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], **Tuomas Sandholm**[1, 4, 5, 6]

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

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.

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), as well as popular multi-player games (Brown and Sandholm 2019). 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, Yang, and Basar 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 the sequel 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, Golowich, and Zhang 2023; Jin, Muthukumar, and Sidford 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, Goldberg, and Papadimitriou 2006). 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 aforedescribed considerations—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, Papadimitriou, and Piliouras 2018), 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, Oikonomou, and Zheng 2022; Gorbunov, Taylor, and Gidel 2022; Golowich et al. 2020), but its role is not well-

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

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.

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, Foster, and Golowich (2020) observed that the Minty property fails even in simple two-player Markov games with a single controller (recalled in Proposition 3). Furthermore, although several relaxations of the Minty property have been proposed, 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 4) 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; this modification already suffices to subsume the counterexample of Daskalakis, Foster, and Golowich (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 5.

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 6), where the notation $O_\epsilon(\cdot)$ here suppresses polynomial factors in all natural parameters of the problem. In the full version, we further establish that this guarantee is robust in the presence of perturbations akin to *relative deterministic noise*—a ubiquitous model in control theory and optimization—and a certain slackness in our condition; 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 the sequel. 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 11). It is well-known that such is the case in two-player zero-sum games, but recent work (Kalogiannis and Panageas 2023; Park, Zhang, and Ozdaglar 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, Golowich, and Zhang 2023; Jin, Muthukumar, and Sidford 2023). For this reason, we further posit that the game admits a *single* controller, a quite classical setting in the literature. The upshot now is that under those two assumptions, our condition that generalizes the Minty property holds (Lemma 13), which brings us to one of our main results.

Theorem 1 (Informal; precise version in Theorem 14)**.** *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 14. In light of existing hardness results for computing stationary NE even in turn-based two-player Markov games (Daskalakis, Golowich, and Zhang 2023; Jin, Muthukumar, and Sidford 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.

Preliminaries on Markov Games

In this section, we provide the necessary preliminaries on Markov games.

Notation We let $\mathbb{N} = \{1, 2, \dots, \}$ denote the set of natural numbers and $\mathbb{N}^* := \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$, we use the shorthand notations $[\![n]\!] := \{1, \dots, n\}$ and $[\![n]\!]^* := \{0, 1, \dots, n\}$. 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 $\boldsymbol{z} \leq \cdot$ 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: $(\boldsymbol{z} \circ \boldsymbol{z}')[r] := \boldsymbol{z}[r] \cdot \boldsymbol{z}'[r]$, 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 $\|F(\boldsymbol{x}) - F(\boldsymbol{x}')\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$, 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 $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \mathbb{P}, \{R_i\}_{i=1}^n, \zeta, \boldsymbol{\rho}) =: \mathcal{G}$, whose constituents are defined as follows.

- $\mathcal{N} := [\![n]\!]$ 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 $\mathcal{A} \coloneqq \bigtimes_{i=1}^n \mathcal{A}_i$;
- $\mathbb{P}$ 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 $\boldsymbol{\pi}_i \mapsto \boldsymbol{x}_i \in \Delta(\mathcal{A}_i)^{\mathcal{S}} =: \mathcal{X}_i$ 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 $\Pi \coloneqq \bigtimes_{i=1}^n \Pi_i$. We will also let $\mathcal{X} \coloneqq \bigtimes_{i=1}^n \mathcal{X}_i$.

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 $\boldsymbol{\pi} \coloneqq (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n) \in \Pi$:

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

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]

Definition 2. A (stationary) *product* policy $\boldsymbol{\pi}^\star \in \Pi$ is an ϵ-Nash equilibrium if $\max_i \left\{V_i^{\dagger,\boldsymbol{\pi}^\star_{-i}}(\boldsymbol{\rho}) - V_i^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho})\right\} \leq \epsilon$.

