Exploiting inter-agent coupling information for efficient reinforcement learning of cooperative LQR

Shahbaz P Qadri Syed SHAHBAZ_QADRI.SYED@OKSTATE.EDU
He Bai HE.BAI@OKSTATE.EDU
Mechanical and Aerospace Engineering, Oklahoma State University, USA.

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

Abstract

Developing scalable and efficient reinforcement learning algorithms for cooperative multi-agent control has received significant attention over the past years. Existing literature has proposed *inexact* decompositions of local Q-functions based on empirical information structures between the agents. In this paper, we exploit inter-agent coupling information and propose a systematic approach to *exactly* decompose the local Q-function of each agent. We develop an approximate least square policy iteration algorithm based on the proposed decomposition and identify two architectures to learn the local Q-function for each agent. We establish that the worst-case sample complexity of the decomposition is equal to the centralized case and derive necessary and sufficient graphical conditions on the inter-agent couplings to achieve better sample efficiency. We demonstrate the improved sample efficiency and computational efficiency on numerical examples.

Keywords: Cooperative Linear quadratic regulator, Multi-agent least square policy iteration, Multi-agent learning for control.

1. Introduction

Owing to the recent advancements in multi-agent reinforcement learning (MARL), there is an increasing interest to investigate MARL as a solution concept for model-free optimal control of uncertain cooperative multi-agent systems (MAS). However, scaling from single agent RL to MARL poses two main challenges: first, a centralized approach for learning is limited by the curse of dimensionality because of the combinatorial joint state-control space of the agents. Several recent works have attempted to address the scalability issue by approximation or relaxed assumptions during learning (e.g., Lowe et al. (2017); Foerster et al. (2017); Zhang et al. (2018); Guestrin et al. (2001); Sunehag et al. (2017); Rashid et al. (2018); Qu et al. (2020); Jing et al. (2024)). Second, each agent typically interacts with only a subset of agents resulting in a partial information structure and *non-stationarity* during learning. In addition, the problem of finding globally optimal controllers with information constraints is known to be NP-hard even in a model-based setting (Witsenhausen (1968); Blondel and Tsitsiklis (2000); Papadimitriou and Tsitsiklis (1986); Ye et al. (2022)). Due to these challenges, the effectiveness of MARL for optimal control of uncertain MAS is not yet fully explored even in the linear quadratic regulator (LQR) setting, where the agent dynamics are linear time-invariant and the cost is quadratic.

In this paper, we study a cooperative LQR problem and propose a systematic approach to decompose the Q-function of each agent and address the issues of information constraint and scalability. We identify *exact* decompositions of the individual Q-function and the gradient of the global Q-function with respect to each agent's control parameters, given the knowledge of state-, cost-,

and observation couplings between the agents. Leveraging inter-agent couplings to develop efficient RL algorithms for MAS is of great interest recently. For instance, Qu et al. (2020) develop a Q-function approximation strategy based on the spatially exponentially decaying (SED) property (Bamieh et al. (2002); Motee and Jadbabaie (2008); Gamarnik et al. (2014); Qu et al. (2020)) which exploits the structure of local interactions in state transitions assuming decentralized observations and cost. Alemzadeh and Mesbahi (2019) consider a distributed Q-learning approach for dynamically decoupled agents with coupled costs. For an LQR problem, Alemzadeh and Mesbahi (2019) show that the controller of each subsystem asymptotically converges to the optimal controller. Other distributed RL methods (e.g., Kar et al. (2013); Zhang et al. (2018); Macua et al. (2018); Zhang and Zavlanos (2023); Zhang et al. (2021); Li et al. (2023); Qu et al. (2019)) have been proposed to employ a consensus algorithm to estimate the global cost with local information from neighbors. The performance of such algorithms depends on convergence of the consensus and the estimation quality of the global cost. Another relevant class of methods are value function factorizatio based methods (Koller and Parr (1999); Guestrin et al. (2001); Sunehag et al. (2017); Rashid et al. (2018); Son et al. (2019)) and coordination graph (CG) based methods (Guestrin et al. (2001, 2002); Kok and Vlassis (2006)), which aim to *approximately* factorize the global Q-function using the individual Q-functions of all the agents. However, the existing methods typically consider one or two types of inter-agent couplings and perform approximate decompositions of the Q-function. In contrast, in this paper we consider that local interactions between the agents occur in the form of state, observation and cost couplings and present an approach to deduce an exact decomposition of the individual Q-function.

