Optimistic Policy Gradient in Multi-Player Markov Games with a Single Controller: Convergence Beyond the Minty Property

Ioannis Anagnostides*[1], Ioannis Panageas[2], Gabriele Farina[3], and Tuomas Sandholm[4]

[1,4]Carnegie Mellon University
[2]University of California Irvine, Archimedes research unit
[3]Massachusetts Institute of Technology
[4]Strategy Robot, Inc.
[4]Strategic Machine, Inc.
[4]Optimized Markets, Inc.
ianagnos@cs.cmu.edu, ipanagea@ics.uci.edu, gfarina@mit.edu, and sandholm@cs.cmu.edu

December 22, 2023

Abstract

Policy gradient methods enjoy strong practical performance in numerous tasks in reinforcement learning. Their theoretical understanding in multiagent settings, however, remains limited, especially beyond two-player competitive and potential Markov games. In this paper, we develop a new framework to characterize *optimistic* policy gradient methods in multi-player Markov games with a *single controller*. Specifically, under the further assumption that the game exhibits an *equilibrium collapse*, in that the marginals of coarse correlated equilibria (CCE) induce Nash equilibria (NE), we show convergence to *stationary* ϵ-NE in $O(1/\epsilon^2)$ iterations, where $O(\cdot)$ suppresses polynomial factors in the natural parameters of the game. Such an equilibrium collapse is well-known to manifest itself in two-player zero-sum Markov games, but also occurs even in a class of multi-player Markov games with *separable interactions*, as established by recent work. As a result, we bypass known complexity barriers for computing stationary NE when either of our assumptions fails. Our approach relies on a natural generalization of the classical *Minty property* that we introduce, which we anticipate to have further applications beyond Markov games.

*Part of this work was performed as an intern at Meta AI.

1 Introduction

Realistic strategic interactions typically occur in stateful multiagent environments in which agents' decisions do not only determine their immediate rewards, but they also shape the next state of the system. Multiagent reinforcement learning (MARL), endowed with game-theoretic principles, furnishes a rigorous framework whereby artificial agents with strong performance guarantees can be developed even in such complex and volatile environments. Indeed, algorithmic advances in MARL have been translated to exciting empirical breakthroughs in grand AI challenges, covering two-player competitive games [Bowling et al., 2015, Brown and Sandholm, 2017, Moravčík et al., 2017, Perolat et al., 2022], as well as popular multi-player games [Brown and Sandholm, 2019, Bakhtin et al., 2022]. In spite of those remarkable developments, our theoretical understanding is still lagging behind, especially in multi-player games; this is precisely the primary focus of our paper.

In particular, we operate in the canonical framework of *Markov* (aka. stochastic) games [Shapley, 1953, Zhang et al., 2019], which captures multiagent Markov decision processes. Such settings have been the subject of intense scrutiny in recent years, with a flurry of results emerging for computing *Nash equilibria (NE)*—the standard game-theoretic equilibrium concept—in either two-player *zero-sum* games or multi-player *cooperative* games; our synopsis in Section 5 features numerous such developments. Algorithmic advances beyond those classes of games are scarce in the literature, and have been considerably impeded by recently established computational barriers for *stationary* NE even in turn-based two-player Markov games [Daskalakis et al., 2023b, Jin et al., 2023]; besides those recent lower bounds, any student of algorithmic game theory should also come to terms with the intrinsic intractability of NE even in one-shot (stateless) general-sum games [Daskalakis et al., 2006, Chen et al., 2009]. Yet, characterizing classes of games that elude those computational barriers is recognized as an important research direction in this line of work.

Our second key motivation—which will naturally coalesce with the considerations described above—is to characterize the behavior of *policy gradient* methods [Agarwal et al., 2021] in Markov games. Such techniques are especially natural from an optimization standpoint, and enjoy strong practical performance in a number of tasks [Schulman et al., 2015, 2017]. Furthermore, unlike other popular methods, they are amenable to function approximation [Sutton et al., 1999], thereby enabling to tackle enormous action spaces under continuous parameterizations.

In light of the inability of traditional gradient-based methods to converge even in normal-form zero-sum games [Mertikopoulos et al., 2018, Vlatakis-Gkaragkounis et al., 2020], we focus here on analyzing *optimistic* gradient descent (henceforth `OGD`). Optimism has been a crucial ingredient in attaining convergence in monotone settings and beyond [Cai et al., 2022b, Gorbunov et al., 2022, Golowich et al., 2020, Vankov et al., 2023], but its role is not well-understood even in two-player zero-sum Markov games. In this paper, we take an important step towards closing this gap, which will uncover as a byproduct a new class of multi-player Markov games for which we can compute efficiently stationary Nash equilibria.

1.1 Our results

To contextualize our approach, we first have to highlight a classical condition in variational inequalities (VIs) which guarantees convergence under certain first-order methods; namely, the so-called *Minty property* [Facchinei and Pang, 2003, Mertikopoulos et al., 2019]. A great number of existing results in optimization—not least in the multiagent setting—leverage that condition to analyze the behavior of learning algorithms. Unfortunately, Daskalakis et al. [2020] observed that the Minty property fails even in simple two-player Markov games with a single controller (recalled in Proposition 3.1). Furthermore, although several relaxations of the Minty property have been proposed (see our overview in Section 5), none has been able to capture such settings, thereby leaving open whether optimistic policy gradient methods converge.

In this context, our first main contribution is to introduce a generalization of the Minty property (Property 3.2) which addresses the aforementioned difficulties by capturing a broad class of multi-player Markov games. Specifically, our condition is more permissive in two crucial aspects. First, it allows distorting the underlying operator by a certain well-behaved function; as we explain in Section 3, this modification already suffices to subsume the counterexample of Daskalakis et al. [2020]—and generalizations thereof. The second modification relaxes the pointwise aspect of the original Minty property into an average guarantee, in the precise sense of Property 3.3.

Now the upshot is that `OGD`—under a suitable parameterization—still converges to an ϵ-strong solution of the induced VI problem after $T = O_\epsilon(1/\epsilon^2)$ iterations even under our more permissive criterion (Theorem 3.4), where the notation $O_\epsilon(\cdot)$ here suppresses polynomial factors in all natural parameters of the problem. We further establish that this guarantee is robust in the presence of perturbations akin to *relative deterministic noise* (Corollary B.9)—a ubiquitous model in control theory and optimization—and a certain slackness in our condition (Corollary B.6); the latter extension turns out to be crucial to capture policy optimization under greedy exploration.

As we have alluded to, the main application of our general theory targets multi-player Markov games, formally introduced in Section 2. In light of the inherent computational barriers described earlier, we need to impose additional structure to obtain meaningful guarantees. Our first assumption is that the underlying Markov game exhibits a certain *equilibrium collapse*, in that the marginals of *coarse correlated equilibria* induce Nash equilibria (Definition 4.4). It is well-known that such is the case in two-player zero-sum games, but recent work [Kalogiannis and Panageas, 2023, Park et al., 2023] has also revealed that equilibrium collapse persists even in a class of multi-player zero-sum games with *separable interactions*—building on a similar result in normal-form *polymatrix* games [Cai et al., 2016]. Yet, perhaps surprisingly and in stark contrast to normal-form games, equilibrium collapse alone does not suffice to enable efficient computation of *stationary* Nash equilibria [Daskalakis et al., 2023b, Jin et al., 2023]. For this reason, we further posit that the game admits a *single* controller, a quite classical setting as surveyed in Section 5. The upshot now is that under those two assumptions, our condition that generalizes the Minty property holds (Lemma 4.6), which brings us to one of our main results.

Theorem 1.1 (Informal; precise version in Theorem 4.7). *Consider any multi-player Markov game $\mathcal{G}$ with a single controller. If $\mathcal{G}$ exhibits equilibrium collapse, there is a* $\mathsf{poly}(|\mathcal{G}|, 1/\epsilon)$ *algorithm that receives gradient feedback and computes a stationary ϵ-Nash equilibrium.*

Above, we denote by $\mathsf{poly}(|\mathcal{G}|)$ a polynomial in the natural parameters of the game; the precise version appears as Theorem 4.7. In light of existing hardness results for computing stationary NE even in turn-based two-player Markov games [Daskalakis et al., 2023b, Jin et al., 2023], it is unlikely that the assumption of having a single controller can be significantly broadened. We also consider our theory investigating tractability beyond the Minty property to have interest beyond Markov games, but this is left for future work.

2 Preliminaries on Markov Games

In this section, we provide the necessary preliminaries on Markov games. For further background on Markov decision processes (MDPs), we refer to the works of Sutton and Barto [2018], Szepesvári [2022], Buşoniu et al. [2010].

Notation We let $\mathbb{N} = \{1, 2, \dots, \}$ denote the set of natural numbers and $\mathbb{N}^* \coloneqq \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$, we use the shorthand notations $[n] \coloneqq \{1, \dots, n\}$ and $[n]^* \coloneqq \{0, 1, \dots, n\}$. For a vector $\boldsymbol{z} \in \mathbb{R}^d$, we often use the variable $r \in [d]$ to index its coordinates, so that the rth coordinate is accessed by $\boldsymbol{z}[r]$. The inequality $\boldsymbol{z} \le \cdot$ is to be interpreted coordinate-wise. For two vectors $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, we denote by $\boldsymbol{z} \circ \boldsymbol{z}' \in \mathbb{R}^d$ their Hadamard product: $(\boldsymbol{z} \circ \boldsymbol{z}')[r] \coloneqq \boldsymbol{z}[r] \cdot \boldsymbol{z}'[r]$, for all $r \in [d]$.

Moreover, we will let $\mathcal{X}$ represent a convex nonempty and compact subset of a Euclidean space. We denote by $D_{\mathcal{X}}$ its ℓ_2 diameter. A function $F : \mathcal{X} \to \mathcal{X}$ is called L-Lipschitz continuous (with respect to the ℓ_2 norm $\|\cdot\|_2$) if $\|F(\boldsymbol{x}) - F(\boldsymbol{x}')\|_2 \le L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$; a differentiable function is called L-smooth if its gradient is L-Lipschitz continuous. Finally, to lighten the exposition, we will often use the $O_n(\cdot)$ notation to indicate the dependency of a function solely on parameter n.

Markov games We are interested in analyzing the convergence of policy gradient methods in multi-player Markov (aka. stochastic) games [Shapley, 1953] in the tabular regime. In such games, each player repeatedly elects actions within a multiagent MDP so as to maximize a reward function. Formally, a multi-player Markov game $\mathcal{G}$ is specified by a tuple $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \mathbb{P}, \{R_i\}_{i=1}^n, \zeta, \boldsymbol{\rho}) \eqqcolon \mathcal{G}$, whose constituents are defined as follows.

- $\mathcal{N} \coloneqq [n]$ is the set of players (or agents);
- $\mathcal{S}$ is a finite *state space*;
- $\mathcal{A}_i$ is the finite and nonempty set of available actions for each player $i \in [n]$ (for simplicity, and without loosing any generality, we posit that the action set does not depend on the underlying state); further, the joint action set is denoted by $\mathcal{A} \coloneqq \times_{i=1}^n \mathcal{A}_i$;
- $\mathbb{P}$ is the transition probability function, so that $\mathbb{P}(s'|s, \boldsymbol{a})$ represents the probability of transitioning to state $s' \in \mathcal{S}$ starting from state $s \in \mathcal{S}$ under the joint action $\boldsymbol{a} \in \mathcal{A}$;

- $R_i : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the (normalized) reward function of player $i \in [n]$, so that $R_i(s, \boldsymbol{a})$ represents the instantaneous reward when players select $\boldsymbol{a} \in \mathcal{A}$ in state $s \in \mathcal{S}$; (For simplicity, the rewards are deterministic.)
- $\zeta \coloneqq \min_{(s,\boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}} (1 - \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \boldsymbol{a})) > 0$ is a lower bound on the probability that the game will terminate at some step of the shared MDP; and
- $\boldsymbol{\rho} \in \Delta(\mathcal{S})$ is the initial distribution over states, assumed to have full support.

Learning algorithms Learning in such multiagent settings proceeds as follows. At every step $h \in \mathbb{N}^*$ each player $i \in [n]$ 1) observes the underlying state $s_h \in \mathcal{S}$; 2) selects an action $a_{i,h} \in \mathcal{A}_i$; and 3) subsequently receives some feedback from the environment, to be specified in the sequel. This process is repeated until the game terminates, which indeed occurs with probability 1 since we assume that $\zeta > 0$; the last step before the game terminates will be denoted by $H \in \mathbb{N}^*$, which is a random variable.

Policies A (potentially randomized) *stationary policy* for player $i \in [n]$ is a mapping $\boldsymbol{\pi}_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$; that is, a stationary policy remains invariant for all steps $h \in \mathbb{N}^*$. We only consider *Markovian* policies throughout this paper, without explicitly mentioning so. We will assume that players follow direct parameterization so that $\boldsymbol{\pi}_i \mapsto \boldsymbol{x}_i \in \Delta(\mathcal{A}_i)^{\mathcal{S}} \eqqcolon \mathcal{X}_i$ with the strategy $\boldsymbol{x}_{i,s}[a_i] \coloneqq \boldsymbol{\pi}_i(a_i|s)$ for all $(a_i, s) \in \mathcal{A}_i \times \mathcal{S}$. As such, strategies and policies will be used interchangeably. The set of all possible (stationary) policies for player $i \in [n]$ will be denoted by Π_i, while $\Pi \coloneqq \times_{i=1}^n \Pi_i$. We will also let $\mathcal{X} \coloneqq \times_{i=1}^n \mathcal{X}_i$.

Value The *value function* $V_i^{\boldsymbol{\pi}}(s)$ with respect to an initial state $s \in \mathcal{S}$ gives the expected reward for player $i \in [n]$ under the joint policy $\boldsymbol{\pi} \coloneqq (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n) \in \Pi$:

$$V_i^{\boldsymbol{\pi}}(s) \coloneqq \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{h=0}^{H} R_i(s_h, \boldsymbol{a}_h) | s_0 = s \right], \tag{1}$$

where the expectation above is taken with respect to the trajectory induced by the joint policy $\boldsymbol{\pi} \in \Pi$. We also generalize (1) by defining $V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \coloneqq \mathbb{E}_{s \sim \boldsymbol{\rho}}[V_i^{\boldsymbol{\pi}}(s)]$, where we recall that $\boldsymbol{\rho} \in \Delta(\mathcal{S})$. Similarly, the Q function with respect to player i is defined as

$$Q_i^{\boldsymbol{\pi}} : (s, \boldsymbol{a}) \mapsto \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{h=0}^{H} R_i(s_h, \boldsymbol{a}_h) | s_0 = s, \boldsymbol{a}_0 = \boldsymbol{a} \right],$$

where the expectation is again taken over the trajectory induced by $\boldsymbol{\pi} \in \Pi$. In this context, we will assume that each player receives as feedback from the environment the gradient of its value function with respect to its strategy.

