}}_{-k}) + \gamma\, \mathbb{P}(s, a, \hat{\boldsymbol{\sigma}}_{-k}) \boldsymbol{w}_k, \\
&\qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}_k.
\end{aligned}$$

Each of these linear programs describes the problem of a single agent MDP —that agent being k. It follows that the optimal $\boldsymbol{w}_k^\dagger$ for every program is unique (each program corresponds to a set of Bellman optimality equations). □

Properties of the NE program. Second, we need to prove that the minimum value of the objective function of the program is nonnegative.

Lemma D.6 (Feasibility of ($\mathrm{P}'_{\mathrm{NE}}$) and global optimum)**.** *The nonlinear program* ($\mathrm{P}'_{\mathrm{NE}}$) *is feasible, has a nonnegative objective value, and its global minimum is equal to* 0*.*

Proof. For the feasibility of the nonlinear program, we invoke the theorem of the existence of a Nash equilibrium. *i.e.*, let a NE product policy, $\boldsymbol{\pi}^\star$, and a vector $\boldsymbol{w}^\star \in \mathbb{R}^{n \times H \times S}$ such that $w^\star_{k,s}(s) = V_k^{\dagger, \boldsymbol{\pi}^\star_{-k}}(s),\ \forall k \in [n] \times \mathcal{S}$.

By Lemma D.5, we know that $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ satisfies all the constraints of ($\mathrm{P}'_{\mathrm{NE}}$). Additionally, because $\boldsymbol{\pi}^\star$ is a NE, $V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) = V_k^{\dagger, \boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho})$ for all $k \in [n]$. Observing that,

$$\boldsymbol{\rho}^\top \left(\boldsymbol{w}_k^\star - (\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^\star))^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}^\star) \right) = V_k^{\dagger, \boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) = 0,$$

concludes the argument that a NE attains an objective value equal to 0.

Continuing, we observe that due to (4.7) the objective function can be equivalently rewritten

as,

$$\begin{aligned}
&\sum_{k\in[n]} \left(\boldsymbol{\rho}^\top \boldsymbol{w}_k - \boldsymbol{\rho}^\top (\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}))^{-1} \boldsymbol{r}_k(\boldsymbol{\pi})\right) \\
&\qquad = \sum_{k\in[n]} \boldsymbol{\rho}^\top \boldsymbol{w}_k - \boldsymbol{\rho}^\top (\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}))^{-1} \sum_{k\in[n]} \boldsymbol{r}_k(\boldsymbol{\pi}_h) \\
&\qquad = \sum_{k\in[n]} \boldsymbol{\rho}^\top \boldsymbol{w}_k.
\end{aligned}$$

Next, we focus on the inequality constraint

$$w_k(s) \geq r_k(s, a, \boldsymbol{\pi}_{-k}) + \gamma\, \mathbb{P}(s, a, \boldsymbol{\pi}_{-k}) \boldsymbol{w}_k$$

which holds for all $s \in \mathcal{S}$, all players $k \in [n]$, and all $a \in \mathcal{A}_k$.

By summing over $a \in \mathcal{A}_k$ while multiplying each term with a corresponding coefficient $\pi_k(a|s)$, the display written in an equivalent element-wise vector inequality reads:

$$\boldsymbol{w}_k \geq \boldsymbol{r}_{k,h}(\boldsymbol{\pi}) + \gamma\, \mathbb{P}(\boldsymbol{\pi}) \boldsymbol{w}_k.$$

Finally, after consecutively substituting $\boldsymbol{w}_k$ with the element-wise lesser term $\boldsymbol{r}_k(\boldsymbol{\pi}) + \gamma\, \mathbb{P}_(\, \boldsymbol{\pi}) \boldsymbol{w}_k$, we end up with the inequality:

$$\boldsymbol{w}_k \geq \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}). \tag{D.7}$$

We note that $\mathbf{I} + \gamma\, \mathbb{P}(\boldsymbol{\pi}) + \gamma^2\, \mathbb{P}^2(\boldsymbol{\pi}) + \cdots = \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1}$.

Summing over k, it holds for the s_1-th entry of the inequality,

$$\sum_{k\in[n]} \boldsymbol{w}_k \geq \sum_{k\in[n]} \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}) = \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \sum_{k\in[n]} \boldsymbol{r}_k(\boldsymbol{\pi}) = 0.$$

Where the equality holds due to the zero-sum property, (4.7). □

Theorem D.4 (NE and global optima of ($\mathrm{P}'_{\mathrm{NE}}$)—infinite-horizon)**.** *If* $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ *yields an* ϵ*-approximate global minimum of* ($\mathrm{P}'_{\mathrm{NE}}$)*, then* $\boldsymbol{\pi}^\star$ *is an* $n\epsilon$*-approximate NE of the infinite-horizon zero-sum polymatrix switching controller MG,* Γ*. Conversely, if* $\boldsymbol{\pi}^\star$ *is an* ϵ*-approximate NE of the MG* Γ *with corresponding value function vector* $\boldsymbol{w}^\star$ *such that* $w_k^\star(s) = V_k^{\boldsymbol{\pi}^\star}(s) \forall (k,s) \in [n] \times \mathcal{S}$*, then* $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ *attains an* ϵ*-approximate global minimum of* ($\mathrm{P}'_{\mathrm{NE}}$)*.*

Proof.