The formulation presented in this paper is closely related to Jing et al. (2024) that also incorporates the three types of inter-agent couplings and proposes a distributed RL framework based on zeroth order optimization. However, there are a few key differences. First, Jing et al. (2024) consider a value function based approach that depends on the global state and action whereas we identify an exact decomposition of the Q-function which is more closely related to the actor-critic methods (Konda and Tsitsiklis (1999)) in RL. Second, Jing et al. (2024) introduce a new graph based on the inter-agent couplings called *learning graph* which represents the information flow during learning. This learning graph is equivalent to the gradient dependency graph introduced in this paper. However, we derive this graph independently based on the decomposition of the Q-function.

LQR is a popular benchmark in RL to analyze the performance and limitations for various algorithms (Recht (2019)). Although the model-free LQR setting (for both structured and unstructured) is well studied in single agent RL (e.g., Bradtke and Barto (1996); Lagoudakis and Parr (2003); Krauth et al. (2019); Park et al. (2020)), it is less explored in the multi-agent setting. For some recent results in the networked LQR setting, see Li et al. (2021); Jing et al. (2021b,a); Alemzadeh and Mesbahi (2019); Zhang et al. (2023); Olsson et al. (2024). Another contribution of this paper is that we analyze the sample complexity and the estimation error in Q-function parameters for the cooperative LQR problem using the least square policy iteration (LSPI) framework proposed in Lagoudakis and Parr (2003). In particular, we propose a *multi-agent LSPI* (MALSPI) algorithm based on the proposed Q-function decomposition and identify two architectures (*direct* and *indirect*) to learn the local Q-function for each agent. We establish that the worst case sample complexity of the direct case is equal to the centralized case, and the worst case sample complexity of the indirect case is equal to the direct case. We also derive the necessary and sufficient graphical conditions on the inter-agent couplings for improved sample complexity of the decomposition. Finally, we validate the sample complexity results using a numerical example.

The rest of the paper is outlined as follows. Section 2 presents the formulation of the cooperative LQR problem. Section 3 introduces the decomposition of the local Q-function and the decomposition of the gradient of the global Q-function. Section 4 describes the MALSPI algorithm, and Section 5 states the main sample complexity results of the algorithm. Simulation results are presented in Section 6. Section 7 concludes the paper.

Notation: We denote a Gaussian distribution with mean μ and covariance Σ as $\mathcal{N}(\mu, \Sigma)$. Let $\mathbb{I}_n$ and $\mathbf{0}_n$ denote an identity matrix and a zero vector of size n respectively. For an n-dimensional symmetric matrix M, let $\text{svec}(M) \in \mathbb{R}^{\frac{n(n+1)}{2}}$ be the vector of the upper triangular entries of M such that $||M||_F^2 = \langle \text{svec}(M), \text{svec}(M) \rangle$ and $\text{smat}(\cdot)$ represent the inverse operation of $\text{svec}(\cdot)$. Let $\mathcal{L}(X, Y)$ be the analytical solution of the discrete-time Lyapunov equation $\mathcal{P} = X\mathcal{P}X^\intercal + Y$.

2. Formulation of the cooperative LQR problem

Consider a multi-agent system comprising N agents with linear time-invariant dynamics and quadratic costs. Let $\mathcal{V} = \{1, \cdots, N\}$. We assume a generic setting, where each agent's dynamics and cost may depend on other agents' states and controls and its observations are the states of a subset of all the agents. Specifically, the dynamics of agent $i \in \mathcal{V}$ is given by

$$x_i(t+1) = \sum_{j \in \mathcal{I}_S^i} A_{ij} x_j(t) + \sum_{j \in \mathcal{I}_S^i} B_{ij} u_j(t) + w_i(t), \tag{1}$$

where $x_i(t) \in \mathbb{R}^{n_x}$, $u_i(t) \in \mathbb{R}^{n_u}$, $w_i(t) \sim \mathcal{N}(0, \sigma_w^2 \, \mathbb{I}_{n_x})$ is the process noise, and the set $\mathcal{I}_S^i$ contains the indices of the agents impacting the dynamics of agent i. For the ease of exposition, we have assumed that the dimensions of $x_i(t)$ and $u_i(t)$ are homogeneous across all the agents. However, the results in this paper can be easily extended to heterogeneous state and control dimensions. Let $x_{\mathcal{I}_S^i}(t) = [\cdots \; x_j^\intercal(t) \; \cdots]^\intercal$ and $u_{\mathcal{I}_S^i}(t) = [\cdots \; u_j^\intercal(t) \; \cdots]^\intercal, \forall \, j \in \mathcal{I}_S^i$.