Nash equilibrium Consider any player $i \in [n]$, and let $\boldsymbol{\mu}_{-i} : \mathcal{S} \to \Delta(\mathcal{A}_{-i})$ be a potentially correlated policy. We denote a stationary *best response policy* of i under $\boldsymbol{\mu}_{-i}$ by $\boldsymbol{\pi}_i^\dagger = \boldsymbol{\pi}_i^\dagger(\boldsymbol{\mu}_{-i}) \in \Pi_i$, so that $V_i^{\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho}) \coloneqq V_i^{\boldsymbol{\pi}_i^\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho})$.[1]

[1] It is well-known that there is always a stationary policy among the set of best response policies [Sutton and Barto, 2018].

Definition 2.1. A (stationary) *product* policy $\boldsymbol{\pi}^\star \in \Pi$ is an ϵ-Nash equilibrium if

$$\max_{1 \le i \le n} \left\{ V_i^{\dagger, \boldsymbol{\pi}^\star_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) \right\} \le \epsilon$$

.

Finally, for $\boldsymbol{\pi} \in \Pi$, we define the *state visitation distribution* $d_{s_0}^{\boldsymbol{\pi}} \in \Delta(\mathcal{S})$ by $d_{s_0}^{\boldsymbol{\pi}}[s] \propto \sum_{h \in \mathbb{N}^*} \mathbb{P}^{\boldsymbol{\pi}}(s_h = s | s_0)$, and $d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}} \coloneqq \mathbb{E}_{s_0 \sim \boldsymbol{\rho}}[d_{s_0}^{\boldsymbol{\pi}}]$. It will also be useful to consider the unnormalized counterparts of those distributions: $\tilde{d}_{s_0}^{\boldsymbol{\pi}}[s] = \sum_{h \in \mathbb{N}^*} \mathbb{P}^{\boldsymbol{\pi}}(s_h = s | s_0)$ and $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}} \coloneqq \mathbb{E}_{s_0 \sim \boldsymbol{\rho}}[\tilde{d}_{s_0}^{\boldsymbol{\pi}}]$.

3 Convergence Beyond the Minty Property

A classical condition that guarantees tractability for a variational inequality (VI) problem is the so-called *Minty property* [Facchinei and Pang, 2003]. To be precise, let $F : \mathcal{X} \to \mathcal{X}$ be a single-valued operator. The Minty property postulates the existence of a point $\boldsymbol{x}^\star \in \mathcal{X}$ such that

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \rangle \ge 0, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{2}$$

By now, there has been significant progress on understanding convergence of first-order methods under the Minty property. Unfortunately, and crucially for the purpose of this work, even two-player zero-sum Markov games fail to satisfy (2), as was first observed by Daskalakis et al. [2020]. In particular, they studied a simple two-player zero-sum Markov game known as Von Neumann's *ratio game* [Neumann, 1945], given by

$$V(\boldsymbol{x}_1, \boldsymbol{x}_2) \coloneqq \frac{\boldsymbol{x}_1^\top \mathbf{R} \boldsymbol{x}_2}{\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2}, \tag{3}$$

where $\boldsymbol{x}_1 \in \Delta(\mathcal{A}_1) \eqqcolon \mathcal{X}_1, \boldsymbol{x}_2 \in \Delta(\mathcal{A}_2) \eqqcolon \mathcal{X}_2$, and $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{\mathcal{A}_1 \times \mathcal{A}_2}$. It is further assumed that $\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2 \ge \zeta$, for some parameter $\zeta > 0$. The following proposition underlies much of the difficulty of analyzing policy gradient methods even under the simple ratio game (3).

Proposition 3.1 (Daskalakis et al., 2020)**.** *Fix any scalars $\epsilon, s \in (0,1)$, and suppose that*

$$\mathbf{R} \coloneqq \begin{pmatrix} -1 & \epsilon \\ -\epsilon & 0 \end{pmatrix} \quad \textit{and} \quad \mathbf{S} \coloneqq \begin{pmatrix} s & s \\ 1 & 1 \end{pmatrix}. \tag{4}$$

Then, the ratio game induced by the matrices in (4) fails to satisfy the Minty property (2).

Notwithstanding the above realization, empirical simulations suggest that optimistic policy gradient methods do in fact exhibit convergent behavior. As a result, a criterion more robust than the Minty property is needed. This is precisely the primary subject of this section.

Before we proceed with our generalized condition, let us make a further observation regarding the ratio game defined in Proposition 3.1 that will be useful in the sequel: that game admits a single controller—the transition probabilities depend solely on the strategy of one of the players; indeed, we note that $\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2 = \boldsymbol{x}_1^\top \boldsymbol{s}$ for any $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, where $\boldsymbol{s} = (s, 1)$—and thereby does not depend on $\boldsymbol{x}_2$.

Now, to address the aforementioned difficulties, we introduce and study a new condition, described below.

Property 3.2 (Generalized Minty property)**.** *Let $F : \mathcal{X} \to \mathcal{X}$ be such that $\mathcal{X} = \bigtimes_{r=1}^{d} \mathcal{Z}_r$ for $d \in \mathbb{N}$, and $\mathbf{1}_{\mathcal{Z}_r}$ be the vector with 1 for all entries corresponding to the component $\mathcal{Z}_r$, and 0 otherwise. Suppose further that $A : \mathcal{X} \to \mathcal{X}$ and $W : \mathcal{X} \to \mathcal{X}$ are functions such that*

- *$A(\boldsymbol{x}) \coloneqq \sum_{r=1}^{d} a_r(\boldsymbol{x}) \mathbf{1}_{\mathcal{Z}_r}$, where each $a_r : \mathcal{X} \to \mathbb{R}$ is α-Lipschitz continuous; $0 < \ell \le A(\boldsymbol{x}) \le h$; and*
- *$W(\boldsymbol{x}) \coloneqq \sum_{r=1}^{d} w_r(\boldsymbol{x}) \mathbf{1}_{\mathcal{Z}_r}$; $0 < \ell \le W(\boldsymbol{x}) \le h$.*

We say that the induced VI problem satisfies the (α, ℓ, h)-generalized Minty property if there exists $\boldsymbol{x}^\star \in \mathcal{X}$ so that

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \circ A(\boldsymbol{x}) \circ W(\boldsymbol{x}^\star) \rangle \ge 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{5}$$

where $\circ$ denotes component-wise multiplication.

Several remarks are in order regarding this property. First, a key assumption is that the underlying joint strategy space $\mathcal{X}$ can be decomposed as a Cartesian product, and that the functions A and W adhere to that structure. It is evident that Property 3.2 is more general than (2) since one can simply take A and W to be constant functions. In fact, when $d = 1$ the two conditions are equivalent; it is precisely the product structure of $\mathcal{X}$—which is inherently present in multi-player games—that makes Property 3.2 interesting. It is also worth noting a related condition appearing in [Harris et al., 2023, Appendix C.5], although it did not have any algorithmic implications.

Let us now relate Property 3.2 to the difficulty exposed by Proposition 3.1 in the context of the ratio game. One can show that if $\boldsymbol{x}^\star \in \mathcal{X}_1 \times \mathcal{X}_2$ is a Nash equilibrium of the ratio game, then if we take $A(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $W(\boldsymbol{x}_1, \boldsymbol{x}_2)$ as

$$\Big(\boldsymbol{x}_1^\top \boldsymbol{s}\, \overbrace{(1, \dots, 1)}^{|\mathcal{A}_1|}, \overbrace{(1, \dots, 1)}^{|\mathcal{A}_2|}\Big), \Big(\frac{1}{\boldsymbol{x}_1^\top \boldsymbol{s}} \overbrace{(1, \dots, 1)}^{|\mathcal{A}_1|}, \overbrace{(1, \dots, 1)}^{|\mathcal{A}_2|}\Big),$$

respectively, then (5) is satisfied (in this particular application, $d = 2$). Furthermore, having assumed that $\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2 \ge \zeta > 0$, we also have control over the lower bound ℓ (as well as the upper bound h); naturally, taking ℓ arbitrarily small trivializes Property 3.2, and so the interesting regime occurs when ℓ is bounded away from 0—this also becomes evident from the guarantee of Theorem 3.4. This observation regarding the VI induced by the ratio game is in fact non-trivial, and it is a byproduct of the minimax theorem shown by Shapley [1953]; in Section 4, we will prove this property in much greater generality.

As we shall see, Property 3.2 is already permissive enough to lead beyond known results. Nevertheless, to obtain as general results as possible, we next introduce a further extension of Property 3.2.

Property 3.3 (Average version of Property 3.2)**.** *Under the preconditions of Property 3.2 with respect to some triple $(\alpha, \ell, h) \in \mathbb{R}^3_{>0}$, we say that the induced VI problem satisfies the* average *(α, ℓ, h)-generalized Minty property if for any sequence $\sigma^{(T)} \coloneqq (\boldsymbol{x}^{(t)})_{1 \le t \le T}$ there exists $\mathcal{X} \ni \boldsymbol{x}^\star = \boldsymbol{x}^\star(\sigma^{(T)})$ so that*

$$\sum_{t=1}^{T} \langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \circ A(\boldsymbol{x}^{(t)}) \circ W(\boldsymbol{x}^\star) \rangle \ge 0. \tag{6}$$

Property 3.2 clearly implies Property 3.3 as a suitable $\boldsymbol{x}^\star \in \mathcal{X}$ would make every term in the summand (6) nonnegative; we have found that the additional generality of the latter property is useful for some applications (see Section 4).

We are now ready to proceed to the main result of this section, which concerns the behavior of the update rule

$$\begin{aligned} \boldsymbol{x}^{(t)} &:= \Pi_{\mathcal{X}}(\hat{\boldsymbol{x}}^{(t)} - \eta A(\boldsymbol{x}^{(t-1)}) \circ F(\boldsymbol{x}^{(t-1)})), \\ \hat{\boldsymbol{x}}^{(t+1)} &:= \Pi_{\mathcal{X}}(\hat{\boldsymbol{x}}^{(t)} - \eta A(\boldsymbol{x}^{(t)}) \circ F(\boldsymbol{x}^{(t)})), \end{aligned} \tag{OGD}$$

for $t \in \mathbb{N}$. Above, $\eta > 0$ is the learning rate; $\Pi_{\mathcal{X}}(\cdot)$ is the Euclidean projection operator; and $\boldsymbol{x}^{(0)} = \hat{\boldsymbol{x}}^{(1)} \in \mathcal{X}$ is an arbitrary initialization. The update rule (OGD) is the familiar optimistic gradient descent method [Chiang et al., 2012, Rakhlin and Sridharan, 2013a], but with an important twist: the operator $F(\boldsymbol{x}^{(t)})$ is now replaced by $A(\boldsymbol{x}^{(t)}) \circ F(\boldsymbol{x}^{(t)})$, where $A : \mathcal{X} \to \mathcal{X}$ is a problem-specific function—in direct correspondence with Property 3.2; this can be simply viewed as incorporating a time-varying but non-vanishing learning rate. We remark that it is assumed that A can be accessed in order to perform the update rule (OGD); this assumption will be discussed and addressed in the context of our applications in Section 4. Below, we show that Property 3.3 is indeed sufficient to guarantee tractability for the induced VI problem, in the following formal sense.

Theorem 3.4. *Let $\mathcal{X} = \times_{r=1}^d \mathcal{Z}_r$ for some $d \in \mathbb{N}$ and $F : \mathcal{X} \to \mathcal{X}$ be an L-Lipschitz continuous operator with $B_F := \max_{1 \le r \le d} \|F_r\|_2$. Suppose further that the average (α, ℓ, h)-generalized Minty property (Property 3.3) holds. Then, for any $\epsilon > 0$, after $T \ge \frac{2D_{\mathcal{X}}^2 h}{\ell \epsilon^2}$ iterations of (OGD) with learning rate $\eta \le \frac{1}{4}\sqrt{\frac{\ell}{h^3 L^2 + h B_F^2 \alpha^2 d}}$ there is a point $\boldsymbol{x}^{(t)} \in \mathcal{X}$ such that for any $\boldsymbol{x}^\star \in \mathcal{X}$,*

$$\langle \boldsymbol{x}^{(t)}, F(\boldsymbol{x}^{(t)}) \rangle - \langle \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \rangle \le 2d \left(\frac{\max_{1 \le r \le d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_F}{\ell} \right) \epsilon.$$

Proof sketch The proof of this theorem is deferred to Appendix B.1, but we briefly describe the key ingredients here. In a nutshell, we analyze optimistic gradient descent (OGD) following the regret analysis of optimistic mirror descent (Proposition B.1) in the context of multi-player games [Rakhlin and Sridharan, 2013b, Syrgkanis et al., 2015]; more precisely, we essentially view each component over $\mathcal{Z}_r$, comprising the Cartesian product $\mathcal{X} := \times_{r=1}^d \mathcal{Z}_r$, as a separate player. The twist is that—in accordance with (OGD)—the observed utility is taken to be $F_r(\boldsymbol{x}^{(t)}) \circ A_r(\boldsymbol{x}^{(t)})$, instead of $F_r(\boldsymbol{x}^{(t)})$, where F_r is the rth component of F. Importantly, the structure imposed on $A(\boldsymbol{x})$ by Property 3.3 enables us to show that a suitable *weighted* notion of regret enjoys a certain upper bound independent of both A and W. Thus, leveraging (6), we are able to show—following earlier work [Anagnostides et al., 2022, Zhang et al., 2022a]—that the second-order path lengths of the dynamics are bounded (Corollary B.3). Then, Theorem 3.4 follows by the assumption that $0 < \ell \le A(\boldsymbol{x}) \le h$; that is, incorporating $A(\boldsymbol{x})$ into the update rule (OGD) does not distort by much the underlying operator F.

A point $\boldsymbol{x}^{(t)}$ such that $\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \rangle \le \epsilon$ for any $\boldsymbol{x}^\star \in \mathcal{X}$—as in the guarantee of Theorem 3.4—is known as an ϵ-approximate solution to the *Stampacchia* VI problem (aka.

an ϵ-approximate strong solution). To make this guarantee more concrete, and to connect it with the forthcoming application in Section 4, let us consider an n-player game so that $F = (F_1, \dots, F_n)$ and $F_i \coloneqq -\nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x})$, where $u_i : \mathcal{X} \to \mathbb{R}$ is the differentiable utility of player $i \in [n]$.