An approximate NE is an approximate global minimum. We show that an ϵ-approximate NE, $\boldsymbol{\pi}^\star$, achieves an $n\epsilon$-approximate global minimum of the program. Utilizing Lemma D.5 by setting $\boldsymbol{w}_k^\star = \mathbf{V}^{\dagger, \boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho})$, feasibility , and the definition of an ϵ-approximate NE we see that,

$$\begin{aligned}\sum_{k\in[n]} \left(\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}^\star)\right) &= \sum_{k\in[n]} \left(\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho})\right) \\ &\leq \sum_{k\in[n]} \epsilon = n\epsilon.\end{aligned}$$

Indeed, this means that $\boldsymbol{\pi}^\star, \boldsymbol{w}^\star$ is an $n\epsilon$-approximate global minimizer of ($\mathrm{P}'_{\mathrm{NE}}$).

An approximate global minimum is an approximate NE. For this direction, we let a feasible ϵ-approximate global minimizer of the program ($\mathrm{P}'_{\mathrm{NE}}$), $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$. Because a global minimum of the program is equal to 0, an ϵ-approximate global optimum must be at most $\epsilon > 0$. We observe that for every $k \in [n]$,

$$\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star \geq \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}^\star), \tag{D.8}$$

which follows from induction on the inequality constraint (D.7).

Consequently, the assumption that

$$\epsilon \geq \boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}^\star)$$

and Equation (D.8), yields the fact that

$$\begin{aligned}\epsilon &\geq \boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1} \boldsymbol{r}_k(\boldsymbol{\pi}^\star) \\ &\geq V_k^{\dagger, \boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}),\end{aligned}$$

where the second inequality holds from the fact that $\boldsymbol{w}^\star$ is also feasible for ($\mathrm{P}'_{\mathrm{BR}}$). The latter concludes the proof, as the display coincides with the definition of an ϵ-approximate NE. □

Theorem D.5 (CCE collapse to NE in polymatrix MG—infinite-horizon)**.** *Let a zero-sum polymatrix switching-control Markov game, i.e., a Markov game for which Assumptions 4.1 and 4.2 hold. Further, let an ϵ-approximate CCE of that game $\boldsymbol{\sigma}$. Then, the marginal product policy $\boldsymbol{\pi}^\sigma$, with $\boldsymbol{\pi}_k^{\boldsymbol{\sigma}}(a|s) = \sum_{\boldsymbol{a}_{-k} \in \mathcal{A}_{-k}} \boldsymbol{\sigma}(a, \boldsymbol{a}_{-k}),\ \forall k \in [n]$ is an $n\epsilon$-approximate NE.*

Proof. Let an ϵ-approximate CCE policy, $\boldsymbol{\sigma}$, of game Γ. Moreover, let the best-response value-vectors of each agent k to joint policy $\boldsymbol{\sigma}_{-k}$, $\boldsymbol{w}_k^\dagger$.

Now, we observe that due to Assumption 4.1,

$$\begin{aligned}w_k^\dagger(s) &\geq r_k(s, a, \boldsymbol{\sigma}_{-k}) + \gamma\, \mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k}) \boldsymbol{w}_k^\dagger \\ &= \sum_{j \in \mathrm{adj}(k)} r_{(k,j),h}(s, a, \boldsymbol{\pi}_j^{\boldsymbol{\sigma}}) + \gamma\, \mathbb{P}(s, a, \boldsymbol{\sigma}_{-k}) \boldsymbol{w}_k^\dagger.\end{aligned}$$

Further, due to Assumption 4.2,

$$\mathbb{P}(s, a, \boldsymbol{\sigma}_{-k}) \boldsymbol{w}_k^\dagger = \mathbb{P}(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}}_{\mathrm{argctrlr}(s)}) \boldsymbol{w}_k^\dagger,$$

or,

$$\mathbb{P}(s, a, \boldsymbol{\sigma}_{-k})\boldsymbol{w}_k^\dagger = \mathbb{P}(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}})\boldsymbol{w}_k^\dagger.$$

Putting these pieces together, we reach the conclusion that $(\boldsymbol{\pi}^\sigma, \boldsymbol{w}^\dagger)$ is feasible for the non-linear program $(\mathrm{P}'_{\mathrm{NE}})$.