The quadratic cost incurred by agent $i \in \mathcal{V}$ at time t is given by

$$c_i(x_{\mathcal{I}_C^i}(t), u_{\mathcal{I}_C^i}(t)) = (x_{\mathcal{I}_C^i}(t))^\intercal S_i (x_{\mathcal{I}_C^i}(t)) + (u_{\mathcal{I}_C^i}(t))^\intercal R_i (u_{\mathcal{I}_C^i}(t)), \tag{2}$$

where the set $\mathcal{I}_C^i$ contains the indices of the agents impacting the cost of agent i, $x_{\mathcal{I}_C^i}(t) = [\cdots \; x_j^\intercal(t) \; \cdots]^\intercal$, and $u_{\mathcal{I}_C^i}(t) = [\cdots \; u_j^\intercal(t) \; \cdots]^\intercal, \; \forall \; j \in \mathcal{I}_C^i$. We assume that S_i is a positive semi-definite matrix and R_i is a positive definite matrix.

Agent i observes the states of the agents in a set $\mathcal{I}_O^i$ and prescribes a structured linear control

$$\pi_i(x_{\mathcal{I}_O^i}(t)) := u_i(t) = K_i x_{\mathcal{I}_O^i}(t), \tag{3}$$

where $K_i = [\cdots \; K_{ij} \; \cdots]$, $x_{\mathcal{I}_O^i}(t) = [\cdots \; x_j^\intercal(t) \; \cdots]^\intercal, \forall \, j \in \mathcal{I}_O^i$.

The agent couplings in state dynamics, costs, and observations are captured by the index sets $\mathcal{I}_S^i$, $\mathcal{I}_C^i$, and $\mathcal{I}_O^i$, $\forall i$. We assume that these sets are time-invariant and further define a *state graph* $\mathcal{G}_S = \{\mathcal{V}, \mathcal{E}_S\}$, an *observation graph* $\mathcal{G}_O = \{\mathcal{V}, \mathcal{E}_O\}$, and a *cost graph* $\mathcal{G}_C = \{\mathcal{V}, \mathcal{E}_C\}$. An edge $(i, j) \in \mathcal{E}_S$, if $i \in \mathcal{I}_S^j$. Similarly, an edge $(i, j) \in \mathcal{E}_O$ if $i \in \mathcal{I}_O^j$ and an edge $(i, j) \in \mathcal{E}_C$ if $i \in \mathcal{I}_C^j$. We now present the combined state dynamics, cost, and controller. Let $x(t) = [x_1^\intercal(t) \; \cdots \; x_N^\intercal(t)]^\intercal \in \mathbb{R}^{Nn_x}$, $u(t) = [u_1^\intercal(t) \; \cdots \; u_N^\intercal(t)]^\intercal \in \mathbb{R}^{Nn_u}$, and $w(t) = [w_1^\intercal(t) \; \cdots \; w_N^\intercal(t)]^\intercal \in \mathbb{R}^{Nn_x}$. The global state of the multi-agent system, $x(t)$, evolves according to

$$x(t+1) = Ax(t) + Bu(t) + w(t), \tag{4}$$

where $A \in \mathbb{R}^{Nn_x \times Nn_x}$ and $B \in \mathbb{R}^{Nn_x \times Nn_u}$ are *unknown* system matrices depending on A_{ij} and B_{ij} in (1), respectively. Observe that $\forall\, i \in \mathcal{V},\ \forall\, j \in \mathcal{V} \setminus \mathcal{I}_S^i$, $A_{ij} = \mathbf{0}_{n_x \times n_x}$, and $B_{ij} = \mathbf{0}_{n_x \times n_u}$.

In our cooperative LQR problem, the global cost $C(x(t), u(t))$ is given by the average of the individual cost functions, i.e., $C(x(t), u(t)) = \frac{1}{N} \sum_{i=1}^N c_i(x_{\mathcal{I}_C^i}(t), u_{\mathcal{I}_C^i}(t))$, which is rewritten as

$$C(x(t), u(t)) = x^\intercal(t) S x(t) + u^\intercal(t) R u(t), \quad \forall\, t, \tag{5}$$

where S and R are cost matrices depending on S_i and R_i in (2). Note that $\forall\, i \in \mathcal{V},\ \forall\, j \in \mathcal{V} \setminus \mathcal{I}_C^i$, $S_{ij} = \mathbf{0}_{n_x \times n_x}$, and $R_{ij} = \mathbf{0}_{n_u \times n_u}$. The prescribed structured static linear controllers (3) represented in a compact form is given by $\pi(x(t)) := u(t) = Kx(t)$, where the control gain K depends on $\mathcal{G}_O$ such that $\forall\, i \in \mathcal{V},\ \forall\, j \in \mathcal{V} \setminus \mathcal{I}_O^i$, $K_{ij} = \mathbf{0}_{n_u \times n_x}$.