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 $\tilde{d}_{\boldsymbol{\rho}}^{\boldsymbol{\pi}} \coloneqq \mathbb{E}_{s_0\sim\boldsymbol{\rho}}[\tilde{d}_{s_0}^{\boldsymbol{\pi}}]$.

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

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

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, Foster, and Golowich (2020). In particular, they studied a simple two-player zero-sum Markov game known as Von Neumann's *ratio game*, 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 $\boldsymbol{x}_1 \in \Delta(\mathcal{A}_1) =: \mathcal{X}_1, \boldsymbol{x}_2 \in \Delta(\mathcal{A}_2) =: \mathcal{X}_2$, and $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{\mathcal{A}_1\times\mathcal{A}_2}$. It is further assumed that $\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2 \geq \zeta$, 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 (Daskalakis, Foster, and Golowich, 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 that will be useful in the sequel: that game

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

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 4 (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; $0 < \ell \leq A(\boldsymbol{x}) \leq h$; and*
- *$W(\boldsymbol{x}) \coloneqq \sum_{r=1}^d w_r(\boldsymbol{x}) \mathbf{1}_{\mathcal{Z}_r}$; $0 < \ell \leq W(\boldsymbol{x}) \leq h$.*

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

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

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 4 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 4 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 4 to the difficulty exposed by Proposition 3 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

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

respectively, then (5) is satisfied (in this particular application, $d = 2$). Furthermore, having assumed that $\boldsymbol{x}_1^\top \mathbf{S} \boldsymbol{x}_2 \geq \zeta > 0$, we also have control over the lower bound ℓ (as well as the upper bound h); naturally, taking ℓ arbitrarily small trivializes Property 4, and so the interesting regime occurs when ℓ is bounded away from 0—this also becomes evident from the guarantee of Theorem 6. 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 the sequel, we will prove this property in much greater generality.

As we shall see, Property 4 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 4.

Property 5 (Average version of Property 4)**.** *Under the preconditions of Property 4 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 $\sigma^{(T)} \coloneqq (\boldsymbol{x}^{(t)})_{1 \leq t \leq T}$ there exists $\mathcal{X} \ni \boldsymbol{x}^\star = \boldsymbol{x}^\star(\sigma^{(T)})$ so that*

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

Property 4 clearly implies Property 5 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.

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

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

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 2013), 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 4; 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. Below, we show that Property 5 is indeed sufficient to guarantee tractability for the induced VI problem, in the following formal sense.

Theorem 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 with $B_F \coloneqq \max_{1 \leq r \leq d} \|F_r\|_2$. Suppose further that the average (α, ℓ, h)-generalized Minty property (Property 5) holds. Then, for any $\epsilon > 0$, after $T \geq \frac{2 D_{\mathcal{X}}^2 h}{\ell \epsilon^2}$ 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 $\boldsymbol{x}^\star \in \mathcal{X}$,*

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

Proof sketch The proof of this theorem is deferred to the full version, 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 in the context of multi-player games (Rakhlin and Sridharan 2013; Syrgkanis et al. 2015); more precisely, we essentially view each component over $\mathcal{Z}_r$, comprising the Cartesian product $\mathcal{X} \coloneqq \bigtimes_{r=1}^d \mathcal{Z}_r$, 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 5 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)—that the second-order path lengths of the dynamics are bounded. Then, Theorem 6 follows by the assumption that $0 < \ell \leq A(\boldsymbol{x}) \leq h$; 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 $\langle \boldsymbol{x}^{(t)} - \boldsymbol{x}^\star, F(\boldsymbol{x}^{(t)}) \rangle \leq \epsilon$ for any $\boldsymbol{x}^\star \in \mathcal{X}$—as in the guarantee of Theorem 6—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 applications, let us consider an n-player game so that $F = (F_1, \dots, F_n)$ and $F_i := -\nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x})$, where $u_i : \mathcal{X} \to \mathbb{R}$ is the differentiable utility of player $i \in [\![n]\!]$.

