Policy Gradient in Linear Quadratic Dynamic Games Has No Convergence Guarantees

Eric V. Mazumdar
Electrical Engineering & Computer Science
University of California, Berkeley
mazumdar@berkeley.edu

Lillian J. Ratliff
Electrical & Computer Engineering
University of Washington
ratliffl@uw.edu

S. Shankar Sastry
Electrical Engineering & Computer Science
University of California, Berkeley
sastry@coe.berkeley.edu

Michael I. Jordan
Computer Science & Statistics
University of California, Berkeley
jordan@cs.berkeley.edu

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 (LQ) games. In such games, the state and action spaces are continuous and a global Nash equilibrium can be found by solving coupled Ricatti equations. Further, gradient-play in LQ games is equivalent to multi-agent policy-gradient. We first prove that despite the non-convexity of the players' objectives, the only critical points of the gradient dynamics in these games 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 LQ games with Nash equilibria that satisfy these conditions. 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 guarantee of convergence in multi-agent settings. Moreover, the ease with which we can generate these counterexamples suggests that such situations 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 [9, 12, 5], multi-agent proximal policy optimization [1], and even simple multi-agent policy-gradients [7] in problems where the various agents have high-dimensional continuous state and action spaces.

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 [6]. 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

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

the problem has allowed for new insights into the behavior of reinforcement learning algorithms in continuous action and state spaces [3, 4, 10].

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 [2]. 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.

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 only global Nash equilibrium 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. In the single-agent setting, it was recently shown that policy-gradient has global convergence guarantees for the LQR problem [4]. These results have recently been extended to projected policy-gradient algorithms in zero-sum LQ games [13].

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. In particular, we show that if each player randomly initializes their policy and then uses a policy-gradient algorithm there exists an LQ game with a Nash equilibrium from which the players would diverge when randomly initialized arbitrarily close to the 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 equilibrium of the game.

2 Preliminaries

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

$$z(t+1) = Az(t) + \textstyle\sum_{i=1}^{N} B_i u_i(t) \quad z(0) = z_0 \sim \mathcal{D}_0, \tag{1}$$

where $z(t) \in \mathbb{R}^m$ is the state at time t, $\mathcal{D}_0$ is the initial state distribution, and $u_i(t) \in \mathbb{R}^{d_i}$ is the control input of player $i \in 1, \dots, N$. For LQ games, it is known that under reasonable assumptions, linear feedback policies for each player that constitute a Nash equilibrium exist. [2]. 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

$$f_i(u_1, \dots, u_N) = \mathbb{E}_{z_0 \sim \mathcal{D}_0}\left[\textstyle\sum_{t=0}^{\infty} z(t)^T Q_i z(t) + u_i(t)^T R_i u_i(t)\right],$$

where Q_i and R_i are the cost matrices for the state and input, respectively. Throughout our work we make the following assumptions which guarantee that a Nash equilibrium exists [2].

Assumption 1. *For each player $i \in \{1, \dots, N\}$, the state and control cost matrices satisfy $Q_i \succ 0$ and $R_i \succ 0$. Further, at least one of $(A, B_i, \sqrt{Q_i})$ is stabilizable-detectable.*

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 form:

$$K_{i,n+1} = K_{i,n} - \gamma_i D_i f_i(K_{1,n}, \dots, K_{N,n}) \quad K_{i,0} \sim \mathcal{D}_i \tag{2}$$

where $D_i f_i(\cdot, \cdot)$ denotes the derivatives of f_i with respect to the i–th argument, $\{\gamma_i\}_{i=1}^{N}$ are the step-sizes of the players, and $\mathcal{D}_i$ is player i's initial distribution over policies. 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, we define $\Sigma_K = \mathbb{E}_{z_0 \sim \mathcal{D}_0}\left[\sum_{t=0}^{\infty} z(t)z(t)^T\right]$. We also denote the covariance of the initial state as $\Sigma_0 = \mathbb{E}_{z_0 \sim \mathcal{D}_0} z(0)z(0)^T$. Direct computation verifies that for player i, $D_i f_i$ is given by:

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

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

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

Given that the players may have different control objectives and do not engage in coordination or cooperation, the natural solution concept is that of a Nash equilibrium.

Definition 1. *A* feedback Nash equilibrium *is a collection of policies $(K_1^*, \dots, K_N^*)$ such that, for each $i \in \{1, \dots, N\}$:*

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

Under Assumption 1, a Nash equilibrium of an LQ game is known to exist and can be found by solving coupled Ricatti equations using the method of Lyapunov iterations. The method is outlined in [8] for continuous time LQ games, and an analogous procedure can be followed for discrete time. Further information on the uniqueness of Nash equilibria in LQ games and the method of Lyapunov iterations can be found in [2] and [8], respectively.

3 Results

Given our setup, we first show that the critical points of the gradient dynamics in LQ games are all Nash equilibria. We then give sufficient conditions under which Nash equilibria will be almost surely avoided. We conclude by showing empirical results that: 1. highlight the frequency of LQ games that admit Nash equilibria which are avoided by policy gradient, and 2. show convergence to limit cycles that have no link to the equilibria of the game.

3.1 All critical points of gradient-play in LQ games are Nash equilibria

Our first theoretical result is on the critical points of gradient play 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 $D_i f_i = \partial f_i / \partial K_i$ and K_i have both been converted into md_i dimensional vectors.