What is left is to prove that it is also an ϵ-approximate global minimum. Indeed, if $\sum_k \boldsymbol{\rho}^\top \boldsymbol{w}_k^\dagger \leq \epsilon$ (by assumption of an ϵ-approximate CCE), then the objective function of $(\mathrm{P}'_{\mathrm{NE}})$ will attain an ϵ-approximate global minimum. In turn, due to Theorem D.4 the latter implies that $\boldsymbol{\pi}^{\boldsymbol{\sigma}}$ is an $n\epsilon$-approximate NE. □

D.3.6.1 No equilibrium collapse with more than one controllers per-state

Example D.1. *We consider the following* 3*-player Markov game that takes place for a time horizon* $H = 3$*. There exist three states,* s_1, s_2*, and* s_3 *and the game starts at state* s_1*. Player* 3 *has a single action in every state, while players* 1 *and* 2 *have two available actions* $\{a_1, a_2\}$ *and* $\{b_1, b_2\}$ *respectively in every state. The initial state distribution* $\boldsymbol{\rho}$ *is the uniform probability distribution over* $\mathcal{S}$*.*

Reward functions. *If player* 1 *(respectively, player* 2*) takes action* a_1 *(resp.,* b_1*), in either of the states* s_1 *or* s_2*, they get a reward equal to* $\frac{1}{20}$*. In state* s_3*, both players get a reward equal to* $-\frac{1}{2}$ *regardless of the action they select. Player* 3 *always gets a reward that is equal to the negative sum of the reward of the other two players. This way, the* zero-sum polymatrix property *of the game is ensured (Assumption 4.1).*

Transition probabilities. *If players* 1 *and* 2 *select the joint action* (a_1, b_1) *in state* s_1*, the game will transition to state* s_2*. In any other case, it will transition to state* s_3*. The*

converse happens if in state s_2 they take joint action (a_1, b_1); the game will transition to state s_3. For any other joint action, it will transition to state s_1. From state s_3, the game transition to state s_1 or s_2 uniformally at random.

At this point, it is important to notice that two players control the transition probability from one state to another. In other words, Assumption 4.2 does not hold.

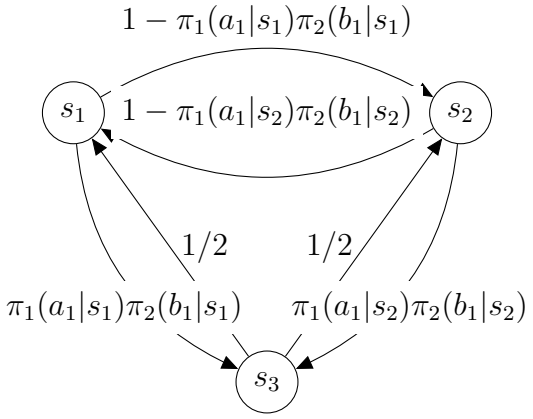

Figure D.1: A graph of the state space with transition probabilities parametrized with respect to the policy of each player.

Next, we consider the joint policy $\boldsymbol{\sigma}$,

$$\boldsymbol{\sigma}(s_1) = \boldsymbol{\sigma}(s_2) = \begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} \end{array}.$$

Claim D.3. *The joint policy $\boldsymbol{\sigma}$ that assigns probability $\frac{1}{2}$ to the joint actions (a_1, b_2) and (a_2, b_1) in both states s_1, s_2 is a CCE and $V_1^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) = V_2^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) = -\frac{1}{10}$.*

Proof.

$$
\begin{aligned}
V_1^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) &= \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\sigma})\right)^{-1} \boldsymbol{r}_1(\boldsymbol{\sigma}) \\
&= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{9}{5} & \frac{6}{5} & 0 \\ \frac{6}{5} & \frac{9}{5} & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{40} \\ \frac{1}{40} \\ -\frac{1}{2} \end{pmatrix} \\
&= -\frac{1}{10}.
\end{aligned}
$$

We check every deviation,

- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{2}{5},$
- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{1}{6},$
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 1 & 0 \end{pmatrix},\ \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{5}{16},$
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 0 & 1 \end{pmatrix},\ \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{5}{16}.$

For every such deviation the value of player 1 is smaller than $-\frac{1}{10}$. For player 2, the same follows by symmetry. Hence, $\boldsymbol{\sigma}$ is indeed a CCE.

□

Yet, the marginalized product policy of $\boldsymbol{\sigma}$ which we note as $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ does not constitute a

NE. The components of this policy are,

$$\begin{cases} \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}(s_1) = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}(s_2) = \begin{array}{c} \begin{matrix} a_1 & a_2 \end{matrix} \\ \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \end{array}, \\ \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}(s_1) = \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}(s_2) = \begin{array}{c} \begin{matrix} b_1 & b_2 \end{matrix} \\ \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \end{array}. \end{cases}$$

I.e., the product policy $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ selects any of the two actions of each player in states s_1, s_2 independently and uniformally at random. With the following claim, it can be concluded that in general when more than one player control the transition the set of equilibria do not collapse.