Corollary 7. *Under the preconditions of Theorem 6, 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 $\|\boldsymbol{x}_i^\star - \boldsymbol{x}_i\|_2 \leq \delta$, $u_i(\boldsymbol{x}) - u_i(\boldsymbol{x}_i^\star, \boldsymbol{x}_{-i}) \geq -\epsilon - \frac{L}{2}\delta^2$;*
2. *if each $u_i(\boldsymbol{x}_i, \cdot)$ is gradient dominant, then for any player $i \in [\![n]\!]$ and $\boldsymbol{x}_i^\star \in \mathcal{X}_i$, $u_i(\boldsymbol{x}) - u_i(\boldsymbol{x}_i^\star, \boldsymbol{x}_{-i}) \geq -\epsilon$.*

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

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

for all $\boldsymbol{x} \in \mathcal{X}$, where $G > 0$ is some parameter. As such, Item 2 follows directly by definition and Theorem 6. 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, Skoulakis, and Zampetakis 2021); more precisely, Item 1 is interesting in the local regime $\delta < \sqrt{\frac{2\epsilon}{L}}$; see (Daskalakis, Skoulakis, and Zampetakis 2021, Definition 1.1); other notions of local optimality have also been studied in the literature, but this is not in our scope here.

Before we conclude this section, let us highlight some interesting extensions of Theorem 6 included in the full version. First, one can further broaden the scope of Property 5 by replacing the right-hand side of (6) by $-\gamma T$, for some parameter $\gamma \in \mathbb{R}_{\geq 0}$. 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. In such settings, one has control over the parameter γ, and so by taking $\gamma := \epsilon^2$ we can generalize the guarantee of Theorem 6.

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. More precisely, for parameters $\rho, \delta > 0$, we assume access to a noisy operator $F^{\delta,\rho} : \mathcal{X} \to \mathcal{X}$ such that $\|F^{\delta,\rho}(\boldsymbol{x}) - F(\boldsymbol{x})\|_2 \leq \delta \cdot \textsc{EqGap}(\boldsymbol{x})$, 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 5 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, we show that the conclusion of Theorem 6 is robust if δ and ρ are small enough.

Optimistic Policy Gradient in Multi-Player Markov Games

In this section, we leverage the theory developed earlier 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 8. *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 8 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, $V_1(\boldsymbol{\rho}) := -V(\boldsymbol{\rho})$ and $V_2(\boldsymbol{\rho}) := V(\boldsymbol{\rho})$ (since the game is zero-sum). The above display establishes Assumption 8 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 8 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 8. However, unlike Assumption 8, the above condition always holds since (stationary) NE always exist.

As it will become clear, Assumption 8 is naturally associated with Property 4. We also introduce a more permissive assumption in direct correspondence with Property 5.

Assumption 9. *Let $\mathcal{G}$ be a Markov game. For any sequence of product policies $\sigma^{(T)} := (\boldsymbol{\pi}^{(t)})_{1 \leq t \leq T}$, 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 9 is satisfied for the class of *zero-sum polymatrix Markov* games (Kalogiannis and Panageas 2023) (see also (Park, Zhang, and Ozdaglar 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) 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 10. *Assumption 9 is satisfied for any polymatrix zero-sum Markov game.*

In fact, this result is a byproduct of a more general characterization that we prove. We first recall the concept of a *coarse correlated equilibrium (CCE)*, which relaxes Definition 2 by allowing *correlated* policies. We will further use the concept of an ϵ-*average* CCE (ϵ-ACCE), a relaxation of CCE in which the sum—instead of the maximum as in CCE—of the players' deviation benefits is at most ϵ.

Definition 11 (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 10 is thus implied by the following result.