We seek to optimize the control $u(t)$ to minimize the expected long term cost. The *global* Q-function for the cooperative LQR problem is defined as $Q(x, u) = \mathbb{E}[\sum_{t=0}^\infty C(x(t), u(t)) | x(0) = x, u(0) = u]$, where the expectation is taken over the state and control distributions, respectively. For agent $i \in \mathcal{V}$, define a *local* Q function as $Q_i(x, u) = \mathbb{E}\left[\sum_{t=0}^\infty c_i(x_{\mathcal{I}_C^i}(t), u_{\mathcal{I}_C^i}(t)) | x(0) = x, u(0) = u\right]$, which satisfies $Q(x, u) = \sum_{i=1}^N Q_i(x, u)$.

The objective of an infinite horizon average cost cooperative LQR problem is to compute an optimal controller $u^*(t) = K_* x(t)$, that minimizes

$$J(x(0), u(0)) = \min_K \lim_{T \to \infty} \mathbb{E}_{x(0), u(0)} \left[Q(x(0), u(0))\right] = \min_K \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^T x^\intercal(t) S x(t) + u^\intercal(t) R u(t)\right],$$

when the system starts from a global state $x(0)$, executes a global control $u(0)$ and follows a policy $\pi(\cdot)$ thereafter. Given a global state $x(t)$, the global controller $u(t)$ and the collection of local controllers $\{u_i(t)\}_{i=1}^N$can be transformed into each other. Therefore, our cooperative LQR problem can be concisely expressed as

$$\begin{aligned} \min_K \quad & \mathbb{E}_{x(0), u(0)} \ [Q(x(0), u(0)) | \mathcal{G}_S, \mathcal{G}_O, \mathcal{G}_C] \\ \text{subject to} \quad & x_i(t+1) \sim \mathcal{N}\left(\textstyle\sum_{j \in \mathcal{I}_S^i} \left[A_{ij} x_j(t) + B_{ij} u_j(t)\right], \sigma_w^2\, \mathbb{I}_{n_x}\right), \\ & u_i(t) = K_i x_{\mathcal{I}_O^i}(t),\ \forall\, i \in \mathcal{V}. \end{aligned} \tag{6}$$

For general linear time-invariant systems, there is no known tractable algorithm for computing optimal K_i (Rotkowitz and Lall (2005)). Moreover, Blondel and Tsitsiklis (2000) showed that the problem of finding stabilizing decentralized output feedback is NP-complete. Tractable algorithms that guarantee a global optimal controller have been developed for specific information structures, such as partially nested information structures (Ho and Chu (1972)), quadratic invariance (Rotkowitz and Lall (2005)), partially ordered sets (Shah and Parrilo (2013)), and decoupled control cost (Kashyap and Lessard (2023)). In this work, we propose an algorithm that aims at solving the general non-convex optimization problem (6) (with a static output feedback) using a policy iteration algorithm. Our main goal is to demonstrate improved sample complexity in estimating Q functions by recognizing structures in agent couplings. Convergence properties of the proposed algorithm will be investigated in the future work.

The system matrices A, B, and the cost matrices S, R are unknown in the model-free setting. However, their sparsity patterns may be known based on the inter-agent couplings (e.g., through

physical and network couplings). Thus, we assume that the agent coupling graphs $\mathcal{G}_S$, $\mathcal{G}_O$, and $\mathcal{G}_C$ are available and develop a systematic procedure to exploit the interplay between the system sparsity and the coupling graphs to improve the sample complexity of the cooperative LQR problem. Specifically, we decompose the Q_i function given $\mathcal{G}_S$, $\mathcal{G}_O$, and $\mathcal{G}_C$ (Section 3), from which a Q-learning algorithm is proposed and analyzed (Section 4 and 5).

