Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games

Eric Mazumdar EMAZUMDAR@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA

Lillian J. Ratliff RATLIFFL@UW.EDU
Department of Electrical and Computer Engineering
University of Washington, Seattle, WA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Division of Computer Science and Department of Statistics
University of California, Berkeley, CA

S. Shankar Sastry SASTRY@COE.BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley, CA

Editor:

Abstract

We show by counterexample that policy-gradient algorithms have no guarantees of even local convergence to Nash equilibria in continuous action and state space multi-agent settings. To do so, we analyze gradient-play in N–player general-sum linear quadratic games, a classic game setting which is recently emerging as a benchmark in the field of multi-agent learning. In such games the state and action spaces are continuous and global Nash equilibria can be found be solving coupled Ricatti equations. Further, gradient-play in LQ games is equivalent to multi-agent policy-gradient. We first show that these games are surprisingly not convex games. Despite this, we are still able to show that the only critical points of the gradient dynamics are global Nash equilibria. We then give sufficient conditions under which policy-gradient will avoid the Nash equilibria, and generate a large number of general-sum linear quadratic games that satisfy these conditions. In such games we empirically observe the players converging to limit cycles for which the time average does not coincide with a Nash equilibrium. The existence of such games indicates that one of the most popular approaches to solving reinforcement learning problems in the classic reinforcement learning setting has no local guarantee of convergence in multi-agent settings. Further, the ease with which we can generate these counterexamples suggests that such situations are not mere edge cases and are in fact quite common.

1. Introduction

Interest in multi-agent reinforcement learning has seen a recent surge of late, and policy-gradient algorithms are championed due to their potential scalability. Indeed, recent impressive successes of multi-agent reinforcement learning have made use of policy optimization algorithms such as multi-agent actor-critic (Lowe et al., 2017; Srinivasan et al., 2018; Jaderberg et al., 2019), multi-agent proximal policy optimization (Bansal et al., 2018), and even simple multi-agent policy-gradients

(Lanctot et al., 2017) in problems where the various agents have high-dimensional continuous state and action spaces like StarCraft II (Vinyals et al., 2019).

Despite these successes, a theoretical understanding of these algorithms in multi-agent settings is still lacking. Missing perhaps, is a tractable yet sufficiently complex setting in which to study these algorithms. Recently, there has been much interest in analyzing the convergence and sample complexity of policy-gradient algorithms in the classic linear quadratic regulator (LQR) problem from optimal control (Kalman, 1960). The LQR problem is a particularly apt setting to study the properties of reinforcement learning algorithms due to the existence of an optimal policy which is a linear function of the state and which can be found by solving a Ricatti equation. Indeed, the relative simplicity of the problem has allowed for new insights into the behavior of reinforcement learning algorithms in continuous action and state spaces (Dean et al., 2017; Fazel et al., 2018; Malik et al., 2019).

An extension of the LQR problem to the setting with multiple agents, known as a *linear quadratic (LQ) game*, has also been well studied in the literature on dynamic games and optimal control (Basar and Olsder, 1998). As the name suggests, an LQ game is a setting in which multiple agents attempt to optimally control a shared linear dynamical system subject to quadratic costs. Since the players have their own costs, the notion of 'optimality' in such games is a Nash equilibrium properties of which have been well analyzed in the literature Engwerda (1998); Possieri and Sassano (2015); Basar (1976); Lukes and Russell (1971).

Like LQR for the classical single-agent setting, LQ games are an appealing setting in which to analyze the behavior of multi-agent reinforcement learning algorithms in continuous action and state spaces since they admit global Nash equilibria in the space of linear feedback policies. Moreover, these equilibria can be found by solving a coupled set of Ricatti equations. As such, LQ games are a natural benchmark problem on which to test policy-gradient algorithms in multi-agent settings. Furthermore, policy gradient methods open up the possibility to new scalable approaches to finding solutions to control problems even with constraints. In the single-agent setting, it was recently shown that policy-gradient has global convergence guarantees for the LQR problem (Fazel et al., 2018). These results have recently been extended to projected policy-gradient algorithms in zero-sum LQ games (Zhang et al., 2019).

Contributions. We present a *negative* result, showing that policy-gradient in general-sum LQ games does not enjoy *even local* convergence guarantees, unlike in LQR and zero-sum LQ games. In particular, we show that, if each player randomly initializes their policy and then uses a policy-gradient algorithm, there exists an LQ game in which the players would almost surely avoid a Nash equilibrium. Further, our numerical experiments indicate that LQ games in which this occurs may be quite common. We also observe empirically that when players fail to converge to the Nash equilibrium they do converge to stable limit cycles. These cycles do not seem to have any readily apparent relationship to the Nash equilibria of the game.

We note that non-convergence to Nash equilibria is not in itself a new phenomenon (see e.g. Mazumdar et al. (2019); Daskalakis et al. (2017); Cesa-Bianchi and Lugosi (2006)) and that the existence of cycles in the dynamics of learning dynamics in games has also been repeatedly observed in various contexts Mazumdar et al. (2018); Mertikopoulos et al. (2018); Papadimitriou and Piliouras. However, we believe that such phenomena have not yet been shown to occur in the dynamics of multi-agent reinforcement learning algorithms in continuous action and state spaces. Since such algorithms have had such striking successes in recent years, we believe a theoretical

understanding of their behaviors can lay the groundwork for the development of more efficient and theoretically sound multi-agent learning algorithms.

Organization. Section 2 introduces N-player general-sum LQ games and presents previous results on the existence of the Nash equilibrium in such games. In Section 3, we show that these games are *not* convex games and that all the stationary points of the joint policy-gradient dynamics are Nash equilibria. Following this, we give sufficient conditions under which policy-gradient almost surely avoids a Nash equilibrium in Section 4. Given these theoretical results, in Section 5 we present empirical results demonstrating that a large number of 2-player LQ games satisfy these sufficient conditions. Numerical experiments showing the existence of limit cycles in the gradient dynamics of general-sum LQ games are also presented. The paper is concluded with a discussion in Section 6.