Proposition 12. *Assumption 9 is satisfied in any Markov game $\mathcal{G}$ exhibiting equilibrium collapse per Definition 11.*

Having justified Assumptions 8 and 9, we now proceed to establishing Property 4. Taking a step back, one might hope that equilibrium collapse (in the sense of Definition 11) would already suffice to efficiently compute stationary NE—as in the case of normal-form games. However, recent lower bounds 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 5. Indeed, this is shown in the following key lemma, which relies on the expression of the difference of the value function and the connection between the Q function and the gradient of the value function. In accordance with our earlier theory, we let $F_{\mathcal{G}}(\boldsymbol{x}) \coloneqq -(\nabla_{\boldsymbol{x}_1} V_1(\boldsymbol{\rho}), \dots, \nabla_{\boldsymbol{x}_n} V_n(\boldsymbol{\rho}))$.

Lemma 13. *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 $\Lambda(\boldsymbol{x}, \boldsymbol{x}^\star) \coloneqq (\Lambda_1(\boldsymbol{x}, \boldsymbol{x}^\star), \dots, \Lambda_n(\boldsymbol{x}, \boldsymbol{x}^\star))$. If Assumption 8 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 4 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 4—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 4 turns out to be crucial to apply our techniques. In fact, the recent hardness result of Park, Zhang, and Ozdaglar (2023) suggests that the general case should be intractable. We further remark that Lemma 13 applies similarly to conclude Property 5 if we substitute Assumption 8 by Assumption 9.

Finally, having established Lemma 13, we can now apply Theorem 6 along with the gradient dominance property to obtain one of our main results. Specifically, we appropriately bound all of the involved parameters appearing in Theorem 6; as usual, this includes a certain *distribution mismatch coefficient* $C_{\mathcal{G}}$—the multi-player analog of the quantity considered by Daskalakis, Foster, and Golowich (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 14. *Let $\mathcal{G}$ be a Markov game that satisfies Assumption 9 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 6 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 13, 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 6.

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 14. Yet, our approach is more restrictive in that it rests on having a single controller.

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; Aydin and Eksin 2023; Ding et al. 2022; Zhang et al. 2022b), or two-player zero-sum Markov games (Daskalakis, Foster, and Golowich 2020; Cen et al. 2023; 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, Doan, and Romberg 2022; Pattathil, Zhang, and Ozdaglar 2023; Yang and Ma 2023), albeit with a few exceptions (Qin and Etesami 2023; Sayin 2023; Giannou et al. 2022; Kalogiannis and Panageas 2023; Kalogiannis et al. 2023; Park, Zhang, and Ozdaglar 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, Golowich, and Zhang 2023; Jin et al. 2021; Erez et al. 2023; Liu, Szepesvári, and Jin 2022; Zhang et al. 2022a).

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, Zhang, and Ozdaglar 2023) investigates the class of polymatrix Markov games, which is based on the homonymous class of normal-form games (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, Chinchali, and Katti 2020; Parise and Ozdaglar 2019), and references therein). Specifically, Kalogiannis and Panageas (2023) and Park, Zhang, and Ozdaglar (2023) leverage the equilibrium collapse of CCE to NE to show that Markov NE can be computed efficiently; in stark contrast, Park, Zhang, and Ozdaglar (2023) showed that computing a *stationary* NE is PPAD-hard; the latter hardness result is based on earlier work by Daskalakis, Golowich, and Zhang (2023); Jin, Muthukumar, and Sidford (2023). In the class of polymatrix zero-sum Markov games, our novelty compared to earlier work (Kalogiannis and Panageas 2023; Park, Zhang, and Ozdaglar 2023) (see also the concurrent paper of Ma, Yang, and Zhang (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; 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. 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 (Lee, Kroer, and Luo 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, Antonakopoulos, and Mertikopoulos 2021; Hussain, Belardinelli, and Piliouras 2023). Even for the latter class of games, it is known that gradient-based methods may fail to converge pointwise to Nash equilibria (Mertikopoulos, Papadimitriou, and Piliouras 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 (2013); Chiang et al. (2012), leads to last-iterate convergence in monotone settings (Cai, Oikonomou, and Zheng 2022; Gorbunov, Taylor, and Gidel 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, Daskalakis, and Jordan (2021). In turn, this has engendered a considerable recent body of work; we refer to the papers of Cai and Zheng (2023); Pethick et al. (2022); Lee and Kim (2021); Mahdavinia et al. (2022), and the 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, Saad, and Niyato 2016; Guan et al. 2016; Qiu et al. 2021; Sayin, Zhang, and Ozdaglar 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, Saad, and Niyato (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.