Corollary 3.5. *Under the preconditions of Theorem 3.4, we can compute a point $\boldsymbol{x} \in \mathcal{X}$ after a sufficiently large $T = O_\epsilon(1/\epsilon^2)$ iterations of (OGD), for any $\epsilon > 0$, such that*

1. *if each $u_i(\boldsymbol{x}_i, \cdot)$ is L-smooth, then for any player $i \in [n]$ and $\boldsymbol{x}_i^\star \in \mathcal{X}_i$ with $\|\boldsymbol{x}_i^\star - \boldsymbol{x}_i\|_2 \le \delta$, $u_i(\boldsymbol{x}) - u_i(\boldsymbol{x}_i^\star, \boldsymbol{x}_{-i}) \ge -\epsilon - \frac{L}{2}\delta^2$;*

2. *if each $u_i(\boldsymbol{x}_i, \cdot)$ is gradient dominant, then for any player $i \in [n]$ and $\boldsymbol{x}_i^\star \in \mathcal{X}_i$, $u_i(\boldsymbol{x}) - u_i(\boldsymbol{x}_i^\star, \boldsymbol{x}_{-i}) \ge -\epsilon$.*

To be precise, the (per-player) gradient dominance property postulates that

$$u_i(\boldsymbol{x}) - \max_{\boldsymbol{x}_i^\star \in \mathcal{X}_i} u_i(\boldsymbol{x}_i^\star, \boldsymbol{x}_{-i}) \ge G \min_{\boldsymbol{x}_i^\star \in \mathcal{X}_i} \langle \boldsymbol{x}_i - \boldsymbol{x}_i^\star, \nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x}) \rangle$$

for all $\boldsymbol{x} \in \mathcal{X}$, where $G > 0$ is some parameter. As such, Item 2 follows directly by definition and Theorem 3.4. Item 1 above is more permissive, but only yields a local optimality guarantee. Still, it turns out that computing such points is hard even in smooth min-max optimization [Daskalakis et al., 2021, 2023a]; more precisely, Item 1 is interesting in the local regime $\delta < \sqrt{\frac{2\epsilon}{L}}$; see [Daskalakis et al., 2021, Definition 1.1]; other notions of local optimality have also been studied in the literature [Jin et al., 2020], but this is not in our scope here.

Before we conclude this section, let us highlight some interesting extensions of Theorem 3.4. First, one can further broaden the scope of Property 3.3 by replacing the right-hand side of (6) by $-\gamma T$, for some parameter $\gamma \in \mathbb{R}_{\ge 0}$. In Corollary B.6, we show that we can then compute an $O_{\epsilon,\gamma}(\sqrt{\gamma} + \epsilon)$-approximate strong solution. This particular relaxation turns out to be crucial to capture policy parameterization under $\Theta_\gamma(\gamma)$-greedy exploration (Remark B.13 elaborates on this point). In such settings, one has control over the parameter γ, and so by taking $\gamma \coloneqq \epsilon^2$ we can generalize the guarantee of Theorem 3.4.

Our second extension concerns the behavior of (OGD) in the presence of noise. Our model of perturbation is akin to the standard relative deterministic noise, wherein the error is proportional to the distance from optimality, for an appropriate notion of distance [Polyak, 1987, Lessard et al., 2016]. More precisely, for parameters $\rho, \delta > 0$, we assume access to a noisy operator $F^{\delta,\rho} : \mathcal{X} \to \mathcal{X}$ such that $\|F^{\delta,\rho}(\boldsymbol{x}) - F(\boldsymbol{x})\|_2 \le \delta \cdot \textsc{EqGap}(\boldsymbol{x})$, where $\textsc{EqGap}(\boldsymbol{x}) : \mathcal{X} \ni \boldsymbol{x} \mapsto \max_{\boldsymbol{x}^\star \in \mathcal{X}} \langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \rangle$ represents the equilibrium gap. We further posit that $F^{\delta,\rho}$ satisfies a relaxed version of Property 3.3 in which the right-hand side of (6) can be as small as $-\rho \sum_{t=1}^T (\textsc{EqGap}(\boldsymbol{x}^{(t)}))^2$. In this context, Corollary B.9 reassures us that the conclusion of Theorem 3.4 is robust if δ and ρ are small enough.

4 Optimistic Policy Gradient in Multi-Player Markov Games

In this section, we leverage the theory developed earlier in Section 3 in order to characterize optimistic policy gradient methods in multi-player Markov games. In light of the inherent hardness of computing Nash equilibria in general-sum games, we will restrict our attention to more structured classes of Markov games. The first assumption we consider can be viewed as a natural counterpart of the Minty property, but with respect to the value functions—without linearizing by taking the gradients.

Assumption 4.1. Let $\mathcal{G}$ be a Markov game. There exists a joint policy $(\boldsymbol{\pi}_1^\star, \dots, \boldsymbol{\pi}_n^\star) \in \Pi$ such that

$$\sum_{i=1}^n V_i^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}(\boldsymbol{\rho}) - \sum_{i=1}^n V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \geq 0, \forall (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n) \in \Pi.$$

Crucially, unlike the Minty property (2), Assumption 4.1 subsumes two-player zero-sum (Markov) games. Indeed, Shapley [1953] proved that there exist policies $(\boldsymbol{\pi}_1^\star, \boldsymbol{\pi}_2^\star) \in \Pi$ such that

$$V^{\boldsymbol{\pi}_1^\star, \boldsymbol{\pi}_2}(\boldsymbol{\rho}) \leq V^{\boldsymbol{\pi}_1^\star, \boldsymbol{\pi}_2^\star}(\boldsymbol{\rho}) \leq V^{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2^\star}(\boldsymbol{\rho}), \quad \forall (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \in \Pi.$$

Here, $V_1(\boldsymbol{\rho}) \coloneqq -V(\boldsymbol{\rho})$ and $V_2(\boldsymbol{\rho}) \coloneqq V(\boldsymbol{\rho})$ (since the game is zero-sum). The above display establishes Assumption 4.1 since $V_1^{\boldsymbol{\pi}_1^\star, \boldsymbol{\pi}_2}(\boldsymbol{\rho}) + V_2^{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2^\star}(\boldsymbol{\rho}) \geq 0$. In other words, Assumption 4.1 is a byproduct of Shapley's minimax theorem.

It is worth noting that any (stationary) Nash equilibrium $(\boldsymbol{\pi}_1^\star, \dots, \boldsymbol{\pi}_n^\star) \in \Pi$ satisfies

$$\sum_{i=1}^n V_i^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) - \sum_{i=1}^n V_i^{\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^\star}(\boldsymbol{\rho}) \geq 0, \forall (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n) \in \Pi,$$

which closely resembles the condition of Assumption 4.1. However, unlike Assumption 4.1, the above condition always holds since (stationary) NE always exist [Fink, 1964].

As it will become clear, Assumption 4.1 is naturally associated with Property 3.2. We also introduce a more permissive assumption in direct correspondence with Property 3.3.

Assumption 4.2. Let $\mathcal{G}$ be a Markov game. For any sequence of product policies $\sigma^{(T)} \coloneqq (\boldsymbol{\pi}^{(t)})_{1 \leq t \leq T}$, there exists $\Pi \ni \boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star(\sigma^{(T)})$ such that

$$\sum_{t=1}^T \sum_{i=1}^n V_i^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}^{(t)}}(\boldsymbol{\rho}) - \sum_{t=1}^T \sum_{i=1}^n V_i^{\boldsymbol{\pi}^{(t)}}(\boldsymbol{\rho}) \geq 0.$$

Beyond the two-player zero-sum setting, we first show that Assumption 4.2 is satisfied for the class of *zero-sum polymatrix Markov* games [Kalogiannis and Panageas, 2023] (see also [Park et al., 2023]).

Polymatrix zero-sum Markov games A polymatrix game is based on an undirected graph $G = (V, E)$. Each node $i \in V$ is (uniquely) associated with a player, while every edge $\{i, i'\} \in E$ represents a pairwise interaction between players i and i'. It is assumed that the reward of each player is given by the sum of the rewards from each game engaged with its neighbors. The zero-sum aspect imposes that the sum of the players' rewards is 0. Such games were investigated by Cai et al. [2016] (see also Even-Dar et al. [2009] for a more general treatment) under the normal form representation. For the Markov setting, Kalogiannis and Panageas [2023] further assumed that in each state there is a single player (not necessarily the same) whose actions determine the transition probabilities to the next state. For that class of games, with a careful examination of their analysis we are able to show the following result.

Proposition 4.3. *Assumption 4.2 is satisfied for any polymatrix zero-sum Markov game.*

In fact, this result is a byproduct of a more general characterization that we prove. To explain our result, we first recall the concept of a *coarse correlated equilibrium (CCE)*, which relaxes Definition 2.1 by allowing *correlated* policies. We will further use the concept of an ϵ-*average* CCE (henceforth ϵ-ACCE), a relaxation of CCE in which the sum—instead of the maximum as in CCE—of the players' deviation benefits must be at most ϵ. Precise definitions are deferred to Appendix A, as they are not crucial for the purpose of this section. In this context, we introduce the following definition.

Definition 4.4 (Equilibrium collapse)**.** Let $\mathcal{G}$ be a Markov game. We say that $\mathcal{G}$ exhibits *equilibrium collapse* if there is a $C = C(\mathcal{G}) \in \mathbb{R}_{>0}$ such that for any stationary ϵ-ACCE $\boldsymbol{\mu} \in \Delta(\mathcal{A})^{\mathcal{S}}$ of $\mathcal{G}$, the marginal policies $(\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n) = (\boldsymbol{\pi}_1(\boldsymbol{\mu}), \dots, \boldsymbol{\pi}_n(\boldsymbol{\mu}))$ form a $(C\epsilon)$-Nash equilibrium of $\mathcal{G}$.

We remark that the prior work on zero-sum polymatrix Markov games established equilibrium collapse with respect to ϵ-CCE [Kalogiannis et al., 2023], but their argument readily carries over for ACCE as well. Proposition 4.3 is thus implied by the following result.

Proposition 4.5. *Assumption 4.2 is satisfied in any Markov game $\mathcal{G}$ exhibiting equilibrium collapse per Definition 4.4.*

Having justified Assumptions 4.1 and 4.2, we now proceed to establishing Property 3.2. Taking a step back, one might hope that equilibrium collapse (in the sense of Definition 4.4) would already suffice to efficiently compute stationary NE—as in the case of normal-form games. However, recent lower bounds [Daskalakis et al., 2023b, Jin et al., 2023] dispel any such hopes, thereby necessitating additional structure in order to elude those intractability barriers. This is precisely where the admission of a single controller comes into play, an assumption crucial for establishing Property 3.3. Indeed, this is shown in the following key lemma, which relies on the expression of the difference of the value function (Lemma B.12) and the connection between the Q function and the gradient of the value function (Lemma B.11). In accordance with our theory in Section 3, we let $F_{\mathcal{G}}(\boldsymbol{x}) \coloneqq -(\nabla_{\boldsymbol{x}_1} V_1(\boldsymbol{\rho}), \dots, \nabla_{\boldsymbol{x}_n} V_n(\boldsymbol{\rho}))$; see Remark B.10 for a clarification regarding differentiability of the value function in spite of the empty interior of $\mathcal{X}$.

Lemma 4.6. *Consider a Markov game $\mathcal{G}$, and let $\Lambda_i(\boldsymbol{x}, \boldsymbol{x}^\star)[s, a_i] \coloneqq \frac{\tilde{d}_{\boldsymbol{\rho}}^{\pi_i^\star, \boldsymbol{\pi}_{-i}}[s]}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s]}$ for $i \in [n]$ and $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$. Further, let $\Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \coloneqq (\Lambda_1(\boldsymbol{x}, \boldsymbol{x}^\star), \dots, \Lambda_n(\boldsymbol{x}, \boldsymbol{x}^\star))$. If Assumption 4.1 holds, then there exists $\boldsymbol{x}^\star \in \mathcal{X}$ such that*

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \circ \Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \rangle \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

In particular, if $\mathcal{G}$ admits a single controller, denoted by $\mathsf{cntrl}_{\mathcal{G}}$, then Property 3.2 holds with

$$A_i(\boldsymbol{x})[s, a_i] \coloneqq \begin{cases} 1 & : \text{if } i \neq \mathsf{cntrl}_{\mathcal{G}} \\ \left(\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i}[s]\right)^{-1} & : \text{if } i = \mathsf{cntrl}_{\mathcal{G}}, \end{cases}$$

and

$$W_i(\boldsymbol{x}^\star)[s, a_i] \coloneqq \begin{cases} 1 & : \text{if } i \neq \mathsf{cntrl}_{\mathcal{G}} \\ \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star}[s] & : \text{if } i = \mathsf{cntrl}_{\mathcal{G}}. \end{cases}$$

We see that (7)—a generalization of Property 3.2—holds without any additional assumptions on the transition probabilities. Yet, decoupling $\Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \coloneqq A(\boldsymbol{x}) \circ W(\boldsymbol{x}^\star)$ in the sense of Property 3.2 turns out to be crucial to apply our techniques. In fact, the recent hardness result of Park et al. [2023] suggests that the general case should be intractable. We further remark that Lemma 4.6 applies similarly to conclude Property 3.3 if we substitute Assumption 4.1 by Assumption 4.2.

Finally, having established Lemma 4.6, we can now apply Theorem 3.4 along with the gradient dominance property (Lemma B.14) to obtain one of our main results. Specifically, in Appendix B we appropriately bound all of the involved parameters appearing in Theorem 3.4; as usual, this includes a certain *distribution mismatch coefficient* $C_{\mathcal{G}}$ (defined in (13) of Appendix B)—the multi-player analog of the quantity considered by Daskalakis et al. [2020]—as well as a dependency on $1/\|\boldsymbol{\rho}\|_\infty$, necessitating that the original distribution $\boldsymbol{\rho}$ assigns a non-negligible probability mass to all states.

Theorem 4.7. *Let $\mathcal{G}$ be a Markov game that satisfies Assumption 4.2 and admits a single controller. Then, (OGD) after $1/\epsilon^2 \cdot \mathsf{poly}(n, \sum_{i=1}^n |\mathcal{A}_i|, |\mathcal{S}|, 1/\zeta, C_{\mathcal{G}}, 1/\|\boldsymbol{\rho}\|_\infty)$ iterations computes a stationary ϵ-NE.*

The importance of Theorem 3.4 stems not just from its computational complexity implications, but also from its applicability in a decentralized environment. Indeed, all players are performing gradient steps without any further information from their environment, with the sole exception of the controller. In particular, as predicted by Lemma 4.6, performing the update rule (OGD) requires some further access to the environment in order to estimate the (unnormalized) state visitation distribution $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[\cdot]$; using standard arguments, this requires $\mathsf{poly}(|\mathcal{G}|, 1/\epsilon)$ time to determine within ϵ-error, which suffices for applying Theorem 3.4 (see Remark B.4).

An illustrative comparison between (OGD) with a time-varying but non-vanishing learning rate—per its update rule—and the vanilla version with a constant learning rate in the context of the ratio game appears in Appendix C.

It is worth noting that our proof technique shares an interesting conceptual similarity with the approach of Erez et al. [2023], also based on a weighted notion of regret. The key point of departure is that we explicitly incorporate the weights into the update rule (OGD), which in turn induces a second-order dependency on the deviation of the weights in lieu of a first-order bound; this turns out to be crucial for establishing Theorem 4.7. Yet, our approach is more restrictive in that it rests on having a single controller.