2. Preliminaries

We consider N-player LQ games subject to a discrete-time dynamical system defined by

$$z(t+1) = Az(t) + \sum_{i=1}^{N} B_i u_i(t) \quad z(0) = z_0 \sim \mathcal{D}_o, \tag{1}$$

where $z(t) \in \mathbb{R}^m$ is the state at time t, $\mathcal{D}_o$ is the initial state distribution, and $u_i(t) \in \mathbb{R}^{d_i}$ is the control input of player $i \in 1, \ldots, N$. For LQ games, it is known that under reasonable assumptions, linear feedback policies for each player that constitute a Nash equilibrium exist and are unique if a set of coupled Ricatti equations admit a unique solution (Basar and Olsder, 1998). Thus, we consider that each player i searches for a linear feedback policy of the form $u_i(t) = -K_i z(t)$ that minimizes their loss, where $K_i \in \mathbb{R}^{d_i \times m}$. We use the notation $d = \sum_{i=1}^{N} d_i$ for the combined dimension of the players' parameterized policies.

As the name of the game implies, the players' loss functions are quadratic functions given by

$$f_i(u_1, \ldots, u_N) = \mathbb{E}_{z_0 \sim \mathcal{D}_o} \left[\sum_{t=0}^{\infty} z(t)^T Q_i z(t) + u_i(t)^T R_i u_i(t) \right],$$

where Q_i and R_i are the cost matrices for the state and input, respectively.

Assumption 1 *For each player $i \in \{1, \ldots, N\}$, the state and control cost matrices satisfy $Q_i \succ 0$ and $R_i \succ 0$.*

We note that the players are coupled through the dynamics since $z(t)$ is constrained to obey the update equation given in (1). We focus on a setting in which all players randomly initialize their strategy and then perform gradient descent simultaneously on their own cost functions with respect to their individual control inputs. That is, the players use policy-gradient algorithms of the following form:

$$K_{i,n+1} = K_{i,n} - \gamma_i D_i f_i(K_{1,n}, \ldots, K_{N,n}) \tag{2}$$

where $D_i f_i(\cdot, \cdot)$ denotes the derivatives of f_i with respect to the i–th argument, and $\{\gamma_i\}_{i=1}^{N}$ are the step-sizes of the players. We note that there is a slight abuse of notation here in the expression of $D_i f_i$ as functions of the parameters K_i as opposed to the control inputs u_i. To ensure there is no confusion between t and n, we also point out that n indexes the policy-gradient algorithm iterations while t indexes the time of the dynamical system.

To simplify notation, define

$$\Sigma_K = \mathbb{E}_{z_0 \sim \mathcal{D}_o} \left[\sum_{t=0}^{\infty} z(t) z(t)^T \right],$$

where we use the subscript notation to denote the dependence on the collection of controllers $K = (K_1, \ldots, K_N)$. Define also the initial state covariance matrix

$$\Sigma_0 = \mathbb{E}_{z_0 \sim D_0}[z_0 z_0^T].$$

Direct computation verifies that for player i, $D_i f_i$ is given by:

$$D_i f_i(K_1, \ldots, K_N) = 2(R_i K_i - B_i^T P_i \bar{A}) \Sigma_K, \tag{3}$$

where $\bar{A} = A - \sum_{i=1}^{N} B_i K_i$, is the closed–loop dynamics given all players' control inputs and, for given $(K_1, \ldots, K_N)$, the matrix P_i is the unique positive definite solution to the Bellman equation:

$$P_i = \bar{A}^T P_i \bar{A} + K_i^T R_i K_i + Q_i, \;\; i \in \{1, \ldots, N\}. \tag{4}$$

Given that the players may have different control objectives and do not engage in coordination or cooperation, the best they can hope to achieve is a Nash equilibrium.

Definition 1 *A* feedback Nash equilibrium *is a collection of policies* $(K_1^*, \ldots, K_N^*)$ *such that:*

$$f_i(K_1^*, \ldots, K_i^*, \ldots, K_N^*) \leq f_i(K_1^*, \ldots, K_i, \ldots, K_N^*), \;\; \forall \, K_i \in \mathbb{R}^{d_i \times m}.$$

for each $i \in \{1, \ldots, N\}$.

Under suitable assumptions on the cost matrices, the Nash equilibrium of an LQ game is known to exist in the space of linear policies Basar and Olsder (1998); Li and Gajic (1995). However, this Nash equilibrium may not be unique. To the best of our knowledge, there are no general set of conditions under which the Nash equilibrium is unique in general-sum LQ games outside of the scalar dynamics setting Engwerda (1998). There are, however, algebraic geometry methods to compute all Nash equilibria in LQ games Possieri and Sassano (2015). We make use of a simpler algorithm to find Nash equilibria which solves coupled Ricatti equations using the method of Lyapunov iterations. The method is outlined in Li and Gajic (1995) for continuous time LQ games, and an analogous procedure can be followed for discrete time. Convergence of this method requires the following assumption.

Assumption 2 *For at least one player* $i \in \{1, \ldots, N\}$, (A, B_i) *is stabilizable.*

Assumption 2 is a necessary condition for the players to be able to stabilize the system. Indeed, the player's costs are finite only if the closed loop system $\bar{A}$ is asymptotically stable, meaning that $|\mathrm{Re}(\lambda)| < 1$ for all $\lambda \in \mathrm{spec}(\bar{A})$, where $\mathrm{Re}(\lambda)$ denotes the real part of λ and $\mathrm{spec}(M)$ is the spectrum of a matrix M.

3. Analyzing the Optimization Landscape of LQ Games

Having introduced the class of games we consider we now analyze the optimization landscape in general-sum LQ games. Letting $x = (K_1, \dots, K_N)$, the object of interest is the map $\omega : \mathbb{R}^{md} \to \mathbb{R}^{md}$ defined as follows:

$$\omega(x) = \begin{bmatrix} D_1 f_1(K_1, \dots, K_N) \\ \vdots \\ D_N f_N(K_1, \dots, K_N) \end{bmatrix}.$$

Note that $D_i f_i = \partial f_i / \partial K_i$ has been converted to an md_i dimensional vector and each K_i has also been vectorized. This is a slight abuse of notation and throughout we treat the K_i's as both vectors and matrices; in general, the shape should be clear from context, and otherwise we make comments where necessary to clarify.

Before analyzing the stationary points of policy-gradient in LQ games, we show that the class of LQ games we consider are *not* convex games. This holds despite the linearity of the dynamics and the positive definiteness of the cost matrices. This fact makes the analysis of such games non-trivial since the lack of strong structural guarantees on the players' costs allows for non-trivial limiting behaviors like cycles, non-Nash equilibria, and chaos in the joint gradient dynamics. Mazumdar et al. (2018).