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. First, our new condition 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, it may break in some settings of interest (Jordan, Lin, and Zampetakis 2023; Goktas and Greenwald 2022). Extending our theory 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 tractability is another promising direction.

Acknowledgments

We thank the anonymous reviewers at AAAI for their feedback. This material is 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.

References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2021. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *JMLR*.

Anagnostides, I.; Panageas, I.; Farina, G.; and Sandholm, T. 2022. On Last-Iterate Convergence Beyond Zero-Sum Games. In *ICML*.

Aydin, S.; and Eksin, C. 2023. Policy Gradient Play with Networked Agents in Markov Potential Games. In *L4DC*.

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

Bowling, M.; Burch, N.; Johanson, M.; and Tammelin, O. 2015. Heads-up Limit Hold'em Poker is Solved. *Science*.

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

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

Cai, Y.; Candogan, O.; Daskalakis, C.; and Papadimitriou, C. H. 2016. Zero-Sum Polymatrix Games: A Generalization of Minmax. *Mathematics of Operations Research*.

Cai, Y.; Oikonomou, A.; and Zheng, W. 2022. Finite-Time Last-Iterate Convergence for Learning in Multi-Player Games. In *NeurIPS*.

Cai, Y.; and Zheng, W. 2023. Accelerated Single-Call Methods for Constrained Min-Max Optimization. In *ICLR*.

Cen, S.; Chi, Y.; Du, S. S.; and Xiao, L. 2023. Faster Last-iterate Convergence of Policy Optimization in Zero-Sum Markov Games. In *ICLR*.

Chiang, C.-K.; Yang, T.; Lee, C.-J.; Mahdavi, M.; Lu, C.-J.; Jin, R.; and Zhu, S. 2012. Online optimization with gradual variations. In *COLT*.

Chu, T.; Chinchali, S.; and Katti, S. 2020. Multi-agent Reinforcement Learning for Networked System Control. In *ICLR*.

Cui, Q.; and Du, S. S. 2022. When are Offline Two-Player Zero-Sum Markov Games Solvable? In *NeurIPS*.

Daskalakis, C.; Foster, D. J.; and Golowich, N. 2020. Independent Policy Gradient Methods for Competitive Reinforcement Learning. In *NeurIPS*.

Daskalakis, C.; Goldberg, P.; and Papadimitriou, C. 2006. The Complexity of Computing a Nash Equilibrium. In *STOC*.

Daskalakis, C.; Golowich, N.; and Zhang, K. 2023. The Complexity of Markov Equilibrium in Stochastic Games. In *COLT*.

Daskalakis, C.; Skoulakis, S.; and Zampetakis, M. 2021. The complexity of constrained min-max optimization. In *STOC*.

Diakonikolas, J.; Daskalakis, C.; and Jordan, M. I. 2021. Efficient Methods for Structured Nonconvex-Nonconcave Min-Max Optimization. In *AISTATS*.

Ding, D.; Wei, C.; Zhang, K.; and Jovanovic, M. R. 2022. Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence. In *ICML*.

Eldosouky, A.; Saad, W.; and Niyato, D. 2016. Single controller stochastic games for optimized moving target defense. In *ICC*.

Erez, L.; Lancewicki, T.; Sherman, U.; Koren, T.; and Mansour, Y. 2023. Regret Minimization and Convergence to Equilibria in General-sum Markov Games. In *ICML*.