5 Further Related Work

Computing and learning equilibria in Markov games has attracted considerable interest recently. Most focus has been on the Nash equilibrium in either identical-interest—or more generally, potential—games [Fox et al., 2022, Leonardos et al., 2022, Alatur et al., 2023, Aydin and Eksin, 2023, Ding et al., 2022, Zhang et al., 2022c, Maheshwari et al., 2022, Macua et al., 2018, Chen et al., 2022], or two-player zero-sum Markov games [Daskalakis et al., 2020, Cen et al., 2023, Cai et al., 2023, Chen et al., 2023c, Wei et al., 2021, Zhang et al., 2020, Sayin et al., 2021, Huang et al., 2022, Cui and Du, 2022, Perolat et al., 2015, Zeng et al., 2022, Pattathil et al., 2023, Yang and Ma, 2023, Arslantas et al., 2023, Chen et al., 2023b], albeit with a few exceptions [Qin and Etesami, 2023, Sayin, 2023, Giannou et al., 2022, Kalogiannis and Panageas, 2023, Kalogiannis et al., 2023, Park et al., 2023, Ma et al., 2023]. In general-sum multi-player games, in light of the intractability of Nash equilibria, most focus has been on computing or indeed learning (coarse) correlated equilibria [Daskalakis et al., 2023b, Jin et al., 2021, Wang et al., 2023, Erez et al., 2023, Liu et al., 2022, Zhang et al., 2022b, Foster et al., 2023].

Nevertheless, an important question has been to identify classes of multi-player games that circumvent the intractability of NE in general games. For example, recent work [Kalogiannis and Panageas, 2023, Park et al., 2023] investigates the class of polymatrix Markov games, which is based on the homonymous class of normal-form games [Cai and Daskalakis, 2011, Cai et al., 2016]; indeed, the topic of network games has been particularly popular in the literature on MARL (see [Zhang et al., 2018, Chu et al., 2020, Parise and Ozdaglar, 2019], and references therein). Specifically, Kalogiannis and Panageas [2023] and Park et al. [2023] leverage the equilibrium collapse of CCE to NE to show that Markov NE can be computed efficiently; in stark contrast, Park et al. [2023] showed that computing a *stationary* NE is PPAD-hard; the latter hardness result is based on earlier work by Daskalakis et al. [2023b], Jin et al. [2023]. In the class of polymatrix zero-sum Markov games, our novelty compared to earlier work [Kalogiannis and Panageas, 2023, Park et al., 2023] (see also the concurrent paper of Ma et al. [2023]) lies in showing convergence to *stationary* Nash equilibria; this does not contradict the aforementioned hardness results since we impose an additional assumption on the transitions. It is worth underscoring that stationarity is a fundamental desideratum with a long history in the literature on repeated games; among other benefits, stationary policies enjoy a much more memory-efficient encoding, which becomes especially crucial when each policy is represented via an enormous neural network with millions of parameters, while stationary policies are also arguably more interpretable.

Beyond games with separable interactions, Kalogiannis et al. [2023] showed that NE can be computed efficiently in a class of games that subsumes both zero-sum and potential games—namely, adversarial team Markov games; see also [Emmons et al., 2022, Wang and Sandholm, 2002] for pertinent results. It is also worth noting that certain refinements of NE—such as *strict* equilibria—have been shown to be attractors under policy gradient methods [Giannou et al., 2022], although such refinements are not universal.

Naturally, gradient-based methods have also received considerable attention in imperfect-information extensive-form games [Hoda et al., 2010, Lee et al., 2021, Piliouras et al., 2022, Zinkevich et al., 2007, Liu et al., 2023], as well as the more tractable class of normal-form games [Hsieh et al., 2021, Hussain et al., 2023]. Even for the latter class of games, it is known that gradient-based methods may fail to converge pointwise to Nash equilibria [Vlatakis-Gkaragkounis et al., 2020, Mertikopoulos et al., 2018]. In stark contrast, it has been documented that *optimism*, a minor modification akin to the extra-gradient method introduced in the online learning literature by Rakhlin and Sridharan [2013a], Chiang et al. [2012], leads to last-iterate convergence in monotone settings [Cai et al., 2022b, Gorbunov et al., 2022, Golowich et al., 2020]. Further, beyond the monotone regime, ample of prior work has endeavored to identify broader classes of tractable VIs, such as the *weak* Minty property put forward by Diakonikolas et al. [2021]. In turn, this has engendered a considerable recent body of work; we refer to the papers of Pethick et al. [2023], Cai and Zheng [2023], Pethick et al. [2022], Lee and Kim [2021], Cai et al. [2022a], Mahdavinia et al. [2022], Vankov et al. [2023], and the many references therein.

Finally, we highlight that Markov games with a single controller have a rich history; see [Parthasarathy and Raghavan, 1981, Başar and Olsder, 1998, Eldosouky et al., 2016, Guan et al., 2016, Qiu et al., 2021, Sayin et al., 2022]; those references contain ample motivation and examples of realistic strategic interactions that can be faithfully modeled as Markov games with a single controller. For example, Eldosouky et al. [2016] cast strategically configuring a wireless network so as to protect against potential attacks as a security game in which the defender serves as the sole controller.

6 Conclusions and Future Work

In conclusion, we have furnished a natural generalization of the classical Minty property, and we showed that computational tractability persists even under our more permissive condition. We also applied our general theory to obtain new convergence results to stationary Nash equilibria for optimistic policy gradient methods in a broad class of multi-player Markov games.

A number of interesting questions arise from our work. First, our new condition (Property 3.2) crucially relies on the product structure of the joint strategy space. While such structure is always present in multi-player games with uncoupled strategy sets, the canonical case treated in the literature, it may break in some settings of interest [Jordan et al., 2023, Goktas et al., 2023, Goktas and Greenwald, 2022]. Extending the theory of Section 3 to capture such settings is an interesting avenue for future work. Furthermore, we have seen that any Markov game that exhibits equilibrium collapse satisfies property (7), without assuming the existence of a single controller. Understanding when property (7) suffices to ensure computational tractability is another promising direction.

Acknowledgments

We are grateful to anonymous reviewers at AAAI for valuable feedback. This material is based on work supported by the Vannevar Bush Faculty Fellowship ONR N00014-23-1-2876, National Science Foundation grants RI-2312342 and RI-1901403, ARO award W911NF2210266, and NIH award A240108S001. Ioannis Panageas would like to acknowledge startup grant from UC Irvine. Part of this work was conducted while Ioannis was visiting Archimedes research unit. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

P. Alatur, G. Ramponi, N. He, and A. Krause. Provably learning nash policies in constrained markov potential games, 2023.

I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 536–581. PMLR, 2022.

Y. Arslantas, E. Yuceel, Y. Yalin, and M. O. Sayin. Convergence of heterogeneous learning dynamics in zero-sum stochastic games, 2023.

S. Aydin and C. Eksin. Policy gradient play with networked agents in markov potential games. In *Learning for Dynamics and Control Conference, L4DC 2023*, volume 211 of *Proceedings of Machine Learning Research*, pages 184–195. PMLR, 2023.

A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, M. Komeili, K. Konath, M. Kwon, A. Lerer, M. Lewis, A. H. Miller, S. Mitts, A. Renduchintala, S. Roller, D. Rowe, W. Shi, J. Spisak, A. Wei, D. Wu, H. Zhang, and M. Zijlstra. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

T. Başar and G. J. Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

M. Bowling, N. Burch, M. Johanson, and O. Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218), January 2015.

N. Brown and T. Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, pages 418–424, Dec. 2017.

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

L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221, 2010.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 2020.

Y. Cai and C. Daskalakis. On minmax theorems for multiplayer games. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 217–234. SIAM, 2011.

Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

Y. Cai, O. Candogan, C. Daskalakis, and C. H. Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization, 2022a.

Y. Cai, A. Oikonomou, and W. Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022b.

Y. Cai, H. Luo, C. Wei, and W. Zheng. Uncoupled and convergent learning in two-player zero-sum markov games, 2023.

S. Cen, Y. Chi, S. S. Du, and L. Xiao. Faster last-iterate convergence of policy optimization in zero-sum markov games. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

D. Chen, Q. Zhang, and T. T. Doan. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games, 2022.

X. Chen, X. Deng, and S.-H. Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, 2009.

X. Chen, E. Minasyan, J. D. Lee, and E. Hazan. Regret guarantees for online deep control. In *Learning for Dynamics and Control Conference, L4DC*, volume 211 of *Proceedings of Machine Learning Research*, pages 1032–1045. PMLR, 2023a.

Z. Chen, K. Zhang, E. Mazumdar, A. Ozdaglar, and A. Wierman. Two-timescale q-learning with function approximation in zero-sum stochastic games, 2023b.

Z. Chen, K. Zhang, E. Mazumdar, A. E. Ozdaglar, and A. Wierman. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games, 2023c.

C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1, 2012.

T. Chu, S. Chinchali, and S. Katti. Multi-agent reinforcement learning for networked system control. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Q. Cui and S. S. Du. When are offline two-player zero-sum markov games solvable? In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 25779–25791, 2022.

C. Daskalakis, P. Goldberg, and C. Papadimitriou. The complexity of computing a Nash equilibrium. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*, 2006.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*, pages 1466–1478. ACM, 2021.

C. Daskalakis, N. Golowich, S. Skoulakis, and E. Zampetakis. Stay-on-the-ridge: Guaranteed convergence to local minimax equilibrium in nonconvex-nonconcave games. In *Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 5146–5198. PMLR, 2023a.

C. Daskalakis, N. Golowich, and K. Zhang. The complexity of markov equilibrium in stochastic games. In *Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 4180–4234. PMLR, 2023b.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 2746–2754. PMLR, 2021.

D. Ding, C. Wei, K. Zhang, and M. R. Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 5166–5220. PMLR, 2022.

A. Eldosouky, W. Saad, and D. Niyato. Single controller stochastic games for optimized moving target defense. In *2016 IEEE International Conference on Communications, ICC 2016*, pages 1–6. IEEE, 2016.

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

L. Erez, T. Lancewicki, U. Sherman, T. Koren, and Y. Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 9343–9373. PMLR, 2023.

E. Even-Dar, Y. Mansour, and U. Nadav. On the convergence of regret minimization dynamics in concave games. In M. Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009*, pages 523–532. ACM, 2009.

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

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.

D. J. Foster, N. Golowich, and S. M. Kakade. Hardness of independent learning and sparse equilibrium computation in markov games. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 10188–10221. PMLR, 2023.

R. Fox, S. M. Mcaleer, W. Overman, and I. Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

A. Giannou, K. Lotidis, P. Mertikopoulos, and E. Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

D. Goktas and A. Greenwald. Exploitability minimization in games and beyond. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

D. Goktas, D. C. Parkes, I. Gemp, L. Marris, G. Piliouras, R. Elie, G. Lever, and A. Tacchetti. Generative adversarial equilibrium solvers. *CoRR*, abs/2302.06607, 2023.

N. Golowich, S. Pattathil, C. Daskalakis, and A. E. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1758–1784. PMLR, 2020.

E. Gorbunov, A. Taylor, and G. Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

P. Guan, M. Raginsky, R. Willett, and D. Zois. Regret minimization algorithms for single-controller zero-sum stochastic games. In *55th IEEE Conference on Decision and Control, CDC 2016*, pages 7075–7080. IEEE, 2016.

K. Harris, I. Anagnostides, G. Farina, M. Khodak, S. Wu, and T. Sandholm. Meta-learning in games. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

S. Hoda, A. Gilpin, J. Peña, and T. Sandholm. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2), 2010.

Y. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In M. Belkin and S. Kpotufe, editors, *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 2388–2422. PMLR, 2021.

B. Huang, J. D. Lee, Z. Wang, and Z. Yang. Towards general function approximation in zero-sum markov games. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

A. A. Hussain, F. Belardinelli, and G. Piliouras. Asymptotic convergence and performance of multi-agent q-learning dynamics. In *Autonomous Agents and Multi-Agent Systems*, pages 1578–1586. ACM, 2023.

A. Jacot, C. Hongler, and F. Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8580–8589, 2018.

C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 2020.

C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning - A simple, efficient, decentralized algorithm for multiagent RL, 2021.

Y. Jin, V. Muthukumar, and A. Sidford. The complexity of infinite-horizon general-sum stochastic games. In Y. T. Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, volume 251 of *LIPIcs*, pages 76:1–76:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

M. I. Jordan, T. Lin, and M. Zampetakis. First-order algorithms for nonlinear generalized nash equilibrium problems. *Journal of Machine Learning Research*, 24:38:1–38:46, 2023.

F. Kalogiannis and I. Panageas. Zero-sum polymatrix markov games: Equilibrium collapse and efficient computation of nash equilibria. *CoRR*, abs/2305.14329, 2023.

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

C. Lee, C. Kroer, and H. Luo. Last-iterate convergence in extensive-form games. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 14293–14305, 2021.

S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 22588–22600, 2021.

S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

L. Lessard, B. Recht, and A. K. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optim.*, 26(1):57–95, 2016.

M. Liu, A. E. Ozdaglar, T. Yu, and K. Zhang. The power of regularization in solving extensive-form games. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

Q. Liu, C. Szepesvári, and C. Jin. Sample-efficient reinforcement learning of partially observable markov games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Z. Ma, J. Yang, and Z. Zhang. Near-optimal last-iterate convergence of policy optimization in zero-sum polymatrix markov games. 2023.

S. V. Macua, J. Zazo, and S. Zazo. Learning parametric closed-loop policies for markov potential games. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

P. Mahdavinia, Y. Deng, H. Li, and M. Mahdavi. Tight analysis of extra-gradient and optimistic gradient methods for nonconvex minimax problems. *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 35:31213–31225, 2022.

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

P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2703–2717. SIAM, 2018.

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

M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, May 2017.

U. Nadav and T. Roughgarden. The limits of smoothness: A primal-dual framework for price of anarchy bounds. In A. Saberi, editor, *International Workshop On Internet And Network Economics (WINE)*, volume 6484 of *Lecture Notes in Computer Science*, pages 319–326. Springer, 2010.

J. V. Neumann. A model of general economic equilibrium. *The Review of Economic Studies*, 13(1):1–9, 1945.

F. Parise and A. E. Ozdaglar. Graphon games. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 457–458. ACM, 2019.

C. Park, K. Zhang, and A. E. Ozdaglar. Multi-player zero-sum markov games with networked separable interactions, 2023.

T. Parthasarathy and T. E. Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *J. Optim. Theory Appl.*, 33(3):375–392, mar 1981.

S. Pattathil, K. Zhang, and A. E. Ozdaglar. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. In F. J. R. Ruiz, J. G. Dy, and J. van de Meent, editors, *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5641–5685. PMLR, 2023.

J. Perolat, B. Scherrer, B. Piot, and O. Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning (ICML)*, pages 1321–1329. PMLR, 2015.

J. Perolat, B. D. Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S. H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, and K. Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher. Solving stochastic weak minty variational inequalities without increasing batch size, 2023.

G. Piliouras, L. J. Ratliff, R. Sim, and S. Skoulakis. Fast convergence of optimistic gradient ascent in network zero-sum extensive form games. In *Algorithmic Game Theory - 15th International Symposium, SAGT 2022*, volume 13584 of *Lecture Notes in Computer Science*, pages 383–399. Springer, 2022.

B. T. Polyak. Introduction to optimization. 1987.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2005.

T. Qin and S. R. Etesami. Scalable and independent learning of nash equilibrium policies in n-player stochastic games with unknown independent chains, 2023.