Proposition 2 *There exists a N-player LQ game satisfying assumptions 1 and 2 that is not a convex game.*

Proof The proof of Proposition 2 follows directly from the non-convexity of the set of stabilizing policies for the single-agent LQR problem which was shown in Fazel et al. (2018). Holding every other players' actions fixed, a player i is faced with a simple LQR problem. Since this problem is non-convex, LQ games are not convex games. ■

In the absence of strong structural guarantees on the players' costs, simultaneous gradient-play in general-sum games can converge to strategies that are not Nash equilibria (Mazumdar et al., 2018). The following theorem shows that, despite the fact that LQ games are not convex for each player, such non-Nash equilibria cannot exist in the gradient dynamics of general-sum LQ games. Indeed, we show that a point x is a critical point of the policy gradient dynamics in a N-player LQ game if and only if it is a Nash equilibrium. We note that critical points of gradient-play are strategies $x = (K_1, \dots, K_N)$ such that $\omega(x) = 0$. Such points are of particular importance since a necessary condition for a point x to be a Nash equilibrium is that it is a critical point.

Theorem 3 *Consider the set of stabilizing policies $x^* = (K_1^*, \dots, K_N^*)$ such that $\Sigma_{K^*} > 0$. $D_i f_i(K_1^*, \dots, K_N^*) = 0$ for each $i \in \{1, \dots, N\}$, if and only if x^* is a Nash equilibrium.*

Proof We prove the forward direction and show that if $D_i f_i(x^*) = 0$ for each $i \in \{1, \dots, N\}$, then x^* is a Nash equilibrium. We show this by contradiction. Suppose the claim does not hold so that $\Sigma_{K^*} > 0$ and $D_i f_i(K_1^*, \dots, K_N^*) = 0$ for each $i \in \{1, \dots, N\}$, yet $(K_1^*, \dots, K_N^*)$ is not a Nash equilibrium. That is, without loss of generality, there exists a $\bar{K}_1$ such that

$$f_1(\bar{K}_1, K_2^*, \dots, K_N^*) < f_1(K_1^*, \dots, K_N^*).$$

Now, fixing $(K_2^*, \ldots, K_N^*)$, player 1 can be seen as facing an LQR problem. Indeed, letting $(K_2^*, \ldots, K_N^*)$ be fixed, player 1 aims to find a 'best response' in the space of linear feedback policies of the form $u_1(t) = Kz(t)$ with $K \in \mathbb{R}^{d_i \times m}$ that minimizes $f_1(\cdot, K_2^*, \ldots, K_N^*)$ subject to the dynamics defined by

$$z(t+1) = \left(A - \sum_{i=2}^N B_i K_i\right) z(t) + B_1 u_1(t).$$

Note that this system is necessarily stabilizable since $\bar{A}$ is stable. Hence, the discrete algebraic Riccati equation for player 1's LQR problem has a positive definite solution P such that $R_1 + B_1^T P B_1 > 0$ since $R_1 > 0$ by assumption. Since $\Sigma_{K^*} > 0$ and $D_1 f_1(K_1^*, \ldots, K_N^*) = 0$, applying Corollary 4 of Fazel et al. (2018), we have that K_1^* must be optimal for player 1's LQR problem so that

$$f_1(K_1^*, \ldots, K_N^*) \leq f_1(K, K_2^*, \ldots, K_N^*), \ \ \forall\, K \in \mathbb{R}^{d_1 \times m}.$$

In particular, the above inequality holds for $\bar{K}_1$, which leads to a contradiction.

To prove the reverse direction, we note that a necessary condition for a point x to be a Nash equilibrium for each player, is that $D_i f_i(x^*) = 0$ for each $i \in \{1, \ldots, N\}$ Ratliff et al. (2013). ■

Theorem 3 shows that, just as in the single-player LQR setting and zero-sum LQ games, the critical points of gradient-play in N–player general-sum LQ games are all Nash equilibria. We note that the condition $\Sigma_K > 0$ can be satisfied by choosing an initial state distribution $\mathcal{D}_o$ with a full-rank covariance matrix.

A simple consequence of Theorem 3 is that when the coupled Ricatti equations characterizing the Nash equilibria of the game have a unique positive definite solution and Assumptions 1 and 2 hold, the gradient dynamics admit a unique critical point.

Corollary 4 *Under Assumption 1 and 2, if the coupled Ricatti equations admit a unique solution and $\Sigma_0 \succ 0$, then the map ω has a unique critical point.*

Given that the critical points of the gradient dynamics in LQ games are Nash equilibria, the aim is to show, via constructing counter-examples, that games in which the gradient dynamics avoid the Nash equilibria do in fact exist. A sufficient condition for this would be to find a game in which gradient-play diverges from neighborhoods of Nash equilibria.

It is demonstrated in Mazumdar et al. (2018) that there may be Nash equilibria that are not even *locally attracting* under the gradient dynamics in N–player general-sum games in which the players' costs are sufficiently smooth (i.e., at least twice continuously differentiable). In games that admit such Nash equilibria, the agents could initialize arbitrarily close to the Nash equilibrium, simultaneously perform individual gradient descent with arbitrarily small step sizes, and still diverge.

The class of N–player LQ games we consider does not, however, satisfy the smoothness assumptions necessary to simply invoke the results in Mazumdar et al. (2018). Indeed, the cost functions are non-smooth and, in fact, are infinite whenever the players have strategies that do not stabilize the dynamics. Further, the set of stabilizing policies for a dynamical system is not even convex (Fazel et al., 2018). Despite these challenges, in the sequel we show that the negative convergence results in Mazumdar et al. (2018) extend to the general-sum LQ setting. In particular, we show that even with arbitrarily small step sizes, players using policy-gradient in LQ games may still diverge from neighborhoods of a Nash equilibrium.

4. Sufficient Conditions for Policy-Gradient to Avoids Nash

We now give sufficient conditions under which gradient-play has no guarantees of even *local*, much less global, convergence to a Nash equilibrium. Towards this end, we first show that ω is sufficiently smooth on the set of stabilizing policies.

Let $\mathcal{S}^{md} \subset \mathbb{R}^{md}$ be the subset of stabilizing md–dimensional matrices.