Claim D.4. *The product policy $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ is not a NE.*

Proof. For $\boldsymbol{\pi}^{\sigma} = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ we get,

$$\begin{aligned} V_1^{\boldsymbol{\pi}^{\sigma}} &= \boldsymbol{\rho}^{\top} \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi}^{\boldsymbol{\sigma}})\right)^{-1} \boldsymbol{r}_1(\boldsymbol{\pi}^{\boldsymbol{\sigma}}) \\ &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{34}{21} & \frac{20}{21} & \frac{3}{7} \\ \frac{20}{21} & \frac{34}{21} & \frac{3}{7} \\ \frac{6}{7} & \frac{6}{7} & \frac{9}{7} \end{pmatrix} \begin{pmatrix} \frac{1}{40} \\ \frac{1}{40} \\ -\frac{1}{2} \end{pmatrix} \\ &= -\frac{3}{10}. \end{aligned}$$

But, for the deviation $\boldsymbol{\pi}_1(a_1|s_1) = \boldsymbol{\pi}_1(a_1|s_2) = 0$, the value funciton of player 1, is equal to $-\frac{1}{6}$. Hence, $\boldsymbol{\pi}^{\sigma}$ is not a NE. □

In conclusion, Assumption 4.1 does not suffice to ensure equilibrium collapse.

Theorem D.6 (No collapse—infinite-horizon)**.** *There exists a zero-sum polymatrix Markov game (Assumption 4.2 is not satisfied) that has a CCE which does not collapse to a NE.*

Proof. The proof follows from the game of Example D.1, and Claims D.3 and D.4. □

D.4 Proof of Extendibility to Nash Equilibria

In this section, we demonstrate how a nearly stationary point $\hat{\boldsymbol{x}}$ of $\phi(\cdot) \coloneqq \max_{\boldsymbol{y} \in \mathcal{Y}} V_{\boldsymbol{\rho}}(\cdot, \boldsymbol{y})$, 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 and Vrieze, 2012, Chapter 3.9). But, unlike the program in (Filar and 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 ϕ, 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, (P_{NE}) —that incorporates an additional quadratic term to the objective function. This allows us to show that the proximal point $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ 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 (P_{NE}). 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 D.4.1 we consider (P_{NE}), 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 D.4.2 we show that (P_{NE}) attains a global optimum at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ (Lemma D.8), where $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ and $\tilde{\boldsymbol{v}}$ is the unique value vector associated with $\tilde{\boldsymbol{x}}$ (Proposition D.2).

(iii) In **??** D.4.3.1 we show that any feasible point of (P_{NE}) satisfies the Arrow-Hurwicz-Uzawa constraint qualification (Lemma D.9). In turn, this implies the existence of nonnegative Lagrange multipliers at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ satisfying the KKT conditions (Corollary D.2).

(iv) In Appendix D.4.4 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 D.10 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 D.11) 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 D.12 and Theorem D.7, 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.

D.4.1 The Quadratic NLP

In this subsection, we describe in more detail the nonlinear program (P_{NE}) we introduced earlier in Section 4.4.4. 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 and Vrieze, 2012, Chapter 3)), introduced below.

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

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

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

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.

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

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, ($\text{NLP}_{\mathcal{G}}$) 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 ($\text{NLP}_{\mathcal{G}}$) 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 $\tilde{\boldsymbol{x}} \coloneqq \text{prox}_{\phi/2\ell}(\hat{\boldsymbol{x}})$ 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 4.2. 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 4.4.4, but we include it below for the convenience of the reader.

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

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

Lemma D.7. *The program* ($\mathrm{P_{NE}}$) *is feasible.*

Proof. Let $\boldsymbol{x} \in \mathcal{X}$ be any directly parameterized policy for the team and $\boldsymbol{v} \coloneqq \frac{1}{1-\gamma}\mathbf{1}$, 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

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

□

D.4.2 The Global Minimum of ($\mathrm{P_{NE}}$)

Here we demonstrate that ($\mathrm{P_{NE}}$) attains a global minimum under $\boldsymbol{x} = \tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$. 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 ϕ 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 ($\mathrm{P_{NE}}$) 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 ℓ-strongly convex, which means that it has a unique minimizer, namely $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$. In turn, this implies that the objective function of ($\mathrm{P_{NE}}$) 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 ($\mathrm{P_{NE}}$).

First, we relate the optimal vector $\boldsymbol{v}$ that arises by fixing $\boldsymbol{x}$ in ($\mathrm{P_{NE}}$) and the function $\phi(\boldsymbol{x})$:

Proposition D.2. *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* ($\mathrm{P_{NE}}$). *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 ($\mathrm{P_{NE}}$) 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 $\boldsymbol{\lambda} \in \mathbb{R}^{S \times B}$:

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

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,

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

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

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

Further, for any such $\boldsymbol{y} \in \mathcal{Y}$ it holds that $d_{\boldsymbol{\rho}}^{\boldsymbol{x},\boldsymbol{y}}(s,b) = \lambda(s,b), \ \forall (s,b) \in \mathcal{S} \times \mathcal{B}$, 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.10 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 D.8. *Let $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$, and $\tilde{\boldsymbol{v}}$ be the unique minimizer for* ($\mathrm{P_{NE}}$) *under a fixed $\boldsymbol{x} = \tilde{\boldsymbol{x}}$. Then, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is a global minimum of* ($\mathrm{P_{NE}}$).