S. Qiu, X. Wei, J. Ye, Z. Wang, and Z. Yang. Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8715–8725. PMLR, 2021.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013a.

A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013b.

M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. Decentralized q-learning in zero-sum markov games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18320–18334, 2021.

M. O. Sayin. Decentralized learning for stochastic games: Beyond zero sum and identical interest, 2023.

M. O. Sayin, K. Zhang, and A. E. Ozdaglar. Fictitious play in markov games with single controller. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 919–936. ACM, 2022.

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

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39: 1095–1100, 1953.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 1999.

V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

C. Szepesvári. *Algorithms for reinforcement learning*. Springer Nature, 2022.

D. Vankov, A. Nedić, and L. Sankar. Last iterate convergence of popov method for non-monotone stochastic variational inequalities, 2023.

E. Vlatakis-Gkaragkounis, L. Flokas, T. Lianeas, P. Mertikopoulos, and G. Piliouras. No-regret learning and mixed nash equilibria: They do not mix. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

X. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2002.

Y. Wang, Q. Liu, Y. Bai, and C. Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation, 2023.

C. Wei, C. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 4259–4299. PMLR, 2021.

Y. Yang and C. Ma. $o(t^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

S. Zeng, T. Doan, and J. Romberg. Regularized gradient descent ascent for two-player zero-sum markov games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 34546–34558, 2022.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 5867–5876. PMLR, 2018.

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

K. Zhang, S. Kakade, T. Basar, and L. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1166–1178, 2020.

M. Zhang, P. Zhao, H. Luo, and Z. Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 26772–26808. PMLR, 2022a.

R. Zhang, Q. Liu, H. Wang, C. Xiong, N. Li, and Y. Bai. Policy optimization for markov games: Unified framework and faster convergence. In *NeurIPS*, 2022b.

R. Zhang, J. Mei, B. Dai, D. Schuurmans, and N. Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022c.

Y. Zhou, F. Kong, and S. Li. Stochastic no-regret learning for general games with variance reduction. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

M. Zinkevich, M. Bowling, M. Johanson, and C. Piccione. Regret minimization in games with incomplete information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

A Additional Preliminaries

In this section, we provide some additional preliminaries omitted from the main body.

Average CCE We first give the definition of an average CCE—in the parlance of Nadav and Roughgarden [2010]. ACCE were also studied recently by Zhou et al. [2023] but under the name "weak CCE."

Definition A.1. A stationary (potentially correlated) policy $\boldsymbol{\mu} \in \Delta(\mathcal{A})^{\mathcal{S}}$ is a stationary ϵ-ACCE if

$$\sum_{i=1}^{n} \left(V_i^{\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\mu}}(\boldsymbol{\rho}) \right) \leq \epsilon.$$

Polymatrix zero-sum Markov games We recall that a polymatrix zero-sum Markov game is based on an undirected graph $G = (V, E)$. Each node $i \in V \coloneqq [n]$ is uniquely associated with a player. We will denote by $\mathcal{N}_i \coloneqq \{i' \in V : \{i, i'\} \in E\}$ the neighborhood of player $i \in [n]$. The set of edges E encodes the underlying pairwise interactions, so that the instantaneous reward of player $i \in [n]$ can be expressed as $R_i(s, \boldsymbol{a}) \coloneqq \sum_{i' \in \mathcal{N}_i} R_{i,i'}(s, a_i, a_{i'})$, where $R_{i,i'} : \mathcal{S} \times \mathcal{A}_i \times \mathcal{A}_{i'} \to \mathbb{R}$. Furthermore, it holds that $\sum_{i=1}^n R_i(s, \boldsymbol{a}) = 0$, for any $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, since the game is assumed to be zero-sum. Furthermore, Kalogiannis and Panageas [2023] also assume that in every state $s \in \mathcal{S}$ there is a single player determining the transition probabilities, denoted by cntrl_s. Interestingly, admitting a switching controller—in the sense of the latter assumption—is necessary to guarantee equilibrium collapse, as shown by Kalogiannis and Panageas [2023].

B Omitted Proofs

In this section, we provide the proofs omitted from the main body. Appendix B.1 below contains the proofs from Section 3, while Appendix B.2 establishes our statements from Section 4.

B.1 Proofs from Section 3

We commence here with the proofs from Section 3. We first have to recall the so-called *RVU property*, crystallized by Syrgkanis et al. [2015]; the version we include below was not explicitly stated by Syrgkanis et al. [2015], Rakhlin and Sridharan [2013a], but follows readily from their arguments.

Proposition B.1 (RVU property [Syrgkanis et al., 2015])**.** *Consider a regret minimization algorithm over $\mathcal{Z}_r$ instantiated with (*`OGD`*) parameterized by a learning rate $\eta > 0$. Then, under any sequence of utilities $(\boldsymbol{u}_r^{(t)})_{0 \le t \le T}$, for some time horizon $T \in \mathbb{N}$, its regret can be upper bounded as*

$$\mathsf{Reg}_{\mathcal{Z}_r}^{(T)} \le \frac{D_{\mathcal{Z}_r}^2}{2\eta} + \eta \sum_{t=1}^T \|\boldsymbol{u}_r^{(t)} - \boldsymbol{u}_r^{(t-1)}\|_2^2 - \frac{1}{2\eta} \sum_{t=1}^T \left(\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2^2 \right).$$

We recall the standard definition of regret: $\mathsf{Reg}_{\mathcal{Z}_r}^{(T)} \coloneqq \max_{\boldsymbol{z}_r^\star \in \mathcal{Z}_r} \sum_{t=1}^T \langle \boldsymbol{z}_r^\star - \boldsymbol{z}_r^{(t)}, \boldsymbol{u}_r^{(t)} \rangle$. We also call attention to the fact that (`OGD`) has access to an auxiliary utility $\boldsymbol{u}^{(0)}$, which also appears in the regret bound of Proposition B.1; this is just made for convenience, and it does not affect the analysis.

We will also use the following lemma, which can be extracted in [Anagnostides et al., 2022].

Lemma B.2. *Suppose that the sequences $(\boldsymbol{z}_r^{(t)})_{0 \le t \le T}$ and $(\hat{\boldsymbol{z}}_r^{(t)})_{1 \le t \le T+1}$ are updated by (*`OGD`*) under a sequence of utilities $(\boldsymbol{u}_r^{(t)})_{0 \le t \le T}$. Then, for any $t \in [T]$,*

$$\max_{\boldsymbol{z}_r^\star \in \mathcal{Z}_r} \langle \boldsymbol{z}_r^\star - \boldsymbol{z}_r^{(t)}, \boldsymbol{u}_r^{(t)} \rangle \le \left(\frac{D_{\mathcal{Z}_r}}{\eta} + \max_{1 \le t \le T} \|\boldsymbol{u}_r^{(t)}\|_2 \right) s_r^{(t)},$$

where $s_r^{(t)} \coloneqq \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2$.

We can proceed with the proof of Theorem 3.4, the statement of which is recalled below.

Theorem 3.4. *Let $\mathcal{X} = \times_{r=1}^{d} \mathcal{Z}_r$ for some $d \in \mathbb{N}$ and $F : \mathcal{X} \to \mathcal{X}$ be an L-Lipschitz continuous operator with $B_F \coloneqq \max_{1 \le r \le d} \|F_r\|_2$. Suppose further that the average (α, ℓ, h)-generalized Minty property (Property 3.3) holds. Then, for any $\epsilon > 0$, after $T \ge \frac{2D_{\mathcal{X}}^2 h}{\ell \epsilon^2}$ iterations of (OGD) with learning rate $\eta \le \frac{1}{4}\sqrt{\frac{\ell}{h^3 L^2 + h B_F^2 \alpha^2 d}}$ there is a point $\boldsymbol{x}^{(t)} \in \mathcal{X}$ such that for any $\boldsymbol{x}^\star \in \mathcal{X}$,*

$$\langle \boldsymbol{x}^{(t)}, F(\boldsymbol{x}^{(t)}) \rangle - \langle \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \rangle \le 2d \left(\frac{\max_{1 \le r \le d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_F}{\ell} \right) \epsilon.$$

Proof. We let $\boldsymbol{x}^{(t)} = (\boldsymbol{z}_r^{(t)})_{r=1}^d$ for any $t \in \mathbb{N}^*$ and $\hat{\boldsymbol{x}}^{(t)} = (\hat{\boldsymbol{z}}_r^{(t)})_{r=1}^d$ for any $t \in \mathbb{N}$. We further let $A = (A_1, \dots, A_d)$ and $F = (F_1, \dots, F_d)$. In light of the Cartesian product structure of $\mathcal{X} = \times_{r=1}^{d} \mathcal{Z}_r$, the update rule of (OGD) can be equivalently written as

$$\begin{aligned} \boldsymbol{z}_r^{(t)} &\coloneqq \Pi_{\mathcal{Z}_r}(\hat{\boldsymbol{z}}_r^{(t)} - \eta A_r(\boldsymbol{x}^{(t-1)}) \circ F_r(\boldsymbol{x}^{(t-1)})), \\ \hat{\boldsymbol{z}}_r^{(t+1)} &\coloneqq \Pi_{\mathcal{Z}_r}(\hat{\boldsymbol{z}}_r^{(t)} - \eta A_r(\boldsymbol{x}^{(t)}) \circ F_r(\boldsymbol{x}^{(t)})), \end{aligned} \tag{8}$$

for all times $t \in \mathbb{N}$ and $r \in [d]$. In turn, (8) can be equivalently expressed so that $\boldsymbol{z}_r^{(t)}$ and $\hat{\boldsymbol{z}}_r^{(t)}$ are solutions to the optimization problems

$$\begin{aligned} &\min_{\boldsymbol{z}_r \in \mathcal{Z}_r} a_r(\boldsymbol{x}^{(t-1)}) \langle \boldsymbol{z}_r, F_r(\boldsymbol{x}^{(t-1)}) \rangle + \frac{1}{2\eta} \|\boldsymbol{z}_r - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2, \\ &\min_{\hat{\boldsymbol{z}}_r \in \mathcal{Z}_r} a_r(\boldsymbol{x}^{(t)}) \langle \hat{\boldsymbol{z}}_r, F_r(\boldsymbol{x}^{(t)}) \rangle + \frac{1}{2\eta} \|\hat{\boldsymbol{z}}_r - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2, \end{aligned}$$

respectively. Let us fix a time horizon $T \in \mathbb{N}$ and $r \in [d]$. Invoking Proposition B.1, it follows that for any reference point $\boldsymbol{z}_r^\star \in \mathcal{Z}_r$ the term $\sum_{t=1}^T a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{(t)} - \boldsymbol{z}_r^\star, F_r(\boldsymbol{x}^{(t)}) \rangle$, which can be viewed as the cumulated regret under the sequence of utilities $(F_r(\boldsymbol{x}^{(t)}) \circ A_r(\boldsymbol{x}^{(t)}))_{1 \le t \le T}$, can be upper bounded by

$$\frac{D_{\mathcal{Z}_r}^2}{2\eta} + \eta \sum_{t=1}^T \|a_r(\boldsymbol{x}^{(t)}) F_r(\boldsymbol{x}^{(t)}) - a_r(\boldsymbol{x}^{(t-1)}) F_r(\boldsymbol{x}^{(t-1)})\|_2^2 - \frac{1}{2\eta} \sum_{t=1}^T \left(\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2^2 \right),$$

where we recall that $D_{\mathcal{Z}_r}$ above represents the ℓ_2 diameter of $\mathcal{Z}_r$. Furthermore, using the assumption of Property 3.2 that $0 < \ell \le w_r(\boldsymbol{x}^\star) \le h$, the term $\sum_{t=1}^T w_r(\boldsymbol{x}^\star) a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{(t)} - \boldsymbol{z}_r^\star, F_r(\boldsymbol{x}^{(t)}) \rangle$ can be in turn upper bounded by

$$\frac{D_{\mathcal{Z}_r}^2 h}{2\eta} + \eta h \sum_{t=1}^T \|a_r(\boldsymbol{x}^{(t)}) F_r(\boldsymbol{x}^{(t)}) - a_r(\boldsymbol{x}^{(t-1)}) F_r(\boldsymbol{x}^{(t-1)})\|_2^2 - \frac{\ell}{2\eta} \sum_{t=1}^T \left(\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2^2 \right), \tag{9}$$

for any $\boldsymbol{x}^\star \in \mathcal{X}$. Now, by selecting a suitable $\boldsymbol{z}_r^\star \in \mathcal{Z}_r$ for each $r \in [d]$, Property 3.3 implies that

$$\sum_{t=1}^T \sum_{r=1}^d w_r(\boldsymbol{x}^\star) a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{(t)} - \boldsymbol{z}_r^\star, F_r(\boldsymbol{x}^{(t)}) \rangle = \sum_{t=1}^T \langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \circ A(\boldsymbol{x}^{(t)}) \circ W(\boldsymbol{x}^\star) \rangle \ge 0,$$

for any $t \in [T]$. Thus, by adding (9) for each $r \in [d]$ we have that

$$0 \le \frac{D_{\mathcal{X}}^2 h}{2\eta} + 2\eta h^3 \sum_{t=1}^{T} \|F(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t-1)})\|_2^2 + 2\eta h \alpha^2 d \max_{1\le r\le d} \|F_r\|_2^2 \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2$$
$$- \frac{\ell}{2\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2 \right),$$

where we used the fact that $\sum_{r=1}^{d} D_{\mathcal{Z}_r}^2 = D_{\mathcal{X}}^2$; $\sum_{r=1}^{d} (\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2^2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2^2) = (\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2)$; and that

$$\|a_r(\boldsymbol{x}^{(t)}) F_r(\boldsymbol{x}^{(t)}) - a_r(\boldsymbol{x}^{(t-1)}) F_r(\boldsymbol{x}^{(t-1)})\|_2^2 \le 2h^2 \|F_r(\boldsymbol{x}^{(t)}) - F_r(\boldsymbol{x}^{(t-1)})\|_2^2$$
$$+ 2\|F_r\|_2^2 |a_r(\boldsymbol{x}^{(t)}) - a_r(\boldsymbol{x}^{(t-1)})|^2.$$

Further, using L-Lipschitz continuity of F, we have that

$$0 \le \frac{D_{\mathcal{X}}^2 h}{2\eta} + 2\eta (h^3 L^2 + h B_F^2 \alpha^2 d) \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 - \frac{\ell}{8\eta} \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2$$
$$- \frac{\ell}{4\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2 \right),$$

where we also used the fact that $\sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 \le 2 \sum_{t=1}^{T} (\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2)$. Thus, for $\eta \le \frac{1}{4}\sqrt{\frac{\ell}{h^3 L^2 + h B_F^2 \alpha^2 d}}$, we conclude the following.