3. Decomposition of the Q-function

Due to the cooperative average quadratic cost, solving (6) is equivalent to minimizing the individual expected $Q_i(\cdot)$, $\forall\ i\ \in\ \mathcal{V}$ which is typically assumed to be dependent on the global state and global control in the literature (see e.g., Jing et al. (2024), Lowe et al. (2017), Zhang et al. (2018)). Such a dependency incurs a combinatorial state-control dimension that grows exponentially with the number of agents leading to the curse of dimensionality. Since $Q_i(x(t),u(t)) = c_i(x_{\mathcal{I}_C^i}(t),u_{\mathcal{I}_C^i}(t)) + \mathbb{E}\left[Q_i(x(t+1),u(t+1))\right]$, $Q_i(\cdot)$ depends only on $\mathcal{I}_C^i$ and the subset of agents required to propagate the states and controls of the agents in $\mathcal{I}_C^i$ through time. That is, $Q_i(\cdot)$ depends on a closed subset (under state transition and control) of agents required to compute $\mathcal{I}_C^i$ at each time instant. We formalize this notion in Lemma 3.1. Let $\mathcal{G}_{SO} = \mathcal{G}_S \cup \mathcal{G}_O$ and define $\mathcal{R}_{SO}^i = \{j \in \mathcal{V} | j \xrightarrow{\mathcal{E}_{SO}} i\} \cup \{i\}$ as the reachability set of i in the $\mathcal{G}_{SO}$. Lemma 3.1[1] shows that $\forall i \in \mathcal{V}$, the set $\mathcal{I}_Q^i \triangleq \{j \in \mathcal{R}_{SO}^k | \forall\ k \in \mathcal{I}_C^i\}$, is closed under state transition and control generation.

Lemma 3.1 *For any $i,j,k \in \mathcal{V}$, if $j \in \mathcal{I}_Q^i$, then for any $k \in \mathcal{R}_{SO}^j$, $k \in \mathcal{I}_Q^i$.*

Theorem 3.1 below establishes that $Q_i(\cdot)$ depends *only* on the states, and controls of the agents in $\mathcal{I}_Q^i$. Thus, we refer to $\mathcal{I}_Q^i$ as the *value dependence set* of agent i and the graph such that the in-neighbors of node i correspond to $\mathcal{I}_Q^i$ as the *value dependency graph* denoted by $\mathcal{G}_Q = \{\mathcal{V}, \mathcal{E}_Q\}$.

Theorem 3.1 (Value decomposition theorem) *Let $\mathcal{R}_{SO}^i = \{j \in \mathcal{V} | j \xrightarrow{\mathcal{E}_{SO}} i\} \cup \{i\}$. Then, $\forall\ i \in \mathcal{V}$, $Q_i(x(t),u(t)) = Q_i(x_{\mathcal{I}_Q^i}(t), u_{\mathcal{I}_Q^i}(t))$, where $\mathcal{I}_Q^i = \{j \in \mathcal{R}_{SO}^k | \forall\ k \in \mathcal{I}_C^i\}$.*

Solving the optimization in (6) requires each agent $i \in \mathcal{V}$ to compute the gradient $\nabla_{K_i} J(\cdot)$ which is again a global computation. Theorem 3.1 implies that K_i affects Q_j only if $i \in \mathcal{I}_Q^j$, from which we pursue a decomposition of the gradient of the global objective w.r.t. (with respect to) K_i. Define the *gradient dependency* graph $\mathcal{G}_{\text{GD}} = \{\mathcal{V}, \mathcal{E}_Q^\intercal\}$ and the corresponding index set $\mathcal{I}_{\text{GD}}^i = \{j \in \mathcal{V} | (j,i) \in \mathcal{E}_Q^\intercal\}$. Theorem 3.2 below shows that the gradient of $Q(\cdot)$ w.r.t. K_i can be decomposed as the sum of the gradients of Q_j w.r.t. K_i, $\forall\ j \in \mathcal{I}_{\text{GD}}^i$.

