Proceedings of Machine Learning Research vol 283:1–14, 2025 7th Annual Conference on Learning for Dynamics and Control

Scalable Natural Policy Gradient for General-Sum Linear Quadratic Games with Known Parameters

Mostafa M. Shibl MABDELNA@PURDUE.EDU
Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Wesley A. Suttle WESLEY.A.SUTTLE.CTR@ARMY.MIL
U.S. Army Research Laboratory, Adelphi, MD, USA

Vijay Gupta GUPTA869@PURDUE.EDU
Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

Consider a general-sum N-player linear-quadratic (LQ) game with stochastic dynamics over a finite time horizon. It is known that under some mild assumptions, the Nash equilibrium (NE) strategies for the players can be obtained by a natural policy gradient algorithm. However, the traditional implementation of the algorithm requires the availability of complete state and action information from all agents and may not scale well with the number of agents. Under the assumption of known problem parameters, we present an algorithm that assumes state and action information from only neighboring agents according to the graph describing the dynamic or cost coupling among the agents. We show that the proposed algorithm converges to an ϵ-neighborhood of the NE where the value of ϵ depends on the size of the local neighborhood of agents.

Keywords: Linear quadratic games, multi-agent systems, learning in games, Nash equilibria.

1. Introduction

Multi-agent systems with self-interested agents interacting through coupled dynamics, costs, or constraints are crucial in various fields. This work focuses on linear-quadratic (LQ) games, an extension of the classical LQ regulator and cooperative distributed LQ regulator problems. LQ games involve a linear time-invariant system controlled by all agents, where each agent aims to minimize its own quadratic cost. For foundational assumptions and theory behind equilibria in LQ games, including the conditions for the uniqueness and existence of Nash equilibria (NE), we refer the reader to Basar and Olsder (1999).

Designing optimal equilibrium strategies in multi-agent systems is challenging. The field of learning in games, particularly with advancements in multi-agent reinforcement learning (MARL), provides a robust framework for this task. The literature in this field is far too numerous to be summarized. However, for comprehensive overviews, we point to surveys such as Busoniu et al. (2008); Li et al. (2022); Yang and Wang (2021); Zhu et al. (2024); Zhang et al. (2021); Canese et al. (2021). As representative examples, policy gradient methods in MARL has shown great success. For instance, in two-player, zero-sum stochastic games Daskalakis et al. (2021), natural policy gradient for constrained nonconcave maximization problems Panageas et al. (2019), neural fictitious play for approximating Nash equilibria in games with imperfect information Heinrich and Silver (2016), and gradient-based learning methods for differentiable games Balduzzi et al. (2018).

In the specific context of LQ games that we consider with known system matrices, one class of algorithms is based on iteration of non-linear equations. For LQ differential games, Riccati-based methods converge to local open-loop and feedback NE in two-player and N-player settings Scarpa and Mylvaganam (2023); Scarpa et al. (2024); Sassano et al. (2025), with reduced complexity in potential games Scarpa and Mylvaganam (2023). Extensions include iterative, data-driven Lyapunov and Riccati-based algorithms for nonzero-sum LQ games in infinite-horizon, discrete-time settings under specified assumptions Nortmann et al. (2024); Nortmann and Mylvaganam (2023); Monti et al. (2024). With unknown system matrices, MARL-based methods are more suitable, such as policy optimization for zero-sum LQ games Zhang et al. (2019), nonzero-sum games with structured interactions Roudneshin et al. (2020), and gradient ascent for general-sum games Song et al. (2019). Additionally, LQ games with a large number of agents are often modeled as continuous-time LQ mean-field games Wang et al. (2021).

This paper focuses particularly on natural policy gradient methods, known for their good convergence rate and applicability across discrete and continuous state and action spaces Kakade and Langford (2002); Kakade (2001); Mnih et al. (2015); Agarwal et al. (2020). In LQ games, policy gradient methods can fail to reach Nash equilibria in general-sum games under deterministic dynamics Mazumdar et al. (2020), but natural policy gradient converges under stochastic dynamics Hambly et al. (2023).

Natural policy gradient algorithms, like many MARL methods, has the underlying assumption that the states and actions of all agents are available at every other agent, which might not be scalable for large scale systems. If agents instead access only their neighbors' states and actions (defined by a coupling graph of dynamics and costs), they can still converge to a neighborhood of equilibrium policies since they should have access to the 'most important information' for the design of their local policies. However, a precise characterization of this intuition has only now begun to arise. In networked Markov decision processes with finite state-action spaces, exponential decay property that quantifies how the effect of distant agents on each other diminishes with their graph distance allows MARL algorithms to rely on local neighborhood information Qu et al. (2020b,a), converging to a neighborhood of the equilibrium policies Shibl and Gupta (2024). Similar results hold for networked systems with spatially decaying dynamics where the effect of a control action decays exponentially with distance Shin et al. (2022, 2023) and cooperative LQ setups (where agents are not self-interested but wish to minimize a team cost function) Olsson et al. (2024).

In this paper, we seek to answer whether such a result is possible for natural policy gradient algorithm in general-sum N-player LQ games with restricted availability of state and action information to be from a neighborhood according to a coupling graph. We redesign the algorithm to ensure agent policies converge to an ϵ-neighborhood of the Nash equilibrium (NE) for scalability as the number of agents grows. Key contributions include developing a scalable distributed policy learning algorithm in LQ games, leveraging local observability, proving convergence to a neighborhood around a NE, and bounding the size of the neighborhood in terms of the problem parameters.

The paper is organized as follows. Section 2 introduces the model used. Our algorithm is proposed and analyzed in Section 3. Section 4 applies the algorithm to a numerical example. Section 5 concludes the paper with some future directions.