Proof. Consider a fixed $\boldsymbol{x} \in \mathcal{X}$. By Proposition D.2, we know that there is a unique optimal vector $\boldsymbol{v}^\star$ in ($\mathrm{P_{NE}}$), which also satisfies the equality

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

Now let us consider the function $\Psi(\boldsymbol{w}) \coloneqq \phi(\boldsymbol{w}) + \ell\|\boldsymbol{w} - \hat{\boldsymbol{x}}\|^2$. Ψ is ℓ-strongly convex and its unique minimum value is attained at $\tilde{\boldsymbol{x}} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\hat{\boldsymbol{x}})$ (Corollary B.1). By (D.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 ($\mathrm{P_{NE}}$) when we set $\boldsymbol{x} = \tilde{\boldsymbol{x}}$, which is feasible for ($\mathrm{P_{NE}}$) (see Lemma D.7 and Proposition D.2). □

D.4.3 KKT Conditions for a Minimizer of Equation ($\mathrm{P_{NE}}$)

As we have shown in the previous subsection, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ is a minimum of the program ($\mathrm{P_{NE}}$). 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 of the constrained minimization problem associated with ($\mathrm{P_{NE}}$):

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

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 ($\mathrm{P_{NE}}$). 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 $(Q1)$, 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 $\overline{k} \in [n], (\overline{s}, \overline{a}) \in \mathcal{S} \times \mathcal{A}_k$,

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

- 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 D.5. *Consider the Lagrangian function $\mathcal{L}$ of* ($\mathrm{P_{NE}}$)*, as introduced in* (D.14)*. Then, for any $\overline{s} \in \mathcal{S}$, the partial derivative of $\mathcal{L}$ with respect to $v(\overline{s})$ reads*

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

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

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

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

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

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

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

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

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

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

D.4.3.1 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 ($\mathrm{P_{NE}}$), there exist (non-negative) Lagrange multipliers that jointly satisfy the KKT conditions. We will first argue in Lemma D.9 below that any feasible point of ($\mathrm{P_{NE}}$) satisfies the Arrow-Hurwicz-Uzawa constraint qualification. Then, we will leverage Corollary A.1 to show that any local minimizer of ($\mathrm{P_{NE}}$)—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 D.9. *Let $(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{X} \times \mathbb{R}^S$ be any feasible point of* ($\mathrm{P_{NE}}$). *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 $\boldsymbol{w}_{\boldsymbol{x}} = \mathbf{0}$. 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 $\boldsymbol{w} = (\mathbf{0}, \boldsymbol{w}_{\boldsymbol{v}})$, 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 $\boldsymbol{w}_{\boldsymbol{v}} \coloneqq -\boldsymbol{v}$. 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 ($\mathrm{P_{NE}}$)—attains nonnegative Lagrange multipliers that satisfy the KKT conditions.

Corollary D.2. *For any local minimum* $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X} \times \mathbb{R}^S$ *of* ($\mathrm{P_{NE}}$)*, 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

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

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

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

D.4.3.2 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 D.12 for controlling the approximation error.

Proposition D.3. *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* (D.19)*. Then, it holds that* $\sum_{b \in \mathcal{B}} \tilde{\lambda}(s,b) > 0$*, for any* $s \in \mathcal{S}$*. Further, if*

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

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 (D.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 $\boldsymbol{d} \in \mathbb{R}^S_{>0}$ such that

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

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

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

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

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

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

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

where (D.21) uses the definition of $\boldsymbol{d}$ given in (D.20); (D.22) follows from the definition of strategy $\boldsymbol{y}$ in the statement of the proposition; and (D.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 (D.23) can be compactly expressed as $\boldsymbol{\rho}^\top = \boldsymbol{d}^\top\left(\mathbf{I} - \gamma\,\mathbb{P}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\right)$ (recall the definition of matrix $\mathbb{P}$), in turn

implying that

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

We note that $(\mathbf{I} - \gamma\, \mathbb{P}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}))$ is invertible (Claim D.7). As a result, by virtue of Claim D.9 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 D.6. *Let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) \in \mathcal{X}\times\mathbb{R}^S$ be a local optimum of the* (P$_{\text{NE}}$)*, and $\{\tilde{\lambda}(s,b)\}, \{\tilde{\psi}(k,s)\}, \{\tilde{\omega}(k,s)\}$ be the associated Lagrange multipliers defined in* (D.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 (D.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