Critical points of gradient-play are strategies $x = (K_1, \dots, K_N)$ such that $\omega(x) = 0$. Recent work has shown that when players perform gradient descent on their own cost functions in general-sum games they may converge to critical points that are not Nash equilibria [11]. The following theorem shows that *despite the non-convexity of each players' individual objective* (see e.g. [4]), such non-Nash equilibria cannot exist in the gradient dynamics of LQ games.

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

Theorem 1 shows that, just as in the single-player LQR setting and zero-sum LQ games [13], 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 $\mathcal{D}_0$ to have a full-rank covariance matrix.

Remark 1. *We remark that a simple consequence of Theorem 1 is that when the Nash equilibrium of the LQ game is unique and Σ_0 is full rank, the gradient dynamics admit a unique critical point.*

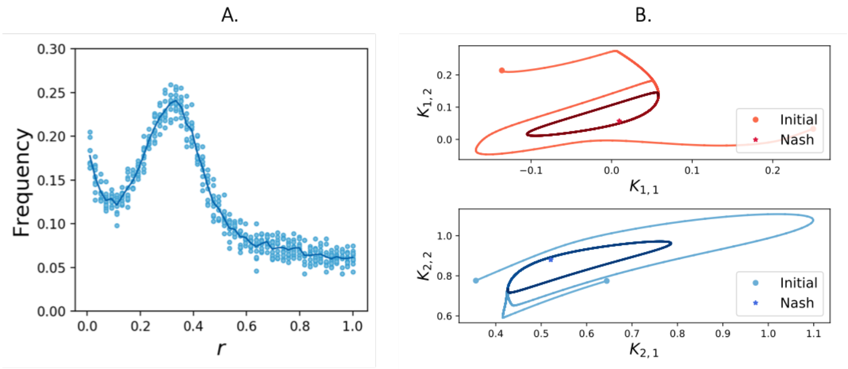

Figure 1: A. Each point represents, for the given parameter value of r, the frequency of LQ games out of 1000 randomly sampled A matrices where policy gradient would avoid a Nash equilibrium. The solid line shows the average frequency of these games. B. In one of the randomly generated LQ games, the two players converge to limit cycles where the Nash strategy is not the time-average.

3.2 Policy-gradient avoids Nash Equilibria that are saddle points of the dynamics

Given that the Nash equilibria are the only critical points of the gradient dynamics in LQ games, we now give sufficient conditions under which gradient-play has no guarantees of even *local*, convergence to these points. We first show that ω is sufficiently smooth on the set of stabilizing policies.

Proposition 1. *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., $\omega \in C^2(\mathcal{S}^{md}, \mathcal{S}^{md})$.*

With the above smoothness result in hand we now give sufficient conditions under which the set of initial conditions from which policy gradient 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 uniformly in a small ball around it. Let the Jacobian of the vector field ω be denoted by $D\omega$. Given a point x, let λ_j be the eigenvalues of $D\omega(x)$, for $j \in \{1, \dots, md\}$, where $d = \sum_{i=1}^n d_i$.

Theorem 2. *Suppose $\mathcal{D}_0$ is chosen such that $\Sigma_0 \succ 0$, and consider an N–player LQ game satisfying Assumption 1 in which the Nash equilibrium is a saddle point of the policy-gradient dynamics: i.e. the Jacobian of ω evaluated at the Nash equilibrium $x^* = (K_1^*, \dots, K_N^*)$ has eigenvalues λ_j is such that $Re(\lambda_j) < 0$ for $j \in \{1, \dots, \ell\}$ and $Re(\lambda_j) > 0$ for $j \in \{\ell + 1, \dots, md\}$ for some ℓ such that $0 < \ell < md$. If each player $i \in \{1, \dots, N\}$ performs policy-gradient with a random initial strategy $K_{i,0} \sim \mathcal{D}_i$ such that the support of $\mathcal{D}_i$ is U, they will almost surely avoid the Nash equilibrium.*

Theorem 2 gives us sufficient conditions under which policy-gradient in general-sum LQ games does not even have *local 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 [4]. In zero-sum LQ games, the structure of the game also precludes any Nash equilibrium from satisfying the conditions of Theorem 2 [11], meaning that local convergence is always guaranteed. In [13], 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.

In Figure 1A, we show the frequency of two-player LQ games in $\mathbb{R}^2$ that admit Nash equilibria that satisfy the conditions of Theorem 2. We fix $\Sigma_0, B_1, B_2, Q_1, Q_2$, and R_1 and set $R_2 = r$. For each value of r we randomly sample 1000 different values of A, find a Nash equilibrium of the game using Lyapunov iterations, and then numerically approximate $D\omega$ to see if the point would be avoided by policy gradient. We find that up to 25% of such randomly sampled LQ games admit Nash equilibria that are saddle points of the policy gradient dynamics. In Figure 1B we show that when policy gradient is used in these games the players converge to stable limit cycles for which the time-average is not Nash. As such, policy gradient *does not even converge in a time-averaged sense.*

These theoretical and numerical results imply that policy-gradient algorithms have no guarantees of local, and consequently global, convergence in general-sum LQ games. Further, they highlight that these issues are not mere edge cases and may be quite common in practice.

References

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

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

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

[4] 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.

[5] M. Jaderberg, W. Czarnecki, I. Dunning, L. Marris, G. Lever, A. Garcia Castaeda, 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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