Notation: $\mathbb{R}$ and $\mathbb{I}$ denote the set of real numbers and the set of integers, respectively. $\mathbf{A} \in \mathbb{R}^{n\times n}$ denotes a real matrix $\mathbf{A}$ of dimensions $n \times n$, while $x \in \mathbb{R}^n$ denotes a real n-dimensional column vector x. We denote $\mathbf{A}(i, j)$ as the i-th block row and j-th block column of $\mathbf{A}$, where the block dimensions are clear from the context. Apart from this usage, superscripts will be used to denote the agent index, and subscripts will be used to denote time. Vector 2-norms and induced 2-norms of matrices are denoted by $\|\cdot\|$. $\mathbf{A} \succ \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is positive definite, and $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. $Tr(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. $\sigma(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ denotes the singular values of $\mathbf{A}$ and the smallest singular value of $\mathbf{A}$, respectively. For any $L > 0$ and $\alpha \in [0, 1)$, a matrix $\mathbf{\Phi}$ is (L, α)-stable if $\|\mathbf{\Phi}^t\| \leq L\alpha^t$ for $t > 0$. The matrix $\mathbf{0}$ denotes the zero matrix and $\mathbf{I}$ denotes the identity matrix with dimensions clear from context. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a pair of a node set $\mathcal{V}$ and an (undirected) edge set $\mathcal{E}$. The distance between i and j on graph $\mathcal{G}$, denoted by $d(i, j)$, is the number of edges in the shortest path connecting i and j, where $i, j \in \mathcal{V}$. $\mathbb{E}$ denotes the expectation operator. We define $\mathbb{I}_{[1,N]} := \{1, 2, ..., N\}$. Finally, we denote $\mathcal{J}_t$ as a time varying set.

2. Model

2.1. N-Player General-Sum Linear-Quadratic Games

We consider a non-cooperative, finite-horizon, general sum LQ game with N agents. Associate each agent with a node in a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ in which $\mathcal{N} := \{0, 1, ..., N-1\}$ is the node set and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ represents the set of undirected edges. An edge $(i, j) \in \mathcal{E}$ indicates the coupling between agents i and j through their dynamics and/or cost as defined below. For each agent i, denote the state by $x^i \in \mathbb{R}^{n^i}$ and the control input by $u^i \in \mathbb{R}^{k^i}$. The state of each agent i evolves as

$$x^i_{t+1} = \sum_{j=0}^{N-1} \mathbf{A}(i,j) x^j_t + \sum_{j=0}^{N-1} \mathbf{B}(i,j) u^j_t + w^i_t, \qquad x^i_0, \tag{1}$$

where w^i_t is the process noise. Non-zero matrices $\mathbf{A}(i, j)$ and $\mathbf{B}(i, j)$ represent the dynamics coupling between agents i and j. We can stack the agent states, control inputs, and process noises into system state, control, and process noise vectors $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^k$, and w_t respectively. By considering $\mathbf{A}(i, j)$ and $\mathbf{B}(i, j)$ as the (i, j)-th blocks of a matrix, we can define the system transition matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B}^i \in \mathbb{R}^{n \times k^i}$, and $\mathbf{B} \in \mathbb{R}^{n \times k}$. We make the following assumption.

Assumption 1 *The initial condition x_0 is a zero-mean Gaussian variable with positive definite covariance matrix $\mathbb{E}[x_0 x_0^\top] \succ 0$. Further, the process noise vectors $\{w_t\}_{t=0}^T$ are independent and identically distributed zero-mean Gaussian random variables with positive definite covariance matrix $\mathbf{W} = \mathbb{E}[w_t w_t^\top] \succ 0$ and independent from x_0, for $t = 0, 1, ..., T$.*

The per time step cost of interest to each agent i is given by

$$J^i_t(x, u) = \begin{cases} x_t^\top \mathbf{Q}^i x_t + (u^i_t)^\top \mathbf{R}^i u^i_t & \text{if } t \neq T, \\ x_t^\top \mathbf{Q}^i x_t & \text{if } t = T, \end{cases} \tag{2}$$

for cost parameterization matrices $\mathbf{Q}^i$ and $\mathbf{R}^i$ satisfying $\mathbf{Q}^i \succeq \frac{L^3(1+L^2)}{(1-\alpha)^2}\mathbf{I}$ and $\mathbf{R}^i \succeq \gamma \mathbf{I}$, where L and α are defined in Assumption 2, and $\gamma \in (0, 1)$. If the appropriate block $\mathbf{Q}^i(i, j)$ and / or $\mathbf{R}^i(i, j)$ (defined in the same manner as the blocks of $\mathbf{A}$ and $\mathbf{B}$) is non-zero, we say that the agents i and j are coupled through their cost functions. The objective function for agent i is to minimize the expected finite horizon sum defined in (3), where the expectation is taken over the system noise and initial state distribution.

$$\begin{aligned} \text{Objective function for agent } i: \qquad & \text{minimize} \quad \mathbb{E}\left[\sum_{t=0}^{T} J_t^i(x,u)\right] \\ & \text{subject to} \quad x_{t+1} = \mathbf{A}x_t + \sum_{i=0}^{N-1} \mathbf{B}^i u_t^i + w_t. \end{aligned} \tag{3}$$

The following definitions will be used in the paper.

Definition 1 *Consider a linear time-invariant system with system matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$*. For any* $L > 0$ *and* $\alpha \in [0, 1)$*,* $(\boldsymbol{A}, \boldsymbol{B})$ *is* (L, α)*-stabilizable if* $\exists \boldsymbol{K} : \|\boldsymbol{K}\| \leq L$ *and* $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}$ *is* (L, α)*-stable. Further,* $(\boldsymbol{A}, \boldsymbol{B})$ *is* (L, α)*-detectable if* $(\boldsymbol{A}^\top, \boldsymbol{B}^\top)$ *is* (L, α)*-stabilizable.*

Definition 2 (Definition 3.3 in Shin et al. (2022)) *Consider a matrix* $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$*, a graph* $\mathcal{G} := (\mathcal{V}, \mathcal{E})$*, and index sets* $\mathcal{I} := \{I_i\}_{i \in \mathcal{V}}$*,* $\mathcal{J} := \{J_i\}_{i \in \mathcal{V}}$ *that partition* $\mathbb{I}_{[1,m]}$*,* $\mathbb{I}_{[1,n]}$*, respectively. We say* $\mathbf{\Phi}$ *induced by* $(\mathcal{G}, \mathcal{I}, \mathcal{J})$ *has bandwidth* B*, if* B *is the smallest non-negative integer satisfying* $\mathbf{\Phi}(I_i, J_j) = \boldsymbol{0}$ *for any* $i, j \in \mathcal{V}$ *with distance* $d(i, j) > B$*.*

We make the following assumptions that are standard in the LQ games setting.