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

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

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

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

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

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

Further, we know from the complementary slackness condition (D.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 (D.24), and recalling that $\sum_{b\in\mathcal{B}}\tilde{\lambda}(s,b)>0$ for any $s \in \mathcal{S}$ (by Proposition D.3), leads to the desired conclusion. □

D.4.4 Efficient Extension to Nash Equilibria

This subsection completes the proof that an ϵ-near stationary point $\hat{\boldsymbol{x}}$ of ϕ 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 $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$:

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

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 ($\mathrm{P}_{\mathrm{NE}}$). 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 (D.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 (D.19b), while ($\mathrm{LP}_{\mathrm{adv}}$.1) is related to the first-order stationary condition (D.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 $(\mathrm{P}_{\mathrm{NE}})$. 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 D.10. *The linear program* $(\mathrm{LP}_{\mathrm{adv}})$ *with variables* $\boldsymbol{\lambda} \in \mathbb{R}^{S\times B}$ *is feasible.*

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

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

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 D.11. *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* ($\mathrm{P_{NE}}$)*. Then, $\tilde{\boldsymbol{\lambda}}$ satisfies all the constraints of* ($\mathrm{LP'_{adv}}$)*.*

Proof. First, ($\mathrm{LP'_{adv}}$.1) is satisfied by the first-order stationarity condition (D.19a); ($\mathrm{LP'_{adv}}$.3) is satisfied by the complementary slackness condition (D.19b); and ($\mathrm{LP'_{adv}}$.4) by the nonnegative of the Lagrange multipliers (D.19c). The rest of the proof is devoted to showing that $\tilde{\boldsymbol{\lambda}}$ also satisfies ($\mathrm{LP'_{adv}}$.2). To this end, we first recall that, by Claim D.6, 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 (D.16) we get that for any $k \in [n], (s,a) \in \mathcal{S}\times\mathcal{A}_k$,

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

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 (D.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

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

Thus,

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

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 (D.27) and the triangle inequality that

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

This inequality uses that $\|\tilde{\boldsymbol{x}}-\hat{\boldsymbol{x}}\|\leq\epsilon$; the fact that $\tilde{\lambda}(s,b)\leq\frac{1}{1-\gamma}$ ($\mathrm{LP}_{\mathrm{adv}}$.5); and the Lipschitz continuity bounds provided in Claim D.15:

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

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

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

Thus, $\tilde{\boldsymbol{\lambda}}$ satisfies ($\mathrm{LP}_{\mathrm{adv}}$.1). Finally, $\tilde{\boldsymbol{\lambda}}$ also satisfies ($\mathrm{LP}_{\mathrm{adv}}$.4) and ($\mathrm{LP}_{\mathrm{adv}}$.5), implied directly by Proposition D.3 and Claim D.13. □

Lemma D.12. *Let $\hat{\boldsymbol{x}}$ be an ϵ-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 ($\mathrm{LP}_{\mathrm{adv}}$) 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 ($\mathrm{LP}_{\mathrm{adv}}$), 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 (P_{NE}), 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

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

where we used Claims D.7, D.8 and D.11, and the notation $\boldsymbol{V}(\hat{\boldsymbol{x}}, \boldsymbol{y})$ to represent the value vector under $(\hat{\boldsymbol{x}}, \boldsymbol{y})$—recall (4.3). Moreover, given that $\boldsymbol{\lambda}$ is a feasible solution of (LP_{adv}), manipulating (LP_{adv}.3) yields that for any $(s, b) \in \mathcal{S} \times \mathcal{B}$,

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

$$\frac{1}{\sum_{b' \in \mathcal{B}} \lambda(s,b')} \lambda(s,b) \left(\left[r\left(s, \hat{\boldsymbol{x}}, b\right) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s' | s, \hat{\boldsymbol{x}}, b\right) \hat{v}(s') \right] - \hat{v}(s) \right) \geq -\frac{c_2 \epsilon}{\sum_{b' \in \mathcal{B}} \lambda(s,b')} \tag{D.29}$$

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

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

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

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

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

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

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

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.8 and D.11. Thus,

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

where (D.32) follows from Claim D.9; (D.33) follows from the feasibility constraint $\sum_{b \in \mathcal{B}} \lambda(s, b) \geq \rho(s)$; and (D.34) uses the definition of mismatch coefficient (Definition 2.6). As a result, combining (D.28) and (D.34), we conclude that for any $\boldsymbol{y} \in \mathcal{Y}$,

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

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 ($\mathrm{LP}_{\mathrm{adv}}$.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.11, 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

$$V_{\boldsymbol{\rho}}(\boldsymbol{x}_k,\hat{\boldsymbol{x}}_{-k}),\hat{\boldsymbol{y}})\geq\boldsymbol{\rho}^\top\hat{\boldsymbol{v}}-c_1DS\epsilon,\tag{D.36}$$

where we used Claim D.9 and Definition 2.6. Next, using ($\mathrm{LP}_{\mathrm{adv}}$.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