Proposition 5 *Consider an N–player LQ game. The vector-valued map ω associated with the game is twice continuously differentiable on $\mathcal{S}^{md}$—i.e., $\omega \in C^2(\mathcal{S}^{md}, \mathcal{S}^{md})$.*

Using our notation, Lemma 6.5 in Zhang et al. (2019) shows for two-player zero-sum LQ games that (P_1, P_2), and Σ_K are continuously differentiable with respect to K_1 and K_2 when $A - B_1K_1 - B_2K_2$ is stable. This, in turn, implies that $\omega(K_1, K_2)$ is continuously differentiable with respect to K_1 and K_2 when the closed loop system $A - B_1K_1 - B_2K_2$ is stable. The result follows by a straightforward application of the implicit function theorem (Abraham et al., 1988). We utilize the same proof technique here in extending the result to N–player general-sum LQ games and, in fact, the proof implies that ω has even stronger regularity properties. Since the proof follows the same techniques as in Zhang et al. (2019), we defer it to Appendix A.

Given that ω is continuously differentiable over the set of stabilizing joint policies $(K_1, \ldots, K_N)$, the following result gives sufficient conditions such that the set of initial conditions in a neighborhood of the Nash equilibrium from which gradient-play converges to the Nash equilibrium is of measure zero. This implies that the players will almost surely avoid the Nash equilibrium even if they randomly initialize in a uniformly small ball around it.

Let the Jacobian of the vector field ω be denoted by $D\omega$. Given a critical point x^*, let λ_j be the eigenvalues of $D\omega(x^*)$, for $j \in \{1, \ldots, md\}$, where $d = \sum_{i=1}^n d_i$. Recall that the state $z(t)$ is dimension m.

Theorem 6 *Suppose that $\Sigma_0 > 0$. Consider any N–player LQ game satisfying Assumptions 1 and 2 that admits a Nash equilibrium that is a saddle point of the policy-gradient dynamics—i.e., LQ games for which the Jacobian of ω evaluated at the Nash equilibrium $x^* = (K_1^*, \ldots, K_N^*)$ has eigenvalues λ_j such that $\mathrm{Re}(\lambda_j) < 0$ for $j \in \{1, \ldots, \ell\}$ and $\mathrm{Re}(\lambda_j) > 0$ for $j \in \{\ell+1, \ldots, md\}$ for some ℓ such that $0 < \ell < md$. Then there exists a neighborhood U of x^* such that policy-gradient converges on a set of measure zero.*

Proof The proof is made up of three parts: (i) we show the existence of an open-convex neighborhood U of x^* on which ω is locally Lipschitz with constant L; (ii) we show that the map $g(x) = x - \Gamma\omega(x)$ is a diffeomorphism on U; and, (iii) we invoke the stable manifold theorem to show that the set of initializations in U on which policy-gradient converges is measure zero.

(i) ω is locally Lipschitz. Proposition 5 shows that ω is continuously differentiable on the set of stabilizing policies $\mathcal{S}^{md}$. Given Assumptions 1 and 2, the Nash equilibrium exists and $x^* \in \mathcal{S}^{md}$. Thus, there must exist an open convex neighborhood U of x^* such that $||D\omega||_2 < L$ for some $L > 0$.

(ii) g is a diffeomorphism. By the preceding argument, ω is locally Lipschitz on U with Lipschitz constant L. Consider the policy-gradient algorithm with $\gamma_i < 1/L$ for each $i \in \{1, \ldots, N\}$. Let $\Gamma = \mathrm{diag}(\Gamma_1, \ldots, \Gamma_N)$ where $\Gamma_i = \mathrm{diag}((\gamma_i)_{j=1}^{md_i})$—that is, Γ_i is an $md_i \times md_i$ diagonal matrix with γ_i repeated on the diagonal md_i times. Now, we claim the mapping $g : \mathbb{R}^{md} \to \mathbb{R}^{md}$:

$x \mapsto x - \Gamma\omega(x)$ is a diffeomorphism on U. If we can show that g is invertible on U and a local diffeomorphism, then the claim follows. Let us first prove that g is invertible.

Consider $x \neq y$ and suppose $g(y) = g(x)$ so that $y - x = \gamma \cdot (\omega(y) - \omega(x))$. Since $\|\omega(y) - \omega(x)\|_2 \leq L\|y-x\|_2$ on U, $\|x-y\|_2 \leq L\|\Gamma\|_2\|y-x\|_2 < \|y-x\|_2$ since $\|\Gamma\|_2 = \max_i |\gamma_i| < 1/L$.

Now, observe that $Dg = I - \Gamma D\omega(x)$. If Dg is invertible, then the implicit function theorem (Abraham et al., 1988) implies that g is a local diffeomorphism. Hence, it suffices to show that $\Gamma D\omega(x)$ does not have an eigenvalue equal to one. Indeed, letting $\rho(A)$ be the spectral radius of a matrix A, we know in general that $\rho(A) \leq \|A\|$ for any square matrix A and induced operator norm $\|\cdot\|$ so that $\rho(\Gamma D\omega(x)) \leq \|\Gamma D\omega(x)\|_2 \leq \|\Gamma\|_2 \sup_{x\in U} \|D\omega(x)\|_2 < \max_i |\gamma_i| L < 1$. Of course, the spectral radius is the maximum absolute value of the eigenvalues, so that the above implies that all eigenvalues of $\Gamma D\omega(x))$ have absolute value less than one.

Since g is injective by the preceding argument, its inverse is well-defined and since g is a local diffeomorphism on U, it follows that g^{-1} is smooth on U. Thus, g is a diffeomorphism.

(iii) Local convergence occurs on a set of measure zero. Let B be the open ball derived from Theorem 9 in Appendix B.

Starting from $x_0 \in U$, if gradient-based learning converges to a strict saddle point, then there exists an n_0 such that $g^n(x_0) \in B$ for all $n \geq n_0$. Applying Theorem 9 (Appendix B), we get that $g^n(x_0) \in W^{cs}_{\text{loc}} \cap B$. Now, using the fact that g is invertible, we can iteratively construct the sequence of sets defined by $W_1(x^*) = g^{-1}(W^{cs}_{\text{loc}} \cap B) \cap U$ and $W_{k+1}(x^*) = g^{-1}(W_k(x^*) \cap B) \cap U$. Then we have that $x_0 \in W_n(x^*)$ for all $n \geq n_0$. The set $U_0 = \cup_{k=1}^{\infty} W_k(x^*)$ contains all the initial points in U such that gradient-based learning converges to a strict saddle.