Assumption 2 *For* $i = 0, 1, ..., N-1$*, we assume that* $\|\boldsymbol{A}\|, \|\boldsymbol{B}^i\|, \|\boldsymbol{Q}^i\|, \|\boldsymbol{R}^i\| \leq L$*. This helps in limiting natural system growth and ensures effective and non-excessive control authority. Further, we assume that the system is stabilizable considering* u *as the control input in that* $(\boldsymbol{A}, \boldsymbol{B})$ *is* (L, α)*-stabilizable. Finally, we assume that the pair* $(\boldsymbol{A}, (\boldsymbol{Q}^i)^{1/2})$ *for each* i *is* (L, α)*-detectable.*

While L and α are not unique, there are unique minimum values $L_{\min}$ and $\alpha_{\min}$. Our results hold for any valid L and α; however, the tightest bounds result from $L_{\min}$ and $\alpha_{\min}$. Also, it is noteworthy to mention that we do not assume that $L_{\min}$ is less than 1.

Assumption 3 *We assume that for* $i = 0, 1, ..., N-1$ *there exists a unique solution* $\{\boldsymbol{K}_t^{i*}\}_{t=0}^{T-1}$*, to the following discrete algebraic Riccati equations:*

$$\boldsymbol{K}_t^{i*} = \left(\boldsymbol{R}^i + (\boldsymbol{B}^i)^\top \boldsymbol{P}_{t+1}^{i*} \boldsymbol{B}^i\right)^{-1} (\boldsymbol{B}^i)^\top \boldsymbol{P}_{t+1}^{i*} \left(\boldsymbol{A} - \sum_{j=1, j \neq i}^{N} \boldsymbol{B}^j \boldsymbol{K}_t^{j*}\right), \tag{4}$$

where $\{\boldsymbol{P}_t^{i*}\}_{t=0}^{T}$ *are obtained recursively backwards from*

$$\boldsymbol{P}_t^{i*} = \boldsymbol{Q}^i + (\boldsymbol{K}_t^{i*})^\top \boldsymbol{R}^i \boldsymbol{K}_t^{i*} + \left(\boldsymbol{A} - \sum_{j=1}^{N} \boldsymbol{B}^j \boldsymbol{K}_t^{j*}\right)^\top \boldsymbol{P}_{t+1}^{i*} \left(\boldsymbol{A} - \sum_{j=1}^{N} \boldsymbol{B}^j \boldsymbol{K}_t^{j*}\right), \tag{5}$$

with terminal condition $\boldsymbol{P}_T^{i*} = \boldsymbol{Q}^i$*.*

In general, an LQ game may have multiple NE. Lemma 3 provides a sufficiency condition for the existence of the unique solution in Assumption 3 that will be referred to as *the NE* in the sequel.

Lemma 3 (Remark 6.5 and Corollary 6.4 in Basar and Olsder (1999)) *For the problem posed above, define the block matrix $\mathbf{\Phi_t}$, for $t = 0, 1, ..., T-1$, with the (i,i)-th block given by $\boldsymbol{R}^i + (\boldsymbol{B}^i)^\top \boldsymbol{P}^{i*}_{t+1}\boldsymbol{B}^i$ and the (i,j)-th block $(i \neq j)$ given by $(\boldsymbol{B}^i)^\top \boldsymbol{P}^{i*}_{t+1}\boldsymbol{B}^j$, for $i, j \in \{0, 1, ..., N-1\}$, and with $\boldsymbol{P}^{i*}_{t+1}$ defined in Assumption 3. Then, a sufficient condition for the existence of a unique solution of (4) is the non-singularity of the block matrix $\mathbf{\Phi_t}$, for $t = 0, 1, ..., T-1$. Further, if such a unique solution exists, then there is a unique NE for the LQ game with*

$$u^{i*}_t = -\boldsymbol{K}^{i*}_t x_t, \qquad \forall t = 0, 1, ..., T-1,$$

where $\boldsymbol{K}^{i*}_t$ is defined in (4). At this NE, the cost (3) for agent i is given by $\mathbb{E}[x_0^\top \boldsymbol{P}^{i*}_0 x_0 + N^{i*}_0]$, where $\{\boldsymbol{P}^{i*}_t\}_{t=0}^T$ are defined in (5) and

$$N^{i*}_t = N^{i*}_{t+1} + \mathbb{E}[w_t^\top \boldsymbol{P}^{i*}_{t+1} w_t] = N^{i*}_{t+1} + Tr(\boldsymbol{W}\boldsymbol{P}^{i*}_{t+1}), \qquad \forall t = 0, 1, ..., T-1$$

with terminal condition $N^{i*}_T = 0$.

This result implies that to obtain the NE of the LQ game, we can focus on linear feedback policies $u^i_t = -\mathbf{K}^i x_t$, $\forall i = 0, 1, ..., N-1$, for the time-invariant case, which means that $\mathbf{A}, \mathbf{B}, \mathbf{Q}$, and $\mathbf{R}$ are not time dependent.

2.2. Natural Policy Gradient Algorithm

In order to obtain the optimal $\mathbf{K}^i$, we assume that all agents utilize natural policy gradient algorithm (Algorithm 1 in Hambly et al. (2023)), which is known to converge to the NE in general sum N-player LQ games, under Assumption 4 for the system noise Hambly et al. (2023). Assumption 4 intuitively means that the system requires a certain level of noise for exploration. It should be mentioned that natural policy gradient algorithm in Hambly et al. (2023) is the only algorithm that has guaranteed convergence in the general sum N-player stochastic LQ game setting.