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

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

$$\begin{aligned}
\boldsymbol{r}\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\right)+\gamma\,\mathbb{P}\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\right)\hat{\boldsymbol{v}} &\leq \hat{\boldsymbol{v}}+B\boldsymbol{\xi}\\
\boldsymbol{V}\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\right) &\leq \hat{\boldsymbol{v}}+B\left(\mathbf{I}-\gamma\,\mathbb{P}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\right)^{-1}\boldsymbol{\xi} &\text{(D.37)}\\
V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}) &\leq \boldsymbol{\rho}^\top\hat{\boldsymbol{v}}+Bc_2\sum_{s\in\mathcal{S}}\frac{d_{\boldsymbol{\rho}}^{\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}}(s)}{\sum_{b\in\mathcal{B}}\lambda(s,b)}\epsilon &\text{(D.38)}\\
V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}) &\leq \boldsymbol{\rho}^\top\hat{\boldsymbol{v}}+c_2BSD\epsilon, &\text{(D.39)}
\end{aligned}$$

where (D.37) follows from Claim D.11; (D.38) follows from Claim D.9; and (D.39) follows from the fact that $\sum_{b\in\mathcal{B}}\lambda(s,b)\geq\rho(s)$ and Definition 2.6. As a result, combining (D.36) and (D.39) we conclude that

$$V_{\boldsymbol{\rho}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}) \leq V_{\boldsymbol{\rho}}\left((\boldsymbol{x}_k;\hat{\boldsymbol{x}}_{-k}),\hat{\boldsymbol{y}}\right)+c_2BSD\epsilon+c_1DS\epsilon. \tag{D.40}$$

□

We state the precise version of Lemma D.12 in Theorem D.7 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 ϵ-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 $\mathrm{LP}_{\mathrm{adv}}$, as introduced in ($\mathrm{LP}_{\mathrm{adv}}$). Then, any feasible solution $\boldsymbol{\lambda}\in\mathbb{R}^{S\times B}$ of ($\mathrm{LP}_{\mathrm{adv}}$) is used to determine the strategy of the adversary as follows.

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

Theorem D.7 (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 ϵ-nearly stationary point of $\phi(\boldsymbol{x})\coloneqq\max_{\boldsymbol{y}} V_{\boldsymbol{\rho}}(\cdot,\boldsymbol{y})$, where $V_{\boldsymbol{\rho}}$ is the value function of $\mathcal{G}$ (4.3). Then, any feasible solution of ($\mathrm{LP}_{\mathrm{adv}}$) $\hat{\boldsymbol{\lambda}}\in\mathbb{R}^{S\times B}_{\geq 0}$ induces a strategy $\hat{\boldsymbol{y}}$, defined as*

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

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

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

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 2.1).

Proof. By Lemma D.10, we know that ($\mathrm{LP}_{\mathrm{adv}}$) 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 ($\mathrm{LP}_{\mathrm{adv}}$) 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 D.12, and in particular (D.35) and (D.40). □

D.5 Convergence to a Nearly Stationary Point

In this section, we establish that IPGMAX reaches to an ϵ-nearly stationary point—in the sense of Definition B.2—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 4.2, which was first introduced in Section 4.4.3. 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 D.13. *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 $L \coloneqq \frac{\sqrt{\sum_{k=1}^{n} A_k + B}}{(1-\gamma)^2}$ and $\ell \coloneqq \frac{2\left(\sum_{k=1}^{n} A_k + B\right)}{(1-\gamma)^3}$. The next key result characterizes the iteration complexity required to reach an ϵ-nearly stationary point of $\phi(\cdot)$. The following analysis follows (Jin et al., 2020).

Proposition 4.2. *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 $0 \leq t^\star \leq T - 1$, such that $\left\|\boldsymbol{x}^{(t^\star)} - \tilde{\boldsymbol{x}}^{(t^\star)}\right\|_2 \leq \epsilon$, where $\tilde{\boldsymbol{x}}^{(t^\star)} \coloneqq \operatorname{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t^\star)})$.*

Proof. By virtue of the ℓ-smoothness of $V_{\boldsymbol{\rho}}(\boldsymbol{x}, \boldsymbol{y})$ (Lemma D.13), it follows that for any $\boldsymbol{x} \in \mathcal{X}$ and $0 \leq t \leq T-1$,

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

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

$$\tilde{\boldsymbol{x}}^{(t)} \coloneqq \arg\min_{\boldsymbol{x}' \in \mathcal{X}} \left\{ \phi(\boldsymbol{x}') + \frac{1}{2\lambda} \left\| \boldsymbol{x}^{(t)} - \boldsymbol{x}' \right\|^2 \right\}, \quad \text{(D.42)}$$

for any $0 \leq t \leq T-1$, where $\lambda \coloneqq \frac{1}{2\ell}$. Using the definition of Moreau envelope (Definition 4.4),