Corollary B.3. *Under the conditions of Theorem 3.4,*

$$\sum_{t=1}^{T} \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2 \right) \le \frac{2 D_{\mathcal{X}}^2 h}{\ell}.$$

As a result, for $T \ge \frac{2 D_{\mathcal{X}}^2 h}{\ell \epsilon^2}$ the above inequality implies that there exists $t \in [T]$ such that $\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2, \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2 \le \epsilon$. Using Lemma B.2, it follows that for any $r \in [d]$,

$$a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{(t)}, F_r(\boldsymbol{x}^{(t)}) \rangle - a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{\star}, F_r(\boldsymbol{x}^{(t)}) \rangle$$
$$\le \left(\frac{D_{\mathcal{Z}_r}}{\eta} + h B_F \right) \left(\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2 \right).$$

Given that $a_r(\boldsymbol{x}^{(t)}) \ge \ell > 0$,

$$\langle \boldsymbol{z}_r^{(t)}, F_r(\boldsymbol{x}^{(t)}) \rangle - \langle \boldsymbol{z}_r^{\star}, F_r(\boldsymbol{x}^{(t)}) \rangle \le 2 \left(\frac{D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_F}{\ell} \right) \epsilon.$$

Adding those inequalities for all $r \in [d]$, we conclude that for $T \ge \frac{2 D_{\mathcal{X}}^2 h}{\ell \epsilon^2}$ it holds that

$$\langle \boldsymbol{x}^{(t)}, F(\boldsymbol{x}^{(t)}) \rangle - \langle \boldsymbol{x}^{\star}, F(\boldsymbol{x}^{(t)}) \rangle \le 2d \left(\frac{\max_{1\le r\le d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_F}{\ell} \right) \epsilon,$$

for any $\boldsymbol{x}^{\star} \in \mathcal{X}$. This concludes the proof. □

Remark B.4. Theorem 3.4 is robust to conceding error in the evaluation of $A(\boldsymbol{x})$. More precisely, let us suppose that $\|A(\boldsymbol{x}^{(t)}) - A'(\boldsymbol{x}^{(t)})\|_\infty \leq \epsilon^{(t)}$, for any $t \in [T]^\star$ and a sufficiently small $\epsilon^{(t)} > 0$. Following our proof of Theorem 3.4, it is easy to see that (OGD) under the sequence $(F(\boldsymbol{x}^{(t)}) \circ A'(\boldsymbol{x}^{(t)}))_{0 \leq t \leq T}$ yields an $O_T\left(\frac{1}{\sqrt{T}} + \frac{1}{T}\sum_{t=0}^T \|A(\boldsymbol{x}^{(t)}) - A'(\boldsymbol{x}^{(t)})\|_\infty\right)$-strong solution to the VI problem. As a result, if we can guarantee that $\epsilon^{(t)} \leq \frac{1}{\sqrt{T}}$, for any $t \in [T]^\star$, we recover the same rate as Theorem 3.4.

We next provide a slight extension of Theorem 3.4 under a more general condition than Property 3.3, described below.

Property B.5 (Extension of Property 3.3)**.** *Under the preconditions of Property 3.2 with respect to some triple (α, ℓ, h), we say that the induced VI problem satisfies the average $(\alpha, \ell, h) \in \mathbb{R}^3_{>0}$-generalized Minty property with slackness $\gamma > 0$ if for any sequence $\sigma^{(T)} \coloneqq (\boldsymbol{x}^{(t)})_{1 \leq t \leq T}$ there exists $\mathcal{X} \ni \boldsymbol{x}^\star = \boldsymbol{x}^\star(\sigma^{(T)})$ so that*

$$\frac{1}{T}\sum_{t=1}^T \langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \circ A(\boldsymbol{x}^{(t)}) \circ W(\boldsymbol{x}^\star)\rangle \geq -\gamma.$$

Corollary B.6. *Let $\mathcal{X} = \bigtimes_{r=1}^d \mathcal{Z}_r$ for some $d \in \mathbb{N}$ and $F : \mathcal{X} \to \mathcal{X}$ be an L-Lipschitz continuous operator. Suppose further that the average (α, ℓ, h)-generalized Minty property with slackness $\gamma > 0$ (Property B.5) holds. Then, for any $\epsilon > 0$, after $T \in \mathbb{N}$ iterations of* (OGD) *with learning rate $\eta \leq \frac{1}{4}\sqrt{\frac{\ell}{h^3 L^2 + h B_F^2 \alpha^2 d}}$ there is a point $\boldsymbol{x}^{(t)} \in \mathcal{X}$ such that for any $\boldsymbol{x}^\star \in \mathcal{X}$,*

$$\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)})\rangle \leq 2d\left(\frac{\max_{1 \leq r \leq d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_F}{\ell}\right)\sqrt{\frac{4\eta\gamma}{\ell} + \frac{2D_{\mathcal{X}}^2 h}{\ell T}}.$$

Proof. Similarly to the proof of Corollary B.3, $\sum_{t=1}^T \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2\right) \leq \frac{2D_{\mathcal{X}}^2 h}{\ell} + \frac{4\eta\gamma}{\ell}T$. As a result, we conclude that there is $t \in [T]$ such that

$$\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2, \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2 \leq \sqrt{\frac{2D_{\mathcal{X}}^2 h}{\ell T} + \frac{4\eta\gamma}{\ell}}.$$

The statement then follows from Lemma B.2, similarly to Theorem 3.4. □

Remark B.7. One application of incorporating slackness—per Property B.5 and variants thereof—with independent interest pertains the convergence of competing neural networks in the neural tangent kernel (NTK) regime [Jacot et al., 2018], wherein the optimization landscape behaves as nearly convex-concave. Indeed, online learning techniques readily extend under near convexity [Chen et al., 2023a], a fact that can be leveraged in conjunction with our approach to provide convergence guarantees in that regime.

We next extend our analysis in the presence of noise in the operator, in a sense that will be made precise very shortly. To this end, we first need to state a slight modification of Property B.5; we recall the notation $\textsc{EqGap}(\boldsymbol{x}) : \mathcal{X} \ni \boldsymbol{x} \mapsto \max_{\boldsymbol{x}^\star \in \mathcal{X}} \langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x})\rangle$.

Property B.8. *Under the preconditions of Property 3.2 with respect to some triple (α, ℓ, h), we say that the induced VI problem satisfies the average $(\alpha, \ell, h, \rho) \in \mathbb{R}^4_{>0}$-generalized Minty property if for any sequence $\sigma^{(T)} \coloneqq (\boldsymbol{x}^{(t)})_{1 \le t \le T}$ there exists $\mathcal{X} \ni \boldsymbol{x}^\star = \boldsymbol{x}^\star(\sigma^{(T)})$ so that*

$$\sum_{t=1}^{T} \langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \circ A(\boldsymbol{x}^{(t)}) \circ W(\boldsymbol{x}^\star) \rangle \ge -\rho \sum_{t=1}^{T} (\textsc{EqGap}(\boldsymbol{x}^{(t)}))^2.$$

Corollary B.9. *Let $\mathcal{X} = \bigtimes_{r=1}^{d} \mathcal{Z}_r$ for some $d \in \mathbb{N}$ and $F : \mathcal{X} \to \mathcal{X}$ be an L-Lipschitz continuous operator. Suppose that we instead have access to an operator $F^{\delta,\rho} : \mathcal{X} \to \mathcal{X}$ such that*

1. *$F^{\delta,\rho}$ satisfies the average (α, ℓ, h, ρ)-generalized Minty property (Property B.8); and*
2. *$\|F^{\delta,\rho}(\boldsymbol{x}) - F(\boldsymbol{x})\|_2 \le \delta \cdot \textsc{EqGap}(\boldsymbol{x})$, for any $\boldsymbol{x} \in \mathcal{X}$.*

If the pair (ρ, δ) is such that

$$\rho + 12\eta h^3 \delta^2 \le \frac{\eta \ell^3}{64(\max_{1 \le r \le d} D_{\mathcal{Z}_r} + \eta h B_{F^{\delta,\rho}})^2}$$

and $\delta \le \frac{1}{2D_{\mathcal{X}}}$, with learning rate $\eta \le \frac{1}{4}\sqrt{\frac{\ell}{3h^3 L^2 + h B^2_{F^{\delta,\rho}} \alpha^2 d}}$, then after $T \in \mathbb{N}$ iterations of (OGD) there is a point $\boldsymbol{x}^{(t)} \in \mathcal{X}$ with equilibrium gap $\textsc{EqGap}(\boldsymbol{x}^{(t)})$ upper bounded by

$$\frac{16\eta K^2}{\sqrt{T}\ell} \left(6\eta h^3 \delta^2 \textsc{EqGap}(\boldsymbol{x}^{(0)}) + \frac{D^2_{\mathcal{X}} h}{2\eta} \right),$$

for any $\epsilon > 0$, where K is defined in (10).

Proof. First, using Lemma B.2, it follows that for any $r \in [d]$,

$$a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^{(t)}, F_r^{\delta,\rho}(\boldsymbol{x}^{(t)}) \rangle - a_r(\boldsymbol{x}^{(t)}) \langle \boldsymbol{z}_r^\star, F_r^{\delta,\rho}(\boldsymbol{x}^{(t)}) \rangle \le \left(\frac{\max_{1 \le r \le d} D_{\mathcal{Z}_r}}{\eta} + h B_{F^{\delta,\rho}} \right) \left(\|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t)}\|_2 + \|\boldsymbol{z}_r^{(t)} - \hat{\boldsymbol{z}}_r^{(t+1)}\|_2 \right),$$

in turn implying that

$$\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F^{\delta,\rho}(\boldsymbol{x}^{(t)}) \rangle \le \left(\frac{\max_{1 \le r \le d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_{F^{\delta,\rho}}}{\ell} \right) \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2 \right).$$

Given that, by assumption, $\|F^{\delta,\rho}(\boldsymbol{x}^{(t)}) - F(\boldsymbol{x}^{(t)})\|_2 \le \delta \textsc{EqGap}(\boldsymbol{x}^{(t)})$, it follows that for $\delta \le \frac{1}{2D_{\mathcal{X}}}$,

$$\textsc{EqGap}(\boldsymbol{x}^{(t)}) \le K \left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2 \right),$$

where we have defined

$$K \coloneqq 2 \left(\frac{\max_{1 \le r \le d} D_{\mathcal{Z}_r}}{\eta \ell} + \frac{h B_{F^{\delta,\rho}}}{\ell} \right). \tag{10}$$

Moreover, similarly to the proof of Theorem 3.4, Property B.8 implies that the term $-\rho\sum_{t=1}^{T}(\textsc{EqGap}(\boldsymbol{x}^{(t)}))^2$ can be upper bounded by

$$\frac{D_{\mathcal{X}}^2 h}{2\eta} + 2\eta h^3 \sum_{t=1}^{T} \|F^{\delta,\rho}(\boldsymbol{x}^{(t)}) - F^{\delta,\rho}(\boldsymbol{x}^{(t-1)})\|_2^2 + 2\eta h \alpha^2 d B_{F^{\delta,\rho}}^2 \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 \\ - \frac{\ell}{2\eta}\sum_{t=1}^{T}\left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2\right).$$

Now the term $-\frac{\ell}{2\eta}\sum_{t=1}^{T}\left(\|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}\|_2^2 + \|\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t+1)}\|_2^2\right)$ can be upper bounded by

$$-\frac{\ell}{8\eta}\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 - \frac{\ell}{8\eta K^2}\sum_{t=1}^{T}(\textsc{EqGap}^{(t)})^2,$$

while the term $\sum_{t=1}^{T}\|F^{\delta,\rho}(\boldsymbol{x}^{(t)}) - F^{\delta,\rho}(\boldsymbol{x}^{(t-1)})\|_2^2$ can be upper bounded by

$$3\delta^2 \textsc{EqGap}(\boldsymbol{x}^{(0)})^2 + 6\delta^2\sum_{t=1}^{T}(\textsc{EqGap}(\boldsymbol{x}^{(t)}))^2 + 3L^2\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2.$$

As a result, for $\eta \le \frac{1}{4}\sqrt{\frac{\ell}{3h^3L^2 + hB_{F^{\delta,\rho}}^2\alpha^2 d}}$ and any pair (ρ,δ) such that

$$\rho + 12\eta h^3\delta^2 \le \frac{\ell}{16\eta K^2},$$

we conclude that there is $\boldsymbol{x}^{(t)} \in \mathcal{X}$ with equilibrium gap $\textsc{EqGap}(\boldsymbol{x}^{(t)})$ upper bounded by

$$\frac{16\eta K^2}{\sqrt{T}\ell}\left(6\eta h^3\delta^2\textsc{EqGap}(\boldsymbol{x}^{(0)}) + \frac{D_{\mathcal{X}}^2 h}{2\eta}\right).$$

□

B.2 Proofs from Section 4

In this section, we provide the proofs deferred from Section 4. We first make a remark regarding differentiability of the value function, following [Daskalakis et al., 2020, Remark 1].

Remark B.10 (Differentiability). Under direct parameterization, the interior of the joint strategy space, denoted by $\text{int}(\mathcal{X})$, is empty. To make sure that the gradient $\nabla_{\boldsymbol{x}_i}V_i(\boldsymbol{\rho})$ is well-defined, we can instead consider a suitable compact and convex set $\mathcal{X}_\delta$, for any $\delta > 0$, so that $\mathcal{X} \subseteq \text{int}(\mathcal{X}_\delta)$ and any point $\mathcal{X}$ is within distance δ from some point in $\mathcal{X}_\delta$. Using continuity and compactness, it is direct to see that by taking the limit $\delta \downarrow 0$ our analysis readily applies.

Now, following the approach of Kalogiannis and Panageas [2023], we prove Proposition 4.3.