Assumption 4 (Assumption 3.3 in Hambly et al. (2023)) *The system parameters satisfy the following inequality for some small constant $\delta > 0$ and initial controller gain $\boldsymbol{K}^{i,(0)}$:*

$$\frac{(\underline{\sigma}_{\boldsymbol{X}})^5}{\|\Sigma_{\boldsymbol{K}^*}\|} > 20(N-1)^2\, T^2\, n\, \frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\} + \theta)^4}{\underline{\sigma}_{\boldsymbol{Q}}^2 \underline{\sigma}_{\boldsymbol{R}}^2} \left(\frac{\bar{\rho}^{2T} - 1}{\bar{\rho}^2 - 1}\right)^2,$$

where $\underline{\sigma}_{\boldsymbol{X}} := \min\{\sigma_{\min}(\mathbb{E}[x_0 x_0^\top]), \sigma_{\min}(\boldsymbol{W})\}$, $\underline{\sigma}_{\boldsymbol{Q}} := \min_i\{\sigma_{\min}(\boldsymbol{Q}^i)\}$, $\underline{\sigma}_{\boldsymbol{R}} := \min_i\{\sigma_{\min}(\boldsymbol{R}^i)\}$, $\Sigma_{\boldsymbol{K}^*} := \sum_{t=0}^T \mathbb{E}[x_t^{\boldsymbol{K}^*}(x_t^{\boldsymbol{K}^*})^\top]$, $\gamma_B := \max_i\{\|\boldsymbol{B}^i\|\}$, $\bar{\rho} := \max\left\{\left\|\boldsymbol{A} - \sum_{i=0}^{N-1}\boldsymbol{B}^i\boldsymbol{K}^{i*}\right\|, 1+\delta\right\} + N\gamma_B\sqrt{\frac{T\theta}{\underline{\sigma}_{\boldsymbol{X}}\underline{\sigma}_{\boldsymbol{R}}}} + \frac{1}{20T^2}$, $C^i(\boldsymbol{K}) := \mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top \boldsymbol{Q}^i x_t + (\boldsymbol{K}^i x_t)^\top \boldsymbol{R}^i(\boldsymbol{K}^i x_t)\right) + x_T^\top \boldsymbol{Q}^i x_T\right]$, and $\theta := \max_i\{C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\}$.

3. Proposed Algorithm

The natural policy gradient algorithm in Hambly et al. (2023) assumes that all agents have access to the states and actions of all other agents, which may not be feasible for large scale problems. In Algorithm 1, we propose an algorithm where each agent only has access to the states and actions of agents within its κ-hop neighborhood, $\mathcal{N}_i^\kappa$, according to the graph $\mathcal{G}$. The κ-hop neighborhood consists of agents whose graph distance to agent i is less than or equal to κ. This means that agent i's control input at time t depends on the states and actions of agents j such that $d(i,j) \leq \kappa$. Thus, the natural policy gradient algorithm must be modified to rely solely on local information. To ensure the error from using only local data is bounded, we show that the dependency of agent i on the states and actions of distant agents decays exponentially with increasing graph distance, which is known as the exponential decay property. Theorem 4 formalizes this exponential decay property for the LQ game controller gain.

Theorem 4 *For the problem formulation in Section 2, it holds that* $\|\boldsymbol{K}^{i*}(i,j)\| \leq \beta\psi^{0.5d(i,j)}$ *for* $i,j \in \mathcal{N}$*, where* β*,* ψ*,* M*,* η*, and* ϕ *are defined below. Based on their definitions,* $\beta \geq 1$ *and* $\psi \in (0,1)$*, which satisfies the exponential decay property.*

$$\beta = \frac{M}{\phi^2} \cdot \psi^{0.5}, \quad \psi = \frac{M^2 - \eta^2}{M^2 + \eta^2}, \quad M = \max\Big(2L+1, \frac{L^3(1+L^2)}{1-\alpha^2}\Big),$$

$$\eta = \frac{\Big(\frac{4M^2L^4(1+L)^2}{(1-\alpha)^2\gamma} + \frac{(1-\alpha)^2\gamma}{2L^4(1+L)^2} + M^2\Big)L^2(1+L)^2}{(1-\alpha)^2},$$

$$\phi = \left(\frac{4L^4(1+L)^2}{(1-\alpha)^2\gamma} + \Big(1 + \frac{8ML^4(1+L)^2}{(1-\alpha)^2\gamma} + \frac{16M^2L^8(1+L)^4}{(1-\alpha)^4\gamma^2}\Big)\frac{M(1+\eta M)(1-\alpha)^2}{L^2(1+L)^2} + \eta\right)^{-1}.$$

Algorithm 1 Scalable Natural Policy Gradient Algorithm for N-player LQ Games

Input: Number of iterations M, time horizon T, initial policies $\boldsymbol{K}^{(0)} = (\boldsymbol{K}^{1,(0)}, ..., \boldsymbol{K}^{N,(0)})$, step size η, model parameters $\{\mathbf{A}\}_{t=0}^{T-1}$, $\{\mathbf{B}^i\}_{t=0}^{T-1}$, $\{\mathbf{Q}^i\}_{t=0}^{T}$, and $\{\mathbf{R}^i\}_{t=0}^{T-1}$ $(i = 0, ..., N-1)$.
for $m \in \{1, ..., M\}$ **do**
  for $t \in \{T-1, ..., 0\}$ **do**
    for $i \in \{1, ..., N\}$ **do**
      Calculate the matrix $\mathbf{P}_{t,i}^{\boldsymbol{K}^{(m-1)}}$ with $\mathbf{P}_{T,i}^{\boldsymbol{K}^{(m-1)}} = \mathbf{Q}^i$ by

$$\mathbf{P}_{t,i}^{\boldsymbol{K}^{(m-1)}} = \mathbf{Q}^i + (\mathbf{K}^{i,(m-1)})^\top \mathbf{R}^i \mathbf{K}^{i,(m-1)} + \Big(\mathbf{A} - \sum_{i \in \mathcal{N}_i^\kappa} \mathbf{B}^i K^{i,(m-1)}\Big)^\top \mathbf{P}_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \cdot \Big(\mathbf{A} - \sum_{i \in \mathcal{N}_i^\kappa} \mathbf{B}^i K^{i,(m-1)}\Big).$$

      Calculate the matrix E_t^i by

$$\mathbf{E}_{t,i}^{\boldsymbol{K}^{(m-1)}} = \mathbf{R}^i \mathbf{K}^{i,(m-1)} - (\mathbf{B}^i)^\top \mathbf{P}_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \Big(\mathbf{A} - \sum_{i \in \mathcal{N}_i^\kappa} \mathbf{B}^i \mathbf{K}^{i,(m-1)}\Big).$$

      Update the policies using the natural policy gradient updating rule:

$$\boldsymbol{K}_t^{i,(m)} = \boldsymbol{K}_t^{i,(m-1)} - 2\eta \mathbf{E}_{t,i}^{\boldsymbol{K}^{(m-1)}}. \tag{6}$$

    end for
  end for
end for
Return the iterates $\boldsymbol{K}^{(M)} = (\boldsymbol{K}^{1,(M)}, ..., \boldsymbol{K}^{N,(M)})$.