Since x^* is a strict saddle, $I - \Gamma D\omega(x^*)$ has an eigenvalue greater than one. This implies that the co-dimension of the unstable manifold is strictly less than md so that $\dim(W^{cs}_{\text{loc}}) < md$. Hence, $W^{cs}_{\text{loc}} \cap B$ has Lebesgue measure zero in $\mathbb{R}^{md}$. Using again that g is a diffeomorphism, $g^{-1} \in C^1$ so that it is locally Lipschitz and locally Lipschitz maps are null-set preserving. Hence, $W_k(x^*)$ has measure zero for all k by induction so that U_0 is a measure-zero set since it is a countable union of measure-zero sets. ■

Theorem 6 gives sufficient conditions under which, with random initializations of K_i, policy-gradient methods would almost surely avoid the critical point. Let each players' initial strategy $K_{i,0}$ be sampled from a distribution $p_{i,0}$ for $i \in \{1, ..., N\}$, and let p_0 be the resulting the joint distribution of $(K_{1,0}, \ldots, K_{N,0})$.

Corollary 7 *Suppose $\mathcal{D}_o$ is chosen such that $\Sigma_0 \succ 0$, and consider an N–player LQ game satisfying Assumptions 1 and 2 in which there is a Nash equilibrium which is a saddle point of the policy-gradient dynamics. If each player $i \in \{1, \ldots, N\}$ performs policy-gradient with a random initial strategy $K_{i,0} \sim p_{i,0}$ such that the support of p_0 is U, they will almost surely avoid the Nash equilibrium.*

Corollary 7 shows that even if the players randomly initialize in a neighborhood of a Nash equilibrium that is a saddle point of the joint gradient dynamics they will almost surely avoid it. The proof follows trivially from the fact that the set of initializations that converge to the Nash equilibrium is of measure zero in U.

In the next section, we generate a large number of LQ games that satisfy the conditions of Corollary 7. Taken together, these theoretical and numerical results imply that policy-gradient algorithms have no guarantees of local, and consequently global, convergence in general-sum LQ games.

Remark 8 *Theorem 6 gives us sufficient conditions under which policy-gradient in general-sum LQ games does not even have* local convergence guarantees*, much less global convergence guarantees. We remark that this is very different from the single-player LQR setting, where policy-gradient will converge from any initialization in a neighborhood of the optimal solution (Fazel et al., 2018). In zero-sum LQ games, the structure of the game also precludes any Nash equilibrium from satisfying the conditions of Theorem 6 (Mazumdar et al., 2018), meaning that local convergence is always guaranteed. In Zhang et al. (2019), the guarantee of local convergence is strengthened to that of global convergence for a class of projected policy-gradient algorithms in zero-sum LQ games.*

5. Generating Counterexamples

Since it is difficult to find a simple closed form for the Jacobian of ω due to the fact that the matrices P_i implicitly depend on all the K_i, we perform random search to find instances of LQ games in which the Nash equilibrium is a strict saddle point of the gradient dynamics. For each LQ game we generate, we use the method of Lyapunov iterations to find a global Nash equilibrium of the LQ game and numerically approximate the Jacobian to machine precision. We then check whether the Nash equilibrium is a strict saddle. Surprisingly, such a simple search procedure finds a large number of LQ games in which policy-gradient avoids Nash equilibria.

For simplicity, we focus on two-player LQ games where $z \in \mathbb{R}^2$ and $d_1 = d_2 = 1$. Thus, each player $i = 1, 2$ has two parameters to learn, which we denote $K_{i,j}$, $j = 1, 2$.

In the remainder of this section, we detail our experimental setup and then present our findings.

5.1 Experimental setup

To search for examples of LQ games in which policy-gradient avoids Nash equilibria, we fix B_1, Q_1, and R_1 and parametrize B_2, Q_2, and R_2 by b, q, and r, respectively. For various values of the parameters b, q, and r, we uniformly sample 1000 different dynamics matrices $A \in \mathbb{R}^{2\times 2}$ such that A, B_1, Q_1 satisfies Assumption 2. Then, for each of the 1000 different LQ games we find the optimal feedback matrices (K_1^*, K_2^*) using the method of Lyapunov iterations (i.e., a discrete time variant of the algorithm outlined in Li and Gajic (1995)), and then numerically approximate $D\omega(K_1^*, K_2^*)$ using auto-differentiation[1] tools and check its eigenvalues.

The exact values of the matrices are defined as follows:

$$A \in \mathbb{R}^{2\times 2} : a_{i,j} \sim \text{Uniform}(0, 1) \quad i, j = 1, 2,$$

$$B_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ B_2 = \begin{bmatrix} b \\ 1 \end{bmatrix},\ Q_1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix},\ Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & q \end{bmatrix}, R_1 = 0.01,\ R_2 = r.$$

5.2 Numerical results

Using the setup outlined in the previous section we randomly generated LQ games to search for counterexamples. We first present results that show that these counterexamples may be quite com-

1. We use auto-differentiation due to the fact that finding an analytical expression for $D\omega$ is unduly arduous even in low dimensions due to the dependence of P_i and Σ_{K_1,K_2} on (K_1, K_2), both of which are implicitly defined.

mon. We then use policy-gradient in two of the LQ games we generated and highlight the existence of limit cycles and the fact that the players' time-averaged strategies do not converge to the Nash equilibrium.

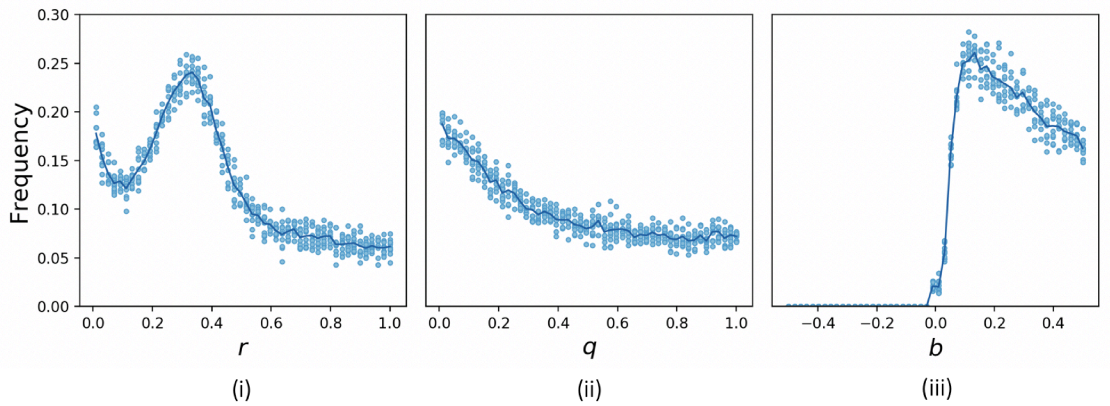

Figure 1: Frequency (out of 1000) of randomly sampled LQ games with global Nash equilibria that are avoided by policy-gradient. Each point represents, for the given parameter value, the frequency of such games out of 1000 randomly sampled A matrices. The solid line shows the average frequency of these games. (i) r is varied in $(0, 1)$, $b = 0$, $q = 0.01$. (ii) q is varied in $(0, 1)$, $b = 0$, $r = 0.1$. (iii) b is varied in $(-0.5, 0.5)$, $q = 0.01$, $r = 0.1$.