Proposition 4.3. *Assumption 4.2 is satisfied for any polymatrix zero-sum Markov game.*

Proof. We consider for each player $i \in [n]$ the following nonlinear program with variables $\boldsymbol{v}_i \in \mathbb{R}^{\mathcal{S}}$ and $\boldsymbol{\mu} \in \Delta(\mathcal{A})^{\mathcal{S}}$.

$$\begin{aligned} &\min \boldsymbol{\rho}^\top \boldsymbol{v}_i \\ &\text{s.t. } \boldsymbol{v}_i[s] \geq \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\mu}_{-i}(\cdot|s)}[R_i(s, \boldsymbol{a}) + \bar{\zeta}_{s,\boldsymbol{a}} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i], \end{aligned}$$

where $\bar{\zeta}_{s,\boldsymbol{a}} \coloneqq 1 - \zeta_{s,\boldsymbol{a}}$, for all $s \in \mathcal{S}$ and $a_i \in \mathcal{A}_i$. In particular, $\boldsymbol{\mu}$ above represents a stationary, potentially correlated joint policy. Now let us fix a player $i \in [n]$ and $\boldsymbol{\mu}_{-i} \in \Delta(\mathcal{A}_{-i})^{\mathcal{S}}$. It is well-known that the induced linear program is feasible, and the optimal objective is equal to the value of player $i \in [n]$ when best responding to $\boldsymbol{\mu}_{-i}$ [Puterman, 2005]. That is, if the optimal value is attained at $\boldsymbol{v}_i^\star(\boldsymbol{\mu}_{-i}) \in \mathbb{R}^{\mathcal{S}}$, it holds that $V_i^{\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho}) = \boldsymbol{\rho}^\top \boldsymbol{v}_i^\star(\boldsymbol{\mu}_{-i})$, for all $i \in [n]$. In particular, if $\boldsymbol{\mu}(\cdot|s) \coloneqq \boldsymbol{\pi}^\star(\cdot|s)$ is a Nash equilibrium policy (Definition 2.1), it holds that $\sum_{i=1}^n V_i^{\dagger, \boldsymbol{\pi}^\star_{-i}}(\boldsymbol{\rho}) = \sum_{i=1}^n V_i^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) = 0$. This in turn implies that the sum of the objectives (over the players) of the original nonlinear programs is nonpositive. Furthermore, let us fix $\boldsymbol{\pi}(\cdot|s)$ to be a product distribution. By feasibility, it follows that for any state $s \in \mathcal{S}$,

$$\boldsymbol{v}_i[s] \geq \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}(\cdot|s)}[R_i(s, \boldsymbol{a}) + \bar{\zeta}_{s,\boldsymbol{a}} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i].$$

Thus, using the fact that $\sum_{i=1}^n R_i(s, \boldsymbol{a}) = 0$ and that $\zeta_{s,\boldsymbol{a}} > 0$ for all $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, it follows that $\sum_{i=1}^n \boldsymbol{v}_i[s] \geq 0$ for all $s \in \mathcal{S}$. So, it follows that when restricting $\boldsymbol{\mu}$ to be a product policy the sum of the objective values is 0.

Now, consider a potentially correlated policy $\boldsymbol{\mu}$, and let $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\mu})$ be the product distribution induced by taking the marginals of $\boldsymbol{\mu}$. By feasibility, for any player $i \in [n]$ and $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$,

$$\boldsymbol{v}_i^\star[s] \geq \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\mu}_{-i}(\cdot|s)}[R_i(s, \boldsymbol{a}) + \bar{\zeta}_{s,\boldsymbol{a}} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i^\star]. \tag{11}$$

By the assumption of separability of the reward function, the first term in the right-hand side of (11) is equal to

$$\sum_{i' \in \mathcal{N}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\mu}_{-i}(\cdot|s)} R_{i,i'}(s, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)} R_i(s, \boldsymbol{a}).$$

Further, by the assumption of having a switching controller, the second term in the right-hand side of (11) is equal to

$$\mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\mu}_{-i}(\cdot|s)} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i^\star = \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i^\star.$$

Thus, we conclude that

$$\boldsymbol{v}_i^\star[s] \geq \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)}[R_i(s, \boldsymbol{a}) + \bar{\zeta}_{s,\boldsymbol{a}} \mathbb{P}(\cdot|s, \boldsymbol{a}) \boldsymbol{v}_i^\star],$$

which means that the pair $(\boldsymbol{v}_i^\star, \boldsymbol{\pi}(\boldsymbol{\mu}))$ constitutes a feasible solution. Given that $\boldsymbol{\pi}(\boldsymbol{\mu})$ is by definition a product distribution, we know that $\sum_{i=1}^n \boldsymbol{\rho}^\top \boldsymbol{v}_i^\star \geq 0$, in turn implying that $\sum_{i=1}^n V_i^{\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho}) \geq 0$. Finally, Assumption 4.2 follows by taking $\boldsymbol{\mu}$ to be a uniform mixture of T product distributions. □

Beyond polymatrix zero-sum Markov games, Assumption 4.2 is satisfied in all games exhibiting equilibrium collapse per Definition 4.4, as we observe next.

Proposition 4.5. *Assumption 4.2 is satisfied in any Markov game $\mathcal{G}$ exhibiting equilibrium collapse per Definition 4.4.*

Proof. For the sake of contradiction, suppose that there exists a sequence of product joint policies $(\boldsymbol{\pi}^{(1)}, \dots, \boldsymbol{\pi}^{(T)})$ such that

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} V_i^{\dagger, \boldsymbol{\pi}_{-i}^{(t)}}(\boldsymbol{\rho}) - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} V_i^{\boldsymbol{\pi}^{(t)}}(\boldsymbol{\rho}) < 0; \tag{12}$$

that is, Assumption 4.2 is violated. If $\boldsymbol{\mu} \in \Delta(\mathcal{A})^{\mathcal{S}}$ represents the uniform mixture over $(\boldsymbol{\pi}^{(1)}, \dots, \boldsymbol{\pi}^{(T)})$, which is potentially a correlated policy, (12) can be rewritten as

$$\sum_{i=1}^{n}\left(V_i^{\dagger, \boldsymbol{\mu}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\mu}}(\boldsymbol{\rho})\right) < 0.$$

In words, $\boldsymbol{\mu}$ constitutes an ϵ-ACCE (Definition A.1) with $\epsilon < 0$. By the assumption that $\mathcal{G}$ exhibits equilibrium collapse (Definition 4.4), it follows that the marginals of $\boldsymbol{\mu}$ induce an ϵ'-Nash equilibrium with $\epsilon' < 0$, which is a contradiction. This completes the proof. □

We next state a number of elementary properties in MDPs [Cai et al., 2020], starting from the connection between the gradient of the value function and the Q function; for completeness, we also provide their proofs.

Lemma B.11. *For any state $s \in \mathcal{S}$ and any joint action profile $(a_i, \boldsymbol{a}_{-i}) = \boldsymbol{a} \in \mathcal{A}$,*

$$\frac{\partial V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})}{\partial \boldsymbol{x}_{i,s}[a_i]} = \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s]\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_{-i}(\cdot|s)}[Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})], \forall i \in [n].$$

Proof. Let $s_0 \in \mathcal{S}$ be any initial state. We have that the gradient $\nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(s_0)$ is equal to

$$\nabla_{\boldsymbol{x}_i}\left(\sum_{a_{i,0}\in\mathcal{A}_i} \boldsymbol{\pi}_i(a_{i,0}|s_0)\mathbb{E}_{\boldsymbol{a}_{-i,0}\sim\boldsymbol{\pi}_{-i}(\cdot|s_0)}[Q_i^{\boldsymbol{\pi}}(s_0, \boldsymbol{a}_0)]\right),$$

where we denoted by $\boldsymbol{a}_0 \coloneqq (a_{1,0}, \dots, a_{n,0})$. The above display is in turn equal to

$$\sum_{a_{i,0}\in\mathcal{A}_i} \boldsymbol{\pi}_i(a_{i,0}|s_0)(\nabla_{\boldsymbol{x}_i} \log \boldsymbol{\pi}_i(a_{i,0}|s_0))\mathbb{E}[Q_i^{\boldsymbol{\pi}}(s_0, \boldsymbol{a}_0)]$$
$$+ \sum_{a_{i,0}\in\mathcal{A}_i} \boldsymbol{\pi}_i(a_{i,0}|s_0)\mathbb{E}\left[\sum_{s_1\in\mathcal{S}} \mathbb{P}(s_1|s_0, \boldsymbol{a}_0)\nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(s_1)\right],$$

where the expectation is taken with respect to $\boldsymbol{a}_{-i,0} \sim \boldsymbol{\pi}_{-i}(\cdot|s_0)$. In particular, the second term above follows from the fact that

$$Q_i^{\boldsymbol{\pi}}(s_0, \boldsymbol{a}_0) = \sum_{s_1\in\mathcal{S}} \mathbb{P}(s_1|s_0, \boldsymbol{a}_0)V_i^{\boldsymbol{\pi}}(s_1).$$

As a result, it follows that the gradient $\nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(s_0)$ can be expressed as

$$\mathbb{E}_{\mathbb{P}^{\boldsymbol{\pi}}(\cdot|s_0)}[\nabla_{\boldsymbol{x}_i} \log \boldsymbol{\pi}_i(a_{i,0}|s_0)) Q_i^{\boldsymbol{\pi}}(s_0, \boldsymbol{a}_0)] + \mathbb{E}_{\mathbb{P}^{\boldsymbol{\pi}}(\cdot|s_0)}\left[\mathbb{1}_{H \geq 1} \nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(s_0)\right].$$

By linearity, the same holds by replacing the initial state $s_0 \in \mathcal{S}$ with any distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$. As a result, by induction and the fact that $\zeta > 0$, we conclude that

$$\nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho}) = \mathbb{E}\left[\sum_{h=0}^{H} \nabla_{\boldsymbol{x}_i} \log \boldsymbol{\pi}_i(a_{i,h}|s_h)) Q_i^{\boldsymbol{\pi}}(s_h, \boldsymbol{a}_h)\right].$$

The above expression is also equal to

$$\sum_{s \in \mathcal{S}} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}(\cdot|s)}[\nabla_{\boldsymbol{x}_i} \log \boldsymbol{\pi}_i(a_i|s)) Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})].$$

The statement of the lemma thus follows by the fact that $\frac{\partial \log \boldsymbol{\pi}_i(a_i|s))}{\partial \boldsymbol{x}_{i,s'}[a_i']} = \frac{1}{\boldsymbol{x}_{i,s}[a_i]}$ if $(s, a_i) = (s', a_i')$, and 0 otherwise. □

Lemma B.12 (Value difference)**.** *For any joint policy $\boldsymbol{\pi} \in \Pi$ and policy $\boldsymbol{\pi}_i' \in \Pi_i$, the value difference $V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})$ is equal to*

$$\sum_{s \in \mathcal{S}} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}[s] \sum_{a_i \in \mathcal{A}_i} (\boldsymbol{x}_{i,s}'[a_i] - \boldsymbol{x}_{i,s}[a_i]) \mathbb{E}[Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})],$$

for any player $i \in [n]$, where the expectation above is taken over $\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)$.

Proof. Let $s \in \mathcal{S}$. We see that the value difference $V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(s) - V_i^{\boldsymbol{\pi}}(s)$ can be expressed as

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=0}^{H} R_i(s_h, \boldsymbol{a}_h)\right] - V_i^{\boldsymbol{\pi}}(s) &= \mathbb{E}\left[\sum_{h=0}^{H} R_i(s_h, \boldsymbol{a}_h) + \mathbb{1}_{h+1 \leq H} V_i^{\boldsymbol{\pi}}(s_h) - V_i^{\boldsymbol{\pi}}(s_h)\right] \\
&= \mathbb{E}\left[\sum_{h=0}^{H} (Q_i^{\boldsymbol{\pi}}(s_h, \boldsymbol{a}_h) - V_i^{\boldsymbol{\pi}}(s_h))\right],
\end{aligned}$$

where the expectation above is taken over trajectories induced by $\mathbb{P}^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}$. As a result, the value difference $V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(s) - V_i^{\boldsymbol{\pi}}(s)$ is equal to

$$\sum_{s' \in \mathcal{S}} \tilde{d}_s^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}[s'] \mathbb{E}_{\boldsymbol{a} \sim (\boldsymbol{\pi}_i'(\cdot|s'), \boldsymbol{\pi}_{-i}(\cdot|s'))}\left[Q_i^{\boldsymbol{\pi}}(s', \boldsymbol{a}) - V_i^{\boldsymbol{\pi}}(s')\right],$$

which leads to the conclusion of the lemma by taking the expectation $\mathbb{E}_{s \sim \boldsymbol{\rho}}[\cdot]$. □

We now combine Lemmas B.11 and B.12 to conclude Lemma 4.6, the statement of which is recalled below.

Lemma 4.6. *Consider a Markov game $\mathcal{G}$, and let $\Lambda_i(\boldsymbol{x}, \boldsymbol{x}^\star)[s, a_i] \coloneqq \frac{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}[s]}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s]}$ for $i \in [n]$ and $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$. Further, let $\Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \coloneqq (\Lambda_1(\boldsymbol{x}, \boldsymbol{x}^\star), \ldots, \Lambda_n(\boldsymbol{x}, \boldsymbol{x}^\star))$. If Assumption 4.1 holds, then there exists $\boldsymbol{x}^\star \in \mathcal{X}$ such that*

$$\langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \circ \Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \rangle \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

In particular, if $\mathcal{G}$ admits a single controller, denoted by $\mathsf{cntrl}_{\mathcal{G}}$*, then Property 3.2 holds with*

$$A_i(\boldsymbol{x})[s, a_i] \coloneqq \begin{cases} 1 & : \textit{if } i \neq \mathsf{cntrl}_{\mathcal{G}} \\ \left(\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i}[s]\right)^{-1} & : \textit{if } i = \mathsf{cntrl}_{\mathcal{G}}, \end{cases}$$

and

$$W_i(\boldsymbol{x}^\star)[s, a_i] \coloneqq \begin{cases} 1 & : \textit{if } i \neq \mathsf{cntrl}_{\mathcal{G}} \\ \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star}[s] & : \textit{if } i = \mathsf{cntrl}_{\mathcal{G}}. \end{cases}$$

Proof. Let us consider a player $i \in [n]$. For $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, Lemma B.11 implies that

$$\Lambda_i(\boldsymbol{x}, \boldsymbol{x}^\star)[s, a_i] \frac{\partial V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})}{\partial \boldsymbol{x}_{i,s}[a_i]} = \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}(s) \mathbb{E}[Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})],$$

where the expectation above is taken over $\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)$. Thus, summing over all $s \in \mathcal{S}$ and $a_i \in \mathcal{A}_i$, it follows that the term $\langle \boldsymbol{x}_i^\star - \boldsymbol{x}_i, \nabla_{\boldsymbol{x}_i} V_i(\boldsymbol{\rho}) \circ \Lambda_i(\boldsymbol{x}, \boldsymbol{x}^\star) \rangle$ is equal to

$$\sum_{s \in \mathcal{S}} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}[s] \sum_{a_i \in \mathcal{A}_i} (\boldsymbol{x}_{i,s}'[a_i] - \boldsymbol{x}_{i,s}[a_i]) \mathbb{E}[Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})],$$

where the expectation above is again taken over $\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)$. By Lemma B.12, the term above is equal to the value difference $V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})$. Thus, summing over all players $i \in [n]$, we find that (5) is a consequence of Assumption 4.1. Analogously, (6) is a consequence of Assumption 4.2. $\square$

Remark B.13 (Greedy exploration). Throughout this paper, we have been operating in the regime of direct parameterization in that $\boldsymbol{\pi}_i(a_i|s) \coloneqq \boldsymbol{x}_{i,s}[a_i]$, for any player $i \in [n]$ and $(s, a_i) \in \mathcal{X} \times \mathcal{A}_i$. This type of parameterization suffices under the assumption that players have complete gradient feedback, but in the more challenging bandit feedback model such a parameterization could cause the variance of the gradient estimator to blow up. One common approach to address this issue consists of incorporating γ-greedy exploration, so that now $\boldsymbol{\pi}_i(a_i|s) \coloneqq (1 - \gamma)\boldsymbol{x}_{i,s}[a_i] + \gamma/|\mathcal{A}_i|$. This in turn leads to variance bounded by $O_\gamma(1/\gamma)$. It is fairly straightforward to see that Lemma 4.6 still implies Property 3.3, with the difference that the right-hand side of (6) is replaced by a term $-\Theta_\gamma(\gamma)T$. By virtue of Corollary B.6, analogous conclusions hold in that case as well.