To prove Theorem 4, we need the supporting lemma shown below. We define $\mathbf{G}^i \in \mathbb{R}^{(T+1)n+Tk^i \times (T+1)n+Tk^i}$ and $\mathbf{F}^i \in \mathbb{R}^{(T+1)n \times (T+1)n+Tk^i}$ below. Also, we define $\mathbf{M}^i$ as the KKT matrix for the equality-constrained quadratic optimization program in (3).

$$\mathbf{G}^i = \begin{bmatrix} \mathbf{Q}^i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^i & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}^i \end{bmatrix}, \mathbf{F}^i = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ -\mathbf{A} & -\mathbf{B}^i & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{A} & -\mathbf{B}^i & \mathbf{I} \end{bmatrix}, \mathbf{M}^i := \begin{bmatrix} \mathbf{G}^i & (\mathbf{F}^i)^\top \\ \mathbf{F}^i & \mathbf{0} \end{bmatrix}$$

Lemma 5 (adapted from Shin et al. (2023)) *Consider a matrix $\boldsymbol{Z}$ induced by the graph $(G = (\mathcal{N}, \mathcal{E}), \mathcal{J} = \{\mathcal{J}_i\}_{i\in\mathcal{N}}, \mathcal{I} = \{\mathcal{I}_i\}_{i\in\mathcal{N}})$, according to Definition 2, with bandwidth not greater than 1. There exists $M, L, \alpha > 0$ such that $\|\boldsymbol{Z}\| \leq M, \boldsymbol{F}^i {\boldsymbol{F}^i}^\top \succeq \frac{(1-\alpha)^2}{L^2(1+L)^2}\boldsymbol{I}, \boldsymbol{N}^\top \boldsymbol{G}^i \boldsymbol{N} \succeq \frac{(1-\alpha)^2\gamma}{2L^4(1+L)^2}\boldsymbol{I}$, where $\boldsymbol{N}$ is the null space of $\boldsymbol{F}^i$. Further, $\phi \leq \sigma(\mathbf{Z}) \leq M$, and $\|\mathbf{Z}^{-1}(i,j)\| \leq \beta\psi^{0.5d(i,j)}$ for $i, j \in \mathcal{N}$, where M, ϕ, β, and ψ are defined in Theorem 4, and L and α are defined in Assumption 2.*

Proof for Theorem 4: Consider a time-varying graph $\mathcal{G}_t = (\mathcal{N}_t, \mathcal{E}_t)$. Note that, by construction, $\mathcal{N}_t$ and $\mathcal{E}_t$ define the nodes and coupling of the graph at the t-th time step that allow the correct partitioning for the respective time step of the time-dependent structures of $\mathbf{F}^i$ and $\mathbf{G}^i$. Thus, the KKT matrix $\mathbf{M}^i$ has a bandwidth not greater than 1 induced by $\mathcal{G}_T = (\mathcal{N}_T, \mathcal{E}_T), \mathcal{J}_T = \{\mathcal{J}_t^i\}_{t\in\mathbb{I}_{[0,T]}}^{i\in\mathcal{N}_T}, \mathcal{I}_T = \{\mathcal{I}_t^i\}_{t\in\mathbb{I}_{[0,T]}}^{i\in\mathcal{N}_T}$. Under Assumption 3, it suffices to show that $\|(\mathbf{M}^i)^{-1}(i,j)\| \leq \beta\psi^{0.5d(i,j)}$ to show that $\|\mathbf{K}^{i*}(i,j)\| \leq \beta\psi^{0.5d(i,j)}$. By Assumption 2 and Lemma 5, we have $\|(\mathbf{M}^i)^{-1}(i,j)\| \leq \beta\psi^{0.5d(i,j)}$ for $i, j \in \mathcal{N}_T$. Lastly, $\psi \in (0,1)$ and $\beta \geq 1$ follow directly from the definitions using the facts that $\phi < 1$ and $M > 1$. ∎

Lemma 6 *Consider the problem formulation in Section 2. Under Assumptions 1, 2, 3 and 4, Algorithm 1 converges to a neighborhood of the optimal solution with a linear convergence rate.*

Proof The proof of convergence follows from standard arguments as in Theorem 3.5 from Hambly et al. (2023) that introduced the original natural policy gradient algorithm for general sum N-player LQ games. ∎

Furthermore, Theorem 7 bounds the error between the true optimal controller gain ($\mathbf{K}^{i*}$) and the truncated optimal controller gain ($\mathbf{K}^{i\kappa}$). We define a sub-exponential function $f(d)$ that satisfies $|\{j \in \mathcal{N} : d(i,j) = d\}| \leq f(d)$.

Theorem 7 *Consider the problem formulation in Section 2, natural policy gradient algorithm in Hambly et al. (2023) for $\boldsymbol{K}^{i*}$, and Algorithm 1 for $\boldsymbol{K}^{i\kappa}$. Under Assumption 2 and 3, $\|\boldsymbol{K}^{i*} - \boldsymbol{K}^{i\kappa}\| \leq \Omega\Psi^\kappa$, where $\Omega = \left(\sup_{d\in\mathbb{I}_{\geq 0}} f(d)(\frac{\psi^{0.5}}{\Psi})^d\right)\beta\frac{\Psi}{1-\Psi}$, $\Psi = \frac{\psi^{0.5}+1}{2}$, and ψ is defined in Theorem 4.*

Proof for Theorem 7: We know that $\sum_{j\in\mathcal{N}\setminus\mathcal{N}_i^\kappa} \|\mathbf{K}(i,j)^{i*}\| \leq \sum_{d=\kappa+1}^{\infty} \beta f(d)$ and $(\frac{\psi^{0.5}}{\Psi})^d\Psi^d \leq (\sup_{d\in\mathbb{I}_{\geq 0}} f(d)(\frac{\psi^{0.5}}{\Psi})^d)\frac{\beta\Psi}{1-\Psi}\Psi^\kappa$ for any $i \in \mathcal{N}$. The first inequality follows from the definition of the sub-exponential function $f(d)$. The second inequality follows since the product of a sub-exponential and exponentially decaying functions is bounded. Further, the supremum is bounded since the product of a sub-exponential function and an exponentially decaying function converges. The effect of multiple nodes being more than κ hops away is exponentially small in κ due to the fact that the exponential decay is stronger than the sub-exponential increase. ∎

Finally, we state our main result which bounds the difference in the costs from using the complete state and action information and using the local state and action information.