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

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

Giannou, A.; Lotidis, K.; Mertikopoulos, P.; and Vlatakis-Gkaragkounis, E. 2022. On the convergence of policy gradient methods to Nash equilibria in general stochastic games. In *NeurIPS*.

Goktas, D.; and Greenwald, A. 2022. Exploitability Minimization in Games and Beyond. In *NeurIPS*.

Golowich, N.; Pattathil, S.; Daskalakis, C.; and Ozdaglar, A. E. 2020. Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems. In *COLT*.

Gorbunov, E.; Taylor, A.; and Gidel, G. 2022. Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities. In *NeurIPS*.

Guan, P.; Raginsky, M.; Willett, R.; and Zois, D. 2016. Regret minimization algorithms for single-controller zero-sum stochastic games. In *CDC*.

Harris, K.; Anagnostides, I.; Farina, G.; Khodak, M.; Wu, S.; and Sandholm, T. 2023. Meta-Learning in Games. In *ICLR*.

Hsieh, Y.; Antonakopoulos, K.; and Mertikopoulos, P. 2021. Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium. In *COLT*.

Huang, B.; Lee, J. D.; Wang, Z.; and Yang, Z. 2022. Towards General Function Approximation in Zero-Sum Markov Games. In *ICLR*.

Hussain, A. A.; Belardinelli, F.; and Piliouras, G. 2023. Asymptotic Convergence and Performance of Multi-Agent Q-learning Dynamics. In *AAMAS*.

Jin, C.; Liu, Q.; Wang, Y.; and Yu, T. 2021. V-Learning - A Simple, Efficient, Decentralized Algorithm for Multiagent RL.

Jin, Y.; Muthukumar, V.; and Sidford, A. 2023. The Complexity of Infinite-Horizon General-Sum Stochastic Games. In *ITCS*.

Jordan, M. I.; Lin, T.; and Zampetakis, M. 2023. First-Order Algorithms for Nonlinear Generalized Nash Equilibrium Problems. *JMLR*.

Kalogiannis, F.; Anagnostides, I.; Panageas, I.; Vlatakis-Gkaragkounis, E.; Chatziafratis, V.; and Stavroulakis, S. A. 2023. Efficiently Computing Nash Equilibria in Adversarial Team Markov Games. In *ICLR*.

Kalogiannis, F.; and Panageas, I. 2023. Zero-sum Polymatrix Markov Games: Equilibrium Collapse and Efficient Computation of Nash Equilibria.

Lee, C.; Kroer, C.; and Luo, H. 2021. Last-iterate Convergence in Extensive-Form Games. In *NeurIPS*.

Lee, S.; and Kim, D. 2021. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*.

Leonardos, S.; Overman, W.; Panageas, I.; and Piliouras, G. 2022. Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games. In *ICLR*.

Liu, M.; Ozdaglar, A. E.; Yu, T.; and Zhang, K. 2023. The Power of Regularization in Solving Extensive-Form Games. In *ICLR*.

Liu, Q.; Szepesvári, C.; and Jin, C. 2022. Sample-Efficient Reinforcement Learning of Partially Observable Markov Games. In *NeurIPS*.

Ma, Z.; Yang, J.; and Zhang, Z. 2023. Near-Optimal Last-iterate Convergence of Policy Optimization in Zero-sum Polymatrix Markov games.

Mahdavinia, P.; Deng, Y.; Li, H.; and Mahdavi, M. 2022. Tight analysis of extra-gradient and optimistic gradient methods for nonconvex minimax problems. *NeurIPS*.

Mertikopoulos, P.; Lecouat, B.; Zenati, H.; Foo, C.; Chandrasekhar, V.; and Piliouras, G. 2019. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*.

Mertikopoulos, P.; Papadimitriou, C. H.; and Piliouras, G. 2018. Cycles in Adversarial Regularized Learning. In *SODA*.