Avoidance of Nash in a nontrivial class of LQ games. As can be seen in Figure 1, across the different parameter values we considered, we found that anywhere from 0% to 25% of randomly sampled LQ games, had Nash equilibria that are strict saddle points of the gradient dynamics. Therefore, in up to 25% of the LQ games we generated policy-gradient would almost surely avoid a Nash solution. Of particular interest, for all values of q and r that we tested, when $b = 0$ at least 5% of the LQ games had a global Nash equilibrium with the strict saddle property.

These empirical observations imply that policy-gradient in competitive settings, even in the relatively straightforward setting of linear dynamics, linear policies, and quadratic costs, could fail to converge to a Nash equilibrium in up to one out of four such problems. This suggests that for more complicated cost functions, policy classes, and dynamics, Nash equilibria may often be avoided by policy-gradient.

We remark that each point in Figure 1 represents the number of counterexamples found (out of 1000) for each parameter value, meaning that for $r \approx 0.35, b = 0$, and $q = 0.01$ we were able to consistently generate around 250 different examples of games where policy-gradient almost surely avoids the only stationary point of the dynamics.

Note also that we were unable to find any counterexamples when b was varied in $(-0.5, 0.5)$ and $q = 0.01$, $r = 0.1$. This suggests that depending on the structure of the dynamical system it may be possible to give stronger convergence guarantees.

Convergence to Cycles. Figures 2–3 show the payoffs and parameter values of the two players when they use policy-gradient in two general-sum LQ games we identified as being counterexamples for convergence to the Nash equilibrium.

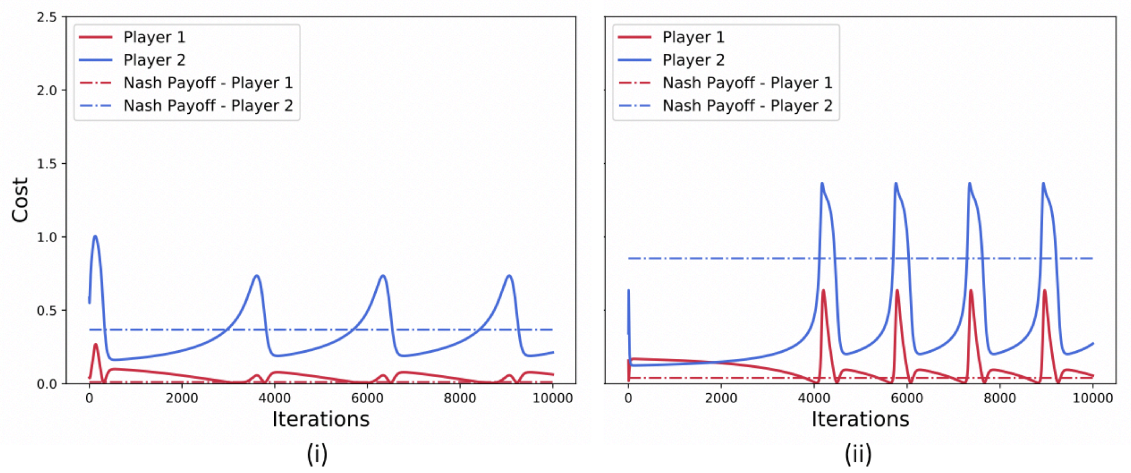

Figure 2: Payoffs of the two players in two general-sum LQ game where there is a Nash equilibrium that is avoided by the gradient dynamics. We observe empirically that in both games the two players diverge from the local Nash equilibrium and converge to a limit cycle around the Nash equilibrium.

In the two games, we initialize both players in a ball of radius 0.25 around their Nash equilibrium strategies and let them perform policy-gradient with step size 0.05. We observe that in both games the players diverge from the Nash equilibrium and converge to limit cycles.

For the two games in Figures 2–4, the game parameters are such that $b = 0$, $r = 0.01$, and $q = 0.147$. The two A matrices are defined as follows:

$$\text{(i):}\ A = \begin{bmatrix} 0.588 & 0.028 \\ 0.570 & 0.056 \end{bmatrix}, \quad \text{(ii):}\ A = \begin{bmatrix} 0.511 & 0.064 \\ 0.533 & 0.993 \end{bmatrix}. \tag{5}$$

We also chose the initial state distribution to be $[1, 1]^T$ or $[1, 1.1]^T$ with probability 0.5 each.

The eigenvalues of the corresponding game Jacobian $D\omega$ evaluated at the Nash equilibrium are as follows:

$$\text{(i):}\ \operatorname{spec}(D\omega(K_1^*, K_2^*)) = \{10.88, 2.02, -0.21, -0.06\}$$
$$\text{(ii):}\ \operatorname{spec}(D\omega(K_1^*, K_2^*)) = \{9.76, 0.54, -0.01 + 0.08j, -0.01 - 0.08j\}.$$

Thus, these games do satisfy the conditions of Corollary 7 for the avoidance of Nash equilibria. We conclude this section by noting that, as shown in Figure 4, the players' average payoffs do not necessarily converge to the Nash equilibrium payoffs.

6. Discussion

We have shown that in the relatively straightforward setting of N–player LQ games, agents performing policy-gradient have no guarantees of local, and therefore global, convergence to the Nash equilibria of the game even if they randomly initialize their first policies in a small neighborhood of the Nash equilibrium. Since we also showed that the Nash equilibria are the only critical points of the gradient dynamics, this means that, for this class of games, policy-gradient algorithms may have no guarantees of convergence to *any* set of stationary policies.