Theorem 8 *Consider the problem formulation in Section 2, the natural policy gradient algorithm in Hambly et al. (2023) for J^{i*}, and Algorithm 1. Under Assumptions 2 and 3, and Theorem 7, $|J^{i*} - J^{i\kappa}| \leq \Delta \Psi^{\kappa} \|x_0\|^2$, for all i and κ,*
where $\Delta = L\Omega \left(\frac{1}{1 - \left(\frac{\psi}{2\beta^2}\right)} \right) \left(\frac{\beta^2}{1-\psi} \right) \cdot \left(\left(1 + \frac{L^4(1+L^2)}{1-\alpha^2} \right) \left(\frac{2L^5(1+L^2)}{\gamma(1-\alpha^2)} + \Omega \right) + \frac{2L^4(1+L^2)}{1-\alpha^2} \right)$.

Proof for Theorem 8: From Theorem 7, we know the following holds:

$$
\begin{aligned}
g(x) &:= (u^{i\kappa}(x))^\top \mathbf{R}^i (u^{i\kappa}(x)) - (u^{i*}(x))^\top \mathbf{R}^i (u^{i*}(x)) \\
&= (u^{i\kappa}(x) + u^{i*}(x))^\top \mathbf{R}^i (u^{i\kappa}(x) - u^{i*}(x)) \\
&\leq L\Omega \Big(2\beta + \Omega \Big) \Psi^{\kappa} \|x\|^2.
\end{aligned}
$$

This implies that:

$$
h(x) := J^{i*}((\mathbf{A} - \mathbf{B}^i \mathbf{K}^{i*})x) - J^{i\kappa}((\mathbf{A} - \mathbf{B}^i \mathbf{K}^{i\kappa})x) \leq \Omega\beta\gamma \Big(2 + 2\beta + \Omega \Big) \Psi^{\kappa} \|x\|^2.
$$

From the definition of $J^{i*}(\cdot)$ and $J^{i\kappa}(\cdot)$, we know that $J^{i\kappa}(x^\kappa(t)) - J^{i*}(x^*(t)) = J^{i\kappa}(x^\kappa(t+1)) - J^{i*}(x^*(t+1)) + g(x^\kappa(t)) + h(x^\kappa(t))$.
Summing up this equality from $t = 0$ to $t = T - 1$, we get:

$$
J^{i\kappa}(x_0) - J^{i*}(x_0) = J^{i\kappa}(x^\kappa(T)) - J^{i*}(x^\kappa(T)) + \sum_{t=0}^{T-1} g(x^\kappa(t)) + \sum_{t=0}^{T-1} h(x^\kappa(t))
$$

From Assumption 2, $J^{i\kappa}(x^\kappa(T)) \to 0$ and $J^{i*}(x^*(T)) \to 0$ as $T \to \infty$. Also, we have $\sum_{t=0}^{\infty} g(x^\kappa(t)) \leq \frac{\frac{\beta^2}{1-\psi} \cdot L\Omega \left(2\beta + \Omega \Psi^{\kappa} \right)}{1 - \left(\frac{\psi}{2\beta^2}\right)} \Psi^{\kappa} \|x_0\|^2$ and $\sum_{t=0}^{\infty} h(x^\kappa(t)) \leq \frac{\frac{\beta^2}{1-\psi} \Omega\beta\gamma(2 + 2\beta + \Omega)}{1 - \left(\frac{\psi}{2\beta^2}\right)} \Psi^{\kappa} \|x_0\|^2$.
Thus, taking $T \to \infty$, we get the required bound. ■

In this result, based on the exponential decay property of the controller gain matrix, we have shown that the norm of the error of the controller gain matrix and the error of the cost from utilizing local neighborhood information only are bounded. This means that Algorithm 1 converges to a bounded neighborhood of the true NE.

4. Numerical Example

We consider a finite horizon, time invariant LQ game. The problem formulation is outlined below, and we consider a linear communication graph with 11 agents. The graph consists of 11 agents. The complete state and action information corresponds to $\kappa = 10$. The time horizon, step size, and number of iterations are set to $T = 5$, $\eta = 0.003$, and $I = 3000$, respectively. The state transition matrices $\mathbf{A}$ and $\{\mathbf{B}^i\}_{i=0}^{10}$, the cost parameterization matrices $\{\mathbf{Q}^i\}_{i=0}^{10}$ and $\{\mathbf{R}^i\}_{i=0}^{10}$, and the system noise covariance matrix $\mathbf{W}$ are defined below.

$$\mathbf{A} = \begin{bmatrix} 0.08 & 0.02 & 0.01 & 0.08 & 0.09 & 0.06 & 0.09 & 0.07 & 0.04 & 0.02 & 0.09 \\ 0.06 & 0.00 & 0.03 & 0.02 & 0.04 & 0.06 & 0.01 & 0.10 & 0.08 & 0.03 & 0.07 \\ 0.09 & 0.07 & 0.04 & 0.03 & 0.04 & 0.09 & 0.09 & 0.08 & 0.05 & 0.08 & 0.03 \\ 0.01 & 0.01 & 0.07 & 0.08 & 0.04 & 0.05 & 0.02 & 0.06 & 0.07 & 0.05 & 0.03 \\ 0.04 & 0.00 & 0.03 & 0.08 & 0.09 & 0.07 & 0.09 & 0.01 & 0.07 & 0.05 & 0.00 \\ 0.02 & 0.07 & 0.09 & 0.00 & 0.05 & 0.04 & 0.03 & 0.06 & 0.01 & 0.03 & 0.05 \\ 0.06 & 0.04 & 0.05 & 0.06 & 0.06 & 0.10 & 0.08 & 0.02 & 0.07 & 0.03 & 0.07 \\ 0.01 & 0.02 & 0.09 & 0.01 & 0.04 & 0.04 & 0.02 & 0.07 & 0.04 & 0.07 & 0.02 \\ 0.10 & 0.09 & 0.09 & 0.04 & 0.02 & 0.02 & 0.08 & 0.02 & 0.07 & 0.00 & 0.08 \\ 0.06 & 0.04 & 0.04 & 0.07 & 0.02 & 0.00 & 0.01 & 0.06 & 0.08 & 0.02 & 0.03 \\ 0.06 & 0.00 & 0.07 & 0.05 & 0.09 & 0.01 & 0.00 & 0.10 & 0.03 & 0.05 & 0.10 \end{bmatrix};$$