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

Parise, F.; and Ozdaglar, A. E. 2019. Graphon Games. In *EC*.

Park, C.; Zhang, K.; and Ozdaglar, A. E. 2023. Multi-Player Zero-Sum Markov Games with Networked Separable Interactions.

Parthasarathy, T.; and Raghavan, T. E. 1981. An Orderfield Property for Stochastic Games When One Player Controls Transition Probabilities. *J. Optim. Theory Appl.*

Pattathil, S.; Zhang, K.; and Ozdaglar, A. E. 2023. Symmetric (Optimistic) Natural Policy Gradient for Multi-Agent Learning with Parameter Convergence. In *AISTATS*.

Perolat, J.; Scherrer, B.; Piot, B.; and Pietquin, O. 2015. Approximate dynamic programming for two-player zero-sum Markov games. In *ICML*.

Pethick, T.; Latafat, P.; Patrinos, P.; Fercoq, O.; and Cevher, V. 2022. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *ICLR*.

Piliouras, G.; Ratliff, L. J.; Sim, R.; and Skoulakis, S. 2022. Fast Convergence of Optimistic Gradient Ascent in Network Zero-Sum Extensive Form Games. In *SAGT*.

Qin, T.; and Etesami, S. R. 2023. Scalable and Independent Learning of Nash Equilibrium Policies in n-Player Stochastic Games with Unknown Independent Chains.

Qiu, S.; Wei, X.; Ye, J.; Wang, Z.; and Yang, Z. 2021. Provably Efficient Fictitious Play Policy Optimization for Zero-Sum Markov Games with Structured Transitions. In *ICML*.

Rakhlin, A.; and Sridharan, K. 2013. Online Learning with Predictable Sequences. In *COLT*.

Sayin, M.; Zhang, K.; Leslie, D.; Basar, T.; and Ozdaglar, A. 2021. Decentralized Q-learning in zero-sum Markov games. In *NeurIPS*, volume 34, 18320–18334.

Sayin, M. O. 2023. Decentralized Learning for Stochastic Games: Beyond Zero Sum and Identical Interest.

Sayin, M. O.; Zhang, K.; and Ozdaglar, A. E. 2022. Fictitious Play in Markov Games with Single Controller. In *EC*.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M. I.; and Moritz, P. 2015. Trust Region Policy Optimization. In *ICML*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms.

Shapley, L. S. 1953. Stochastic Games. *PNAS*.

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

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *NIPS*.

Syrgkanis, V.; Agarwal, A.; Luo, H.; and Schapire, R. E. 2015. Fast convergence of regularized learning in games. In *NeurIPS*.

Wei, C.; Lee, C.; Zhang, M.; and Luo, H. 2021. Last-iterate Convergence of Decentralized Optimistic Gradient Descent/Ascent in Infinite-horizon Competitive Markov Games. In *COLT*.

Yang, Y.; and Ma, C. 2023. $O(T^{-1})$ Convergence of Optimistic-Follow-the-Regularized-Leader in Two-Player Zero-Sum Markov Games. In *ICLR*.

Zeng, S.; Doan, T.; and Romberg, J. 2022. Regularized gradient descent ascent for two-player zero-sum Markov games. In *NeurIPS*.

Zhang, K.; Kakade, S.; Basar, T.; and Yang, L. 2020. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In *NeurIPS*.

Zhang, K.; Yang, Z.; and Basar, T. 2019. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. In *ICML*.

Zhang, R.; Liu, Q.; Wang, H.; Xiong, C.; Li, N.; and Bai, Y. 2022a. Policy Optimization for Markov Games: Unified Framework and Faster Convergence. In *NeurIPS*.

Zhang, R.; Mei, J.; Dai, B.; Schuurmans, D.; and Li, N. 2022b. On the Global Convergence Rates of Decentralized Softmax Gradient Play in Markov Potential Games. In *NeurIPS*.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *NIPS*.