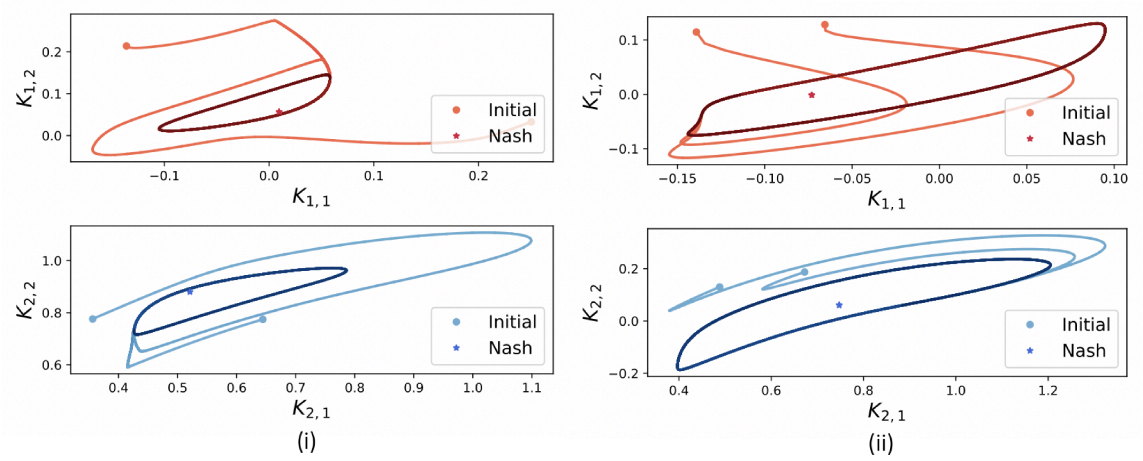

Figure 3: Parameter values of the two players in two general-sum LQ game where the Nash equilibrium is avoided by the gradient dynamics. We empirically observe in both games described in (5) that players converge to the same cycle from different initializations. Time is shown by the progressive darkening of the players' strategies.

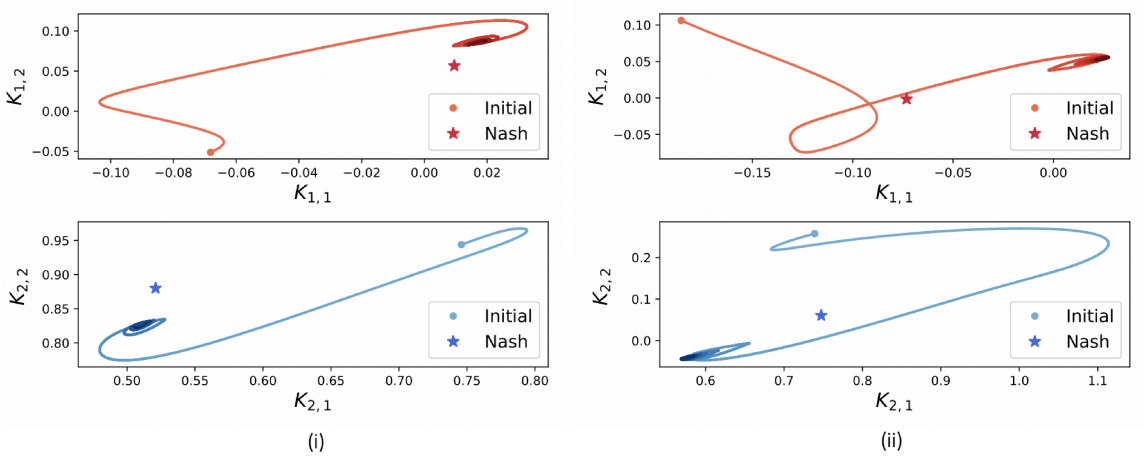

Figure 4: Time average parameter values of the two players in the general-sum LQ game with dynamics given in (5). We empirically observe that in both games the players' time average strategy does not converge to the Nash equilibrium strategy. Time is shown by progressive darkening of the players' strategies.

Since linear dynamics, quadratic costs, and linear policies are a relatively simple setup compared to many recent deep multi-agent reinforcement learning problems (Bansal et al., 2018; Jaderberg et al., 2019), we believe that the issues of non-convergence are likely to be present in more complex scenarios involving more complex dynamics and parametrizations of the policies. This can be viewed as a cautionary note, but it also suggests that the algorithms that have yielded impressive results in multi-agent settings can be further improved by leveraging the underlying game-theoretic structure.

We remark that we only analyzed the deterministic policy-gradient setting, though the findings extend to settings in which players construct unbiased estimates of their gradients (Sutton and Barto, 2017) and even actor-critic methods (Srinivasan et al., 2018). Indeed all of these algorithms will suffer the same problems since they all seek to track the same limiting continuous-time dynamical system (Mazumdar et al., 2018).

Our numerical experiments also highlight the existence of limit cycles in the policy-gradient dynamics. Unlike in classical optimization settings in which oscillations are normally caused by the choice of step sizes, the cycles we highlight are behaviors that can occur even with arbitrarily small step sizes. They are a fundamental feature of learning in multi-agent settings and have been observed in the dynamics of many learning algorithms (Mazumdar et al., 2018; Papadimitriou and Piliouras; Hommes and Ochea, 2012; Mertikopoulos et al., 2018). We remark, however, that there is no obvious link between the limit cycles that arise in the gradient dynamics of the LQ games and the Nash equilibrium of the game. Indeed, unlike with other game dynamics in more simple games, such as the well-studied replicator dynamics in bilinear games (Mertikopoulos et al., 2018) or multiplicative weights in rock-paper-scissors (Hommes and Ochea, 2012), the time average of the players' strategies does not coincide with the Nash equilibrium. This may be due to the fact that the Nash equilibrium is a saddle point of the gradient dynamics and not simply marginally stable, though the issue warrants further investigation.

This paper highlights how algorithms developed for classical optimization or single-agent optimal control settings may not behave as expected in multi-agent and competitive environments. Algorithms and approaches that have provable convergence guarantees and performance in competitive settings, while retaining the scalability and ease of implementation of simple policy-gradient methods, are therefore a crucial and promising open area of research.