$$\mathbf{B} = \begin{bmatrix} 0.03 & 0.09 & 0.03 & 0.06 & 0.05 & 0.06 & 0.07 & 0.04 & 0.10 & 0.00 & 0.02 \\ 0.03 & 0.09 & 0.08 & 0.06 & 0.09 & 0.04 & 0.01 & 0.01 & 0.08 & 0.00 & 0.04 \\ 0.07 & 0.02 & 0.06 & 0.08 & 0.02 & 0.02 & 0.02 & 0.09 & 0.07 & 0.05 & 0.01 \\ 0.05 & 0.08 & 0.00 & 0.03 & 0.02 & 0.08 & 0.06 & 0.06 & 0.03 & 0.10 & 0.07 \\ 0.02 & 0.10 & 0.08 & 0.04 & 0.03 & 0.07 & 0.07 & 0.03 & 0.03 & 0.05 & 0.06 \\ 0.10 & 0.02 & 0.07 & 0.08 & 0.09 & 0.04 & 0.03 & 0.09 & 0.04 & 0.09 & 0.05 \\ 0.07 & 0.01 & 0.05 & 0.05 & 0.07 & 0.05 & 0.08 & 0.00 & 0.07 & 0.07 & 0.05 \\ 0.01 & 0.02 & 0.01 & 0.08 & 0.05 & 0.05 & 0.01 & 0.05 & 0.07 & 0.05 & 0.06 \\ 0.09 & 0.07 & 0.04 & 0.01 & 0.06 & 0.08 & 0.05 & 0.02 & 0.04 & 0.08 & 0.04 \\ 0.07 & 0.05 & 0.04 & 0.04 & 0.10 & 0.05 & 0.01 & 0.02 & 0.10 & 0.06 & 0.07 \\ 0.01 & 0.00 & 0.00 & 0.09 & 0.02 & 0.04 & 0.01 & 0.05 & 0.03 & 0.05 & 0.09 \end{bmatrix};$$

$$\mathbf{W} = \begin{bmatrix} 0.1 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.01 \\ 0.01 & 0.2 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.02 & 0.01 & 0.1 & 0.02 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.02 & 0.2 & 0.02 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.01 \\ 0.02 & 0.01 & 0.01 & 0.02 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.2 & 0.01 & 0.02 & 0.01 & 0.02 & 0.02 \\ 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.1 & 0.01 & 0.02 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.2 & 0.01 & 0.01 & 0.01 \\ 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.1 & 0.02 & 0.01 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.02 & 0.2 & 0.02 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.01 & 0.02 & 0.01 \end{bmatrix};$$

$$\mathbf{Q}^i = 0.2 \cdot \mathbf{I}; \mathbf{R}^i = 0.5 \text{ for } i = 0, 1, ..., 10.$$

The initial states are sampled from a Gaussian distribution with the means and variances shown below, and the initial controller gains $\{\mathbf{K}^i\}_{i=0}^{10}$ are set as shown below.

$$x_0^0 = x_0^2 = x_0^3 = x_0^5 = x_0^6 = x_0^8 = x_0^9 = N(0.3, 0.2); x_0^1 = x_0^4 = x_0^7 = x_0^{10} = N(0.2, 0.3);$$
$$\mathbf{K}^0 = \mathbf{K}^3 = \mathbf{K}^6 = \mathbf{K}^9 = (0.35, 0.01, 0.1, 0.35, 0.01, 0.1, 0.35, 0.01, 0.1, 0.35, 0);$$
$$\mathbf{K}^1 = \mathbf{K}^4 = \mathbf{K}^7 = \mathbf{K}^{10} = (-0.3, -0.2, 0, -0.3, -0.2, 0, -0.3, -0.2, 0, -0.3, 0);$$
$$\mathbf{K}^2 = \mathbf{K}^5 = \mathbf{K}^8 = (-0.3, 0.1, 0, -0.3, 0.1, 0, -0.3, 0.1, 0, -0.3, 0).$$

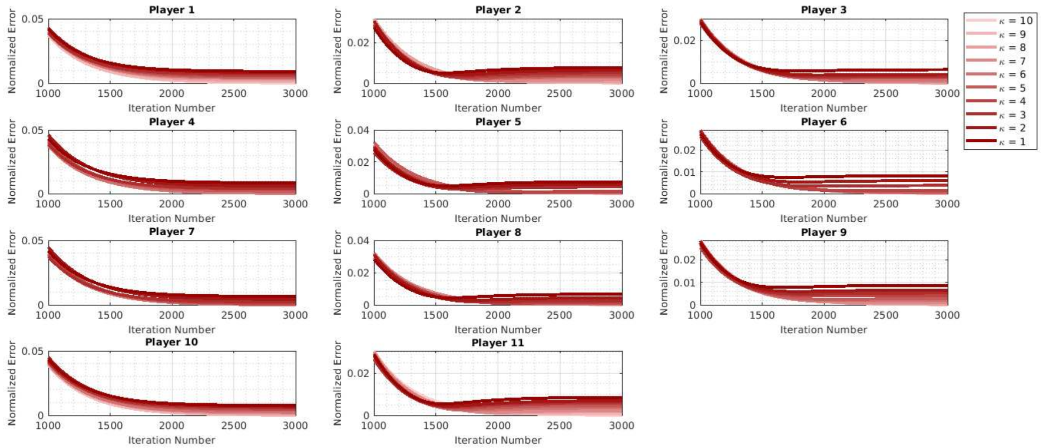

Figure 1: Convergence Results of Scalable Natural Policy Gradient for 11-Player LQ Game for Different Values of κ

Figure 1 shows the convergence results of the scalable natural policy gradient for the synthetic 11-player LQ game across different values of κ. The algorithm converges for all players with minimal error. As κ increases, the normalized error decreases, as indicated by the red gradient color code, due to the increased availability of state and action information, leading to more accurate policy approximations.

Figure 2 shows the cost error due to incomplete state and action information using scalable natural policy gradient for the synthetic 11-player LQ game across different values of κ. As κ increases, the cost error decreases for all players as expected. The figure also demonstrates the algorithm's feasibility in converging to a bounded ϵ-neighborhood of the NE, with a maximum cost error of 0.009. At $\kappa = 10$, all agents have complete information, resulting in zero cost error.