In Corollary 3.5, we saw that the guarantee of Theorem 3.4 in general smooth multi-player games yields only a local optimality guarantee; to arrive at global optimality, as claimed in Theorem 4.7, we will show that the gradient dominance property (Item 2) holds; the proof below follows that of [Daskalakis et al., 2020, Lemma 1].

Lemma B.14 (Gradient dominance)**.** *Let $\boldsymbol{\pi} \in \Pi$ and $\boldsymbol{\pi}_i' \in \Pi_i$, for some player $i \in [n]$. Then, the value difference $\max_{\boldsymbol{\pi}_i' \in \Pi_i} V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})$ is upper bounded by*

$$\min_{\boldsymbol{\pi}_i^\star \in \Pi_i^\star(\boldsymbol{\pi}_{-i})} \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}}{\boldsymbol{\rho}} \right\|_\infty \frac{1}{\zeta} \max_{\boldsymbol{x}_i' \in \mathcal{X}_i} \langle \boldsymbol{x}_i' - \boldsymbol{x}, \nabla_{\boldsymbol{x}_i} V_i(\boldsymbol{\rho}) \rangle.$$

Before we proceed with the proof, let us point out that the ratio above is defined coordinate-wise, which is well-defined since we have assumed that $\boldsymbol{\rho}$ has full support. We also note that the nomenclature $\Pi_i^\star(\boldsymbol{\pi}_{-i})$ above denotes the set of (stationary) best response policies of player $i \in [n]$ under $\boldsymbol{\pi}_{-i}$.

Proof of Lemma B.14. By Lemma B.12, we have that the value difference $\max_{\boldsymbol{\pi}_i' \in \Pi_i} V_i^{\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho})$ can be upper bounded by

$$\sum_{s \in \mathcal{S}} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}[s] \max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)} [V_i^{\boldsymbol{\pi}}(s) - Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})],$$

where $\boldsymbol{\pi}_i^\star(\boldsymbol{\pi}_{-i}) \in \Pi_i^\star(\boldsymbol{\pi}_{-i})$ is a policy minimizing $\left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}}{\boldsymbol{\rho}} \right\|_\infty$. Since $\max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)} [V_i^{\boldsymbol{\pi}}(s) - Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})] \geq 0$ for any state $s \in \mathcal{S}$, the last displayed term can be in turn upper bounded by

$$\left\| \frac{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}} \right\|_\infty \sum_{s \in \mathcal{S}} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] \max_{a_i \in \mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)} [V_i^{\boldsymbol{\pi}}(s) - Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})].$$

Moreover, the first term above can be bounded as

$$\left\| \frac{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}} \right\|_\infty \leq \frac{1}{\zeta} \left\| \frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_{-i}}}{\boldsymbol{\rho}} \right\|_\infty,$$

while the second term is equal to

$$\begin{aligned} &\max_{\boldsymbol{x}_i' \in \mathcal{X}_i} \sum_{s \in \mathcal{S}} \sum_{a_i \in \mathcal{A}_i} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] \boldsymbol{x}_{i,s}'[a_i] \mathbb{E}[V_i^{\boldsymbol{\pi}}(s) - Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})] \\ &\qquad = \max_{\boldsymbol{x}_i' \in \mathcal{X}_i} \sum_{s \in \mathcal{S}} \sum_{a_i \in \mathcal{A}_i} \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] (\boldsymbol{x}_{i,s}[a_i] - \boldsymbol{x}_{i,s}'[a_i]) \mathbb{E}[Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})], \end{aligned}$$

where the expectation is taken over $\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}_{-i}(\cdot|s)$. By Lemma B.11, the last term can be recognized as $\max_{\boldsymbol{x}_i' \in \mathcal{X}_i} \langle \boldsymbol{x}_i - \boldsymbol{x}_i', \nabla_{\boldsymbol{x}_i} V_i(\boldsymbol{\rho}) \rangle$, concluding the proof. □

Finally, to conclude Theorem 3.4 using Theorem 4.7, we appropriately bound all the involved parameters. We first point out a standard bound on the smoothness of the value function.

Lemma B.15. *For any joint policies $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$,*

$$\|\nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}}(\boldsymbol{\rho}) - \nabla_{\boldsymbol{x}_i} V_i^{\boldsymbol{\pi}'}(\boldsymbol{\rho})\|_2 \leq \frac{4|\mathcal{A}_i|}{\zeta^3} \|\boldsymbol{x} - \boldsymbol{x}'\|_2,$$

for any player $i \in [n]$.

Theorem 4.7. *Let $\mathcal{G}$ be a Markov game that satisfies Assumption 4.2 and admits a single controller. Then,* (OGD) *after $1/\epsilon^2 \cdot \mathsf{poly}(n, \sum_{i=1}^n |\mathcal{A}_i|, |\mathcal{S}|, 1/\zeta, C_{\mathcal{G}}, 1/\|\boldsymbol{\rho}\|_\infty)$ iterations computes a stationary ϵ-NE.*

Proof. In light of Lemma 4.6, we will apply Theorem 3.4 with the following parameters:

- $\mathcal{X}_i \coloneqq \Delta(\mathcal{A}_i)^{\mathcal{S}}$ and $\mathcal{X} \coloneqq \times_{i=1}^n \mathcal{X}_i$. As such, we have that $\mathcal{X} = \times_{i\in[n],s\in\mathcal{S}} \mathcal{Z}_{i,s}$ with $\mathcal{Z}_{i,s} = \Delta(\mathcal{A}_i)$;
- $d \coloneqq n|\mathcal{S}|$;
- $D_{\mathcal{X}}^2 = 2n|\mathcal{S}|$;
- $h \coloneqq \max\left\{\frac{1}{\zeta}, \frac{1}{\|\boldsymbol{\rho}\|_\infty}\right\}$ and $\ell \coloneqq \min\{\zeta, \|\boldsymbol{\rho}\|_\infty\}$. This follows given that for any joint policy $\boldsymbol{\pi} \in \Pi$ it holds that $\tilde{d}_{s_0}^{\boldsymbol{\pi}}[s] = \sum_{h\in\mathbb{N}^*} \mathbb{P}^{\boldsymbol{\pi}}(s_h = s|s_0) \leq \sum_{h=0}^{\infty}(1-\zeta)^h \leq \frac{1}{\zeta}$ and that $\tilde{d}_{s_0}^{\boldsymbol{\pi}}[s] \geq \mathbb{P}^{\boldsymbol{\pi}}(s_0 = s|s_0)$, for any $(s, s_0) \in \mathcal{S}\times\mathcal{S}$, in turn implying that $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] \leq \frac{1}{\zeta}$ and $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] \geq \boldsymbol{\rho}[s]$, for any $s \in \mathcal{S}$. Hence, the claimed bounds on ℓ and h follow directly by virtue of Lemma 4.6;
- $L \coloneqq \frac{4\sqrt{\sum_{i=1}^n |\mathcal{A}_i|^2}}{\zeta^3}$. Indeed, having taken $F(\boldsymbol{x}) \coloneqq -(\nabla_{\boldsymbol{x}_1} V_1(\boldsymbol{\rho}), \dots, \nabla_{\boldsymbol{x}_n} V_n(\boldsymbol{\rho}))$, the claimed bound on the Lipschitz continuity of F follows directly by Lemma B.15;
- $B_F \coloneqq \frac{\max_{1\leq i\leq n}\sqrt{|\mathcal{A}_i|}}{\zeta^2}$. This follows given that, by Lemma B.11, $\|F_{i,s}\|_\infty \leq \frac{1}{\zeta^2}$, for any $(i,s) \in [n]\times\mathcal{S}$, in turn implying that $\|F_{i,s}\|_2 \leq \frac{\sqrt{|\mathcal{A}_i|}}{\zeta^2}$; and
- $\alpha \coloneqq \frac{\sqrt{|\mathcal{S}|\sum_{i=1}^n |\mathcal{A}_i|}}{\zeta^2\|\boldsymbol{\rho}\|_\infty^2}$. Indeed, for any two joint policies $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \Pi$ it holds that

  $$\left|\frac{1}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s]} - \frac{1}{\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}'}[s]}\right| \leq \frac{1}{\|\boldsymbol{\rho}\|_\infty^2}|\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] - \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}'}[s]|,$$

  for any $s \in \mathcal{S}$. Let us fix the state $s \in \mathcal{S}$. To bound the term $|\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] - \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}'}[s]|$, we consider a fictitious Markov game $\tilde{\mathcal{G}}$ such that for any player $i \in [n]$ the reward is defined so that $\tilde{R}_i(s', \boldsymbol{a}) = 1$ if $s' = s$, and 0 otherwise. Then, it follows that the value function takes the form $\tilde{V}_i^{\boldsymbol{\pi}}(\boldsymbol{\rho}) = \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s]$, for any player $i \in [n]$ and joint policy $\boldsymbol{\pi} \in \Pi$. Thus, the term $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] - \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}'}[s]$ is equal to $\tilde{V}_1^{\boldsymbol{\pi}}(\boldsymbol{\rho}) - \tilde{V}_1^{\boldsymbol{\pi}_1', \boldsymbol{\pi}_{-1}}(\boldsymbol{\rho}) + \cdots + \tilde{V}_n^{\boldsymbol{\pi}_n, \boldsymbol{\pi}_{-n}'}(\boldsymbol{\rho}) - \tilde{V}_n^{\boldsymbol{\pi}'}(\boldsymbol{\rho})$, and in turn Lemma B.12 yields that

  $$\begin{aligned}|\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}}[s] - \tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}'}[s]| &\leq \frac{1}{\zeta^2}\sum_{i=1}^n \sum_{s\in\mathcal{S}} \sum_{a_i\in\mathcal{A}_i} |\boldsymbol{x}_{i,s}[a_i] - \boldsymbol{x}_{i,s}'[a_i]| \\ &= \frac{1}{\zeta^2}\|\boldsymbol{x} - \boldsymbol{x}'\|_1.\end{aligned}$$

  The conclusion then follows from the equivalence between the ℓ_1 and the ℓ_2 norms.

As a result, Theorem 3.4 along with Lemma 4.6 imply that after a sufficiently large number of iterations $T = \mathsf{poly}(n, \sum_{i=1}^n |\mathcal{A}_i|, |\mathcal{S}|, 1/\zeta, 1/\|\boldsymbol{\rho}\|_\infty) \cdot 1/\epsilon^2$, we have computed a point $\boldsymbol{x}^{(t)}$ such that

$$\langle \boldsymbol{x}^{(t)}, F(\boldsymbol{x}^{(t)})\rangle - \min_{\boldsymbol{x}^\star\in\mathcal{X}}\langle \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)})\rangle \leq \epsilon.$$

Finally, Lemma B.14 implies that

$$\sum_{i=1}^{n}\left(V_i^{\boldsymbol{\pi}^{(t)}}(\boldsymbol{\rho}) - V_i^{\dagger,\boldsymbol{\pi}_{-i}^{(t)}}(\boldsymbol{\rho})\right) \geq -G\epsilon,$$

where $G \coloneqq \frac{C_{\mathcal{G}}}{\zeta}$, in accordance to Lemma B.14. In particular, here we defined $C_{\mathcal{G}}$ as

$$\max_{1\leq i\leq n}\max_{\boldsymbol{\pi}_{-i}\in\Pi_{-i}}\left\{\min_{\boldsymbol{\pi}_i^\star\in\Pi_i^\star(\boldsymbol{\pi}_{-i})}\left\|\frac{d_{\boldsymbol{\rho}}^{\boldsymbol{\pi}_i^\star,\boldsymbol{\pi}_{-i}}}{\boldsymbol{\rho}}\right\|_\infty\right\}. \tag{13}$$

Thus, rescaling $\epsilon' \coloneqq G\epsilon$ concludes the proof. □

C Illustrative Experiments

Our main result concerns the behavior of (OGD) under a time-varying but non-vanishing learning rate—captured by the term $A(\boldsymbol{x})$ in the update rule of (OGD). In this section, we present some illustrative experiments that juxtapose the performance of the variant we analyze and the standard optimistic gradient descent algorithm under a constant learning rate $\eta > 0$.

Specifically, we conduct experiments on the ratio game (3), where $\mathbf{R} \in \mathbb{R}^{100\times 120}$ and $\mathbf{S} \coloneqq \boldsymbol{s} \otimes \mathbf{1}_{120}$ for some $\boldsymbol{s} \in \mathbb{R}^{100}$. Each entry of $\mathbf{R}$ and $\boldsymbol{s}$ are selected uniformly at random from $(0,1)$. We execute each algorithm for 10^3 iterations with $\eta \coloneqq 0.1$. The results for 9 different random realizations are illustrated in Figure 1. Overall, we see that the two algorithms attain similar performance, although there are no theoretical guarantees for the performance of (OGD) with a constant learning rate.

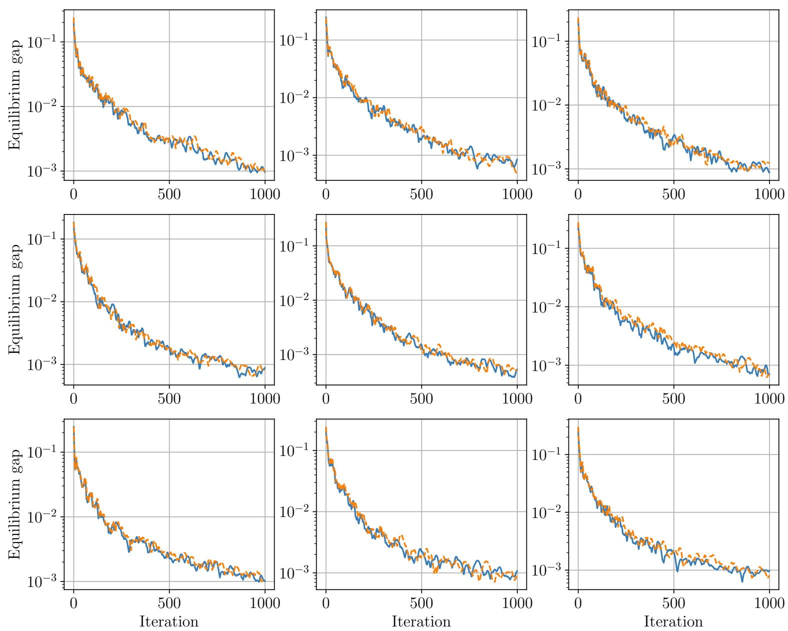

Figure 1: Optimistic gradient descent with a constant learning rate $\eta \coloneqq 0.1$ (blue curve) versus optimistic gradient descent with a time-varying learning rate per (OGD) (orange curve). Each figure corresponds to a separate random ratio game. The equilibrium gap of a joint strategy $\boldsymbol{x} \in \mathcal{X}$ is defined as $\max_{\boldsymbol{x}^\star \in \mathcal{X}} \langle \boldsymbol{x} - \boldsymbol{x}^\star, F(\boldsymbol{x}) \rangle$.