Appendix A. Proofs of Auxiliary Results

Proposition 5 *Consider an N–player LQ game. The vector-valued map ω twice continuously differentiable on $\mathcal{S}^{md}$; i.e., $\omega \in C^2(\mathcal{S}^{md}, \mathcal{S}^{md})$.*

Proof Following the proof technique of Zhang et al. (2019), we show the regularity of ω using the implicit function theorem (Abraham et al., 1988). In particular, we show that $\Sigma_K = \mathbb{E}_{z_0 \sim \mathcal{D}_o}\left[\sum_{t=0}^{\infty} z(t) z(t)^T\right]$ and P_i for $i \in \{1, \ldots, N\}$ are C^1 with respect to each K_i on the space of stabilizing matrices.

For any stabilizing $(K_1, \ldots, K_N)$, Σ_K is the unique solution to the following discrete-time Lyapunov equation:

$$\bar{A} \Sigma_K \bar{A}^T + \Sigma_0 = \Sigma_K, \tag{6}$$

where $\Sigma_0 = \mathbb{E}_{z_0 \sim \mathcal{D}_o}[z(0) z(0)^T] > 0$ and $\bar{A} = A - \sum_{i=1}^{N} B_i K_i$. Both sides of this expression can be vectorized. Indeed, using the same notation as in Zhang et al. (2019), let $\mathrm{vect}(\cdot)$ be the map that vectorizes its argument and let $\Psi : \mathbb{R}^{m^2} \times \mathbb{R}^{d_1 \times m} \times \cdots \times \mathbb{R}^{d_N \times m} \to \mathbb{R}^{m^2}$ be defined by

$$\Psi(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_N) = \left[\bar{A} \otimes \bar{A}\right] \cdot \mathrm{vect}(\Sigma_K) + \mathrm{vect}(\Sigma_0).$$

Then, (6) can be written as

$$\begin{aligned} F(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_N) &= \Psi(\mathrm{vect}(\Sigma_K), K_1, \ldots, K_N) - \mathrm{vect}(\Sigma_K) \\ &= 0. \end{aligned}$$

The map F implicitly defines Σ_K. Moreover, letting I denote the appropriately sized identity matrix, we have that

$$\frac{\partial F(\text{vect}(\Sigma_K), K_1, \ldots, K_N)}{\partial \text{vect}^T(\Sigma_K)} = \left[\bar{A} \otimes \bar{A}\right] - I.$$

For stabilizing $(K_1, \ldots, K_N)$, this matrix is an isomorphism since $\text{spec}(\bar{A})$ is inside the unit circle. Thus, using the implicit function theorem, we conclude that $\text{vect}(\Sigma_K) \in C^1$. As noted in Zhang et al. (2019), the proof for each P_i, $i \in \{1, \ldots, N\}$ is completely analogous. Since Σ_K and P_i are C^1 and ω is linear in these terms, the result of the proposition follows. ■

Appendix B. Additional Mathematical Preliminaries and Results

The following theorem is the celebrated center manifold theorem from geometry. We utilize it in showing avoidance of saddle point equilibria of the dynamics.

Theorem 9 (Stable Manifold Theorem (Shub, 1978, Thm. III.7), Smale (1967)) *Let x_0 be a fixed point for the C^r local diffeomorphism $\phi : U \to \mathbb{R}^n$ where $U \subset \mathbb{R}^n$ is an open neighborhood of x_0 in $\mathbb{R}^n$ and $r \geq 1$. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^n$ into generalized eigenspaces of $D\phi(x_0)$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $D\phi(x_0)$ invariant subspace $E^s \oplus E^c$ there is an associated local ϕ–invariant C^r embedded disc W^{cs}_{loc} called the local stable center manifold of dimension $\dim(E^s \oplus E^c)$ and ball B around x_0 such that $\phi(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}$, and if $\phi^n(x) \in B$ for all $n \geq 0$, then $x \in W^{sc}_{loc}$.*

References

R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. Springer, 1988.

T. Bansal, J. Pachocki, S. Sidor, I. Sutskever, and I. Mordatch. Emergent complexity via multi-agent competition. In *International Conference on Learning Representations*, 2018.

T. Basar. On the uniqueness of the nash solution in linear-quadratic differential games. *International Journal of Game Theory*, 1976.

T. Basar and G. Olsder. *Dynamic Noncooperative Game Theory*. Society for Industrial and Applied Mathematics, 2 edition, 1998.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Traning GANs with Optimism. *arxiv:1711.00141*, 2017.

S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *ArXiv e-prints*, 2017.

J. Engwerda. On scalar feedback nash equilibria in the infinite horizon lq-game. *IFAC Proceedings Volumes*, 1998.

M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

C. H. Hommes and M. I. Ochea. Multiple equilibria and limit cycles in evolutionary games with logit dynamics. *Games and Economic Behavior*, 74, 2012.

M. Jaderberg, W. Czarnecki, I. Dunning, L. Marris, G. Lever, A. Garcia Castaneda, C. Beattie, N. C. Rabinowitz, A. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, and T. Graepel. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364, 2019.

R. E. Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5, 1960.

M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Perolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems 30*. 2017.

T. Li and Z. Gajic. Lyapunov iterations for solving coupled algebraic Riccati equations of Nash differential games and algebraic Riccati equations of zero-sum games. In *New Trends in Dynamic Games and Applications*, 1995.

R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems 30*. 2017.

D. L. Lukes and D. L. Russell. A global theory for linear-quadratic differential games. *Journal of Mathematical Analysis and Applications*, 1971.

D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *Proceedings of Machine Learning Research*, 2019.

E. Mazumdar, L. J. Ratliff, and S Sastry. On the convergence of gradient-based learning in continuous games. *ArXiv e-prints*, 2018.

E. Mazumdar, M. I. Jordan, and S. S. Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *CoRR*, 2019.

P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

C. Papadimitriou and G. Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*.

C. Possieri and M. Sassano. An algebraic geometry approach for the computation of all linear feedback Nash equilibria in lq differential games. In *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015.

L. J. Ratliff, S. A. Burden, and S. S. Sastry. Characterization and computation of local Nash equilibria in continuous games. In *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*, pages 917–924, Oct 2013.

M. Shub. *Global Stability of Dynamical Systems*. Springer-Verlag, 1978.

S. Smale. Differentiable dynamical systems. *Bull. Amer. Math. Soc.*, 73, 1967.

S. Srinivasan, M. Lanctot, V. Zambaldi, J. Perolat, K. Tuyls, R. Munos, and M. Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural Information Processing Systems 31*. 2018.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2017.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wunsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, Oct 2019.

K. Zhang, Z. Yang, and T. Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games, 2019.