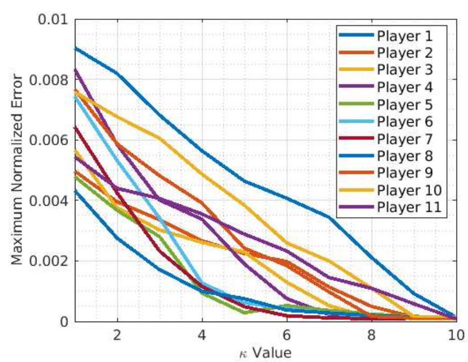

Figure 2: Cost Error for 11-Player LQ Game for Different Values of κ

5. Conclusion & Future Work

We considered policy gradient algorithm in a general-sum LQ game. The traditional implementation requires state and action information from all other agents which may not be scalable. Instead, we proposed and analyzed an algorithm that converges to an ϵ-neighborhood of the NE with local information. Future work could extend this method to the setting with unknown system parameters.

Acknowledgments

The work was partially supported by NSF and ARO.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 354–363. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balduzzi18a.html.

T. Basar and G.J. Olsder. *Dynamic Noncooperative Game Theory: Second Edition*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1999. ISBN 9781611971132. URL https://books.google.com/books?id=nry8U3CfF-gC.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. doi: 10.1109/TSMCC.2007.913919.

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11), 2021. ISSN 2076-3417. doi: 10.3390/app11114948. URL https://www.mdpi.com/2076-3417/11/11/4948.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *CoRR*, abs/2101.04233, 2021.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods find the nash equilibrium in n-player general-sum linear-quadratic games. *Journal of Machine Learning Research*, 24(139): 1–56, 2023. URL http://jmlr.org/papers/v24/21-0842.html.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *CoRR*, abs/1603.01121, 2016.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey.

IEEE Communications Surveys & Tutorials, 24(2):1240–1279, 2022. doi: 10.1109/COMST.2022.3160697.

Eric Mazumdar, Lillian J. Ratliff, Michael I. Jordan, and S. Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 860–868, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, and et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. doi: 10.1038/nature14236.

Andrea Monti, Benita Nortmann, Thulasi Mylvaganam, and Mario Sassano. Feedback and open-loop nash equilibria for lq infinite-horizon discrete-time dynamic games. *SIAM Journal on Control and Optimization*, 62(3):1417–1436, 2024. doi: 10.1137/23M1579960. URL https://doi.org/10.1137/23M1579960.

Benita Nortmann and Thulasi Mylvaganam. Approximate nash equilibria for discrete-time linear quadratic dynamic games. *IFAC-PapersOnLine*, 56(2):1760–1765, 2023. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2023.10.1886. URL https://www.sciencedirect.com/science/article/pii/S2405896323022954. 22nd IFAC World Congress.

Benita Nortmann, Andrea Monti, Mario Sassano, and Thulasi Mylvaganam. Nash equilibria for linear quadratic discrete-time dynamic games via iterative and data-driven algorithms. *IEEE Transactions on Automatic Control*, 69(10):6561–6575, 2024. doi: 10.1109/TAC.2024.3375249.

Johan Olsson, Runyu Cathy Zhang, Emma Tegling, and Na Li. Scalable reinforcement learning for linear-quadratic control of networks. In *2024 American Control Conference (ACC)*, pages 1813–1818, 2024. doi: 10.23919/ACC60939.2024.10644413.

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. Multiplicative weights updates as a distributed constrained optimization algorithm: Convergence to second-order stationary points almost always. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4961–4969. PMLR, 09–15 Jun 2019.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2074–2086. Curran Associates, Inc., 2020a.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 256–266. PMLR, 10–11 Jun 2020b.

Masoud Roudneshin, Jalal Arabneydi, and Amir G. Aghdam. Reinforcement learning in nonzero-sum linear quadratic deep structured games: Global convergence of policy optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 512–517, 2020. doi: 10.1109/CDC42340.2020.9303950.

Mario Sassano, Thulasi Mylvaganam, and Alessandro Astolfi. Ol-ne for lq differential games: A port-controlled hamiltonian system perspective and some computational strategies. *Automatica*, 171:111953, 2025. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2024.111953. URL https://www.sciencedirect.com/science/article/pii/S0005109824004473.

M. L. Scarpa and T. Mylvaganam. Open-loop and feedback lq potential differential games for multi-agent systems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 6283–6288, 2023. doi: 10.1109/CDC49753.2023.10384220.

M. L. Scarpa, B. Nortmann, M. Sassano, and T. Mylvaganam. Feedback nash equilibrium solutions of two-player lq differential games: Synthesis and analysis via a state/costate interpretation. *IEEE Control Systems Letters*, 8:1451–1456, 2024. doi: 10.1109/LCSYS.2024.3410630.

Mostafa M. Shibl and Vijay Gupta. A scalable game theoretic approach for coordination of multiple dynamic systems. *IEEE Control Systems Letters*, 2024. doi: 10.1109/LCSYS.2024.3501155.

Sungho Shin, Mihai Anitescu, and Victor M. Zavala. Exponential decay of sensitivity in graph-structured nonlinear programs. *SIAM Journal on Optimization*, 32(2):1156–1183, 2022. doi: 10.1137/21M1391079.

Sungho Shin, Yiheng Lin, Guannan Qu, Adam Wierman, and Mihai Anitescu. Near-optimal distributed linear-quadratic regulator for networked systems. *SIAM Journal on Control and Optimization*, 61(3):1113–1135, 2023. doi: 10.1137/22M1489836.

Xinliang Song, Tonghan Wang, and Chongjie Zhang. Convergence of multi-agent learning with a finite step size in general-sum games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 935–943, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10772–10782. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21j.html.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective, 2021. URL https://arxiv.org/abs/2011.00583.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.,

2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5446f217e9504bc593ad9dcf2ec88dda-Paper.pdf.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, pages 321–384. Springer International Publishing, Cham, 2021. doi: 10.1007/978-3-030-60990-0_12. URL https://doi.org/10.1007/978-3-030-60990-0_12.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication, 2024. URL https://arxiv.org/abs/2203.08975.