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

where

- (D.43) uses the fact that $\boldsymbol{x}_k^{(t+1)} \coloneqq \mathrm{Proj}_{\mathcal{X}_k} \left\{ \boldsymbol{x}_k^{(t)} - \eta \nabla_{\boldsymbol{x}_k} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) \right\}$ 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 $\tilde{\boldsymbol{x}}^{(t)} \in \mathcal{X}$;
- (D.44) follows from the fact that the projection operator is nonexpansive (Fact D.2);
- (D.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
- (D.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 defini-

tion of $\tilde{\boldsymbol{x}}^{(t)}$ in (D.42) and the definition of Moreau envelope with $\lambda = \frac{1}{2\ell}$ (Definition 4.4);

(ii) $V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) + \langle \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}), \tilde{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)} \rangle - \frac{\ell}{2} \left\| \tilde{\boldsymbol{x}} - \boldsymbol{x}^{(t)} \right\|^2 \leq \phi(\tilde{\boldsymbol{x}}^{(t)})$, which is an application of (D.41) for $\boldsymbol{x} \coloneqq \tilde{\boldsymbol{x}}^{(t)}$; and

(iii) $\left\| \nabla_{\boldsymbol{x}} V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)}) \right\|^2 \leq L^2$ by L-Lipschitz continuity of $V_{\boldsymbol{\rho}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t+1)})$ (Lemma D.13) combined with Fact D.1.

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

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(\phi(\boldsymbol{x}^{(t)}) - \phi(\tilde{\boldsymbol{x}}^{(t)}) - \frac{\ell}{2} \left\| \boldsymbol{x}^{(t)} - \tilde{\boldsymbol{x}}^{(t)} \right\|^2 \right) \leq \frac{\phi_\lambda(\boldsymbol{x}^{(0)}) - \phi_\lambda(\boldsymbol{x}^{(T)})}{2\eta\ell T} + \frac{\eta L^2}{2} \leq \frac{1}{2(1-\gamma)\eta\ell T} + \frac{\eta L^2}{2}, \tag{D.47}$$

since $\phi_\lambda(\boldsymbol{x}^{(T)}) \geq 0$, directly by Definition 4.4, and $\phi_\lambda(\boldsymbol{x}^{(0)}) \leq \phi(\boldsymbol{x}^{(0)}) \leq \frac{1}{1-\gamma}$, where the last inequality follows from Claim D.14. Therefore we conclude that there exists an iterate $t^\star$, with $0 \leq t^\star \leq T-1$, so that

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

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

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

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

A limitation of this proposition is that it only establishes a "best-iterate" guarantee. However, as we explained in Section 4.4.3, 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 D.3. *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 $\tilde{\boldsymbol{x}}^{(t')} \coloneqq \mathrm{prox}_{\phi/(2\ell)}(\boldsymbol{x}^{(t')})$.*

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

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

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

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

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

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

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 (D.50) given that $g^{(t)} \geq 0$ for all t. In turn, this implies our claim in light of (D.49). Finally, the proof of the corollary follows from a standard boosting argument, as well as rescaling ϵ by $\frac{1}{\sqrt{2}}$. □

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

Proof. In place of ϵ of Proposition 4.2 we set

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

which allows us to compute an ϵ-approximate Nash equilibrium by virtue of Theorem D.7. Then, the number of iterations reads

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

with a learning rate

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

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

D.5.1 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.7. *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.8. *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.7), 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.9. *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.10. *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 $s \in \mathcal{S}$,

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

Similarly, the value function can be written as

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

□

Claim D.11. *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.7 and D.8 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.12. *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 ϵ-approximate Nash equilibrium in $\mathcal{G}'$ is also an ϵ-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 ((4.4)). □

Claim D.13. *Let $\boldsymbol{\pi} \in \Pi$ be a joint stationary policy, and $\boldsymbol{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}$ be the induced visitation measure. Then, for every $s \in \mathcal{S}$,*

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

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

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

and

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

□

Claim D.14. *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 (4.3), 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.15. *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* ($\mathrm{P_{NE}}$). *The following inequalities hold:*

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

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

where (D.51) follows from the triangle inequality and the fact that $|r(s,\boldsymbol{a},b)| \leq 1$, and (D.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 and 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.14 and Proposition D.2).

- For Item 3, we begin by noting that $\hat{\boldsymbol{v}}$ and $\tilde{\boldsymbol{v}}$ are the unique optimal vectors of ($\mathrm{P_{NE}}$) for $\hat{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$ respectively (recall Proposition D.2). Further, by Proposition D.2, 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 2.1 and Lemma B.1).

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

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

  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: $\sum_{(a_1,\ldots,a_n)\in\mathcal{A}}\prod_{k=1}^{n}\hat{\boldsymbol{x}}_{k,s,a_k}=1$.

□