Theorem 3.2 (Gradient decomposition theorem) *For the cooperative LQR problem defined in Section 2, we have $\forall\ t \geq 0$, $\forall\ i \in \mathcal{V}$,*

$$\nabla_{K_i} Q(x(t),u(t)) = \nabla_{K_i} \left(\sum_{j \in \mathcal{I}_{GD}^i} Q_j(x_{\mathcal{I}_Q^j}(t), u_{\mathcal{I}_Q^j}(t)) \right).$$

Owing to Theorem 3.2, we further define $\mathcal{I}_{\widehat{Q}}^i \triangleq \{k \in \mathcal{V} | k \in \bigcup_{j \in \mathcal{I}_{\text{GD}}^i} \mathcal{I}_Q^j\}$ and $\widehat{Q}_i(x_{\mathcal{I}_{\widehat{Q}}^i}, u_{\mathcal{I}_{\widehat{Q}}^i}) \triangleq \sum_{j \in \mathcal{I}_{\text{GD}}^i} Q_j(x_{\mathcal{I}_Q^j}, u_{\mathcal{I}_Q^j})$. The multi-agent deterministic policy gradient theorem Lowe et al. (2017)

1. The proofs of Lemma 3.1 and the other theoretical results can be found in the appendix available at https://coral-osu.github.io/assets/pdf/MALQR_supplemental.pdf.

is an extension of the deterministic policy gradient theorem Silver et al. (2014) to an MARL setting that gives the gradient of the objective $J(\cdot)$ w.r.t. the policy parameters of agent i as $\nabla_{K_i} J(x,u) = \mathbb{E}\left[\nabla_{u_i} Q(x,u) \nabla_{K_i} u_i\right]$. Applying Theorem 3.1 and 3.2 yields our cooperative deterministic policy gradient: $\forall\, i \in \mathcal{V}$,

$$\nabla_{K_i} J(x,u) = \mathbb{E}\left[\nabla_{K_i} u_i \cdot \nabla_{u_i} \widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})\right] = \mathbb{E}\left[\nabla_{K_i} u_i \cdot \nabla_{u_i} \sum_{j \in \mathcal{I}^i_{\text{GD}}} Q_j(x_{\mathcal{I}^j_Q}, u_{\mathcal{I}^j_Q})\right]. \quad (7)$$

In the remainder of the paper, we focus on two architectures for computing the policy gradient: 1) the *direct* case (using the first equality in (7)), where agent i, $\forall i \in \mathcal{V}$, directly estimates $\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})$ as a quadratic function of $x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}}$; 2) the *indirect* case (using the second equality in (7)), where agent i, $\forall i \in \mathcal{V}$, estimates Q_i as a quadratic function of $x_{\mathcal{I}^i_Q}, u_{\mathcal{I}^i_Q}$ and communicates with the agents in $\mathcal{G}_{\text{GD}}$ to compute $\sum_{j \in \mathcal{I}^i_{\text{GD}}} Q_j(x_{\mathcal{I}^j_Q}, u_{\mathcal{I}^j_Q})$.

4. Multi-agent structured least square policy iteration

In this section, we present the *multi-agent least square policy iteration (MALSPI)* to solve (6) in a model-free setting. This algorithm extends the least square policy iteration (Lagoudakis and Parr (2003))–which is well understood both theoretically and empirically in a single agent setting–to MAS utilizing the decomposition in Theorem 3.1 and 3.2. The proposed MALSPI algorithm is an off-policy algorithm that employs a shared experience buffer. In each iteration, a trajectory rollout of T samples is collected using a stabilizing policy $K_{\text{play}} (\neq K$, in general$)$. We assume that K_{play} is either known (e.g., for open-loop stable dynamics) or learned in a model-free setting (e.g., Jing et al. (2021b)). Then each agent performs consecutive policy evaluation and policy improvement steps in parallel. The proposed MALSPI algorithm for the *direct* case is summarized in Algorithm 1.

Algorithm 1 Multi-agent Least Square Policy Iteration (MALSPI) - the direct case

Input: Initial stabilizing controller K_0, number of policy iterations n, length of trajectory rollout T, exploration noise variance σ_η^2, lower eigenvalue bound ζ , learning rate parameter α, direct VD set $\mathcal{I}^i_{\widehat{Q}}, \forall\, i \in \mathcal{V}$, global initial state mean x_0, and covariance Σ_0.

for $l = 0, \cdots, n$ **do**

  Starting from a global state $x^{(l)}(0) \sim \mathcal{N}(x_0, \Sigma_0)$, and using an arbitrary policy $u^{(l)}(t) = K_0 x^{(l)}(t) + \eta_t^{(l)}$, $\eta^{(l)}(t) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbb{I}_{N n_u})$, collect sample global trajectory $\mathcal{D}^{(l)} = \{x^{(l)}(t), u^{(l)}(t), x^{(l)}(t+1)\}_{t=1}^T$.

  for $i = 1, 2, \cdots, N$ *(in parallel)* **do**

    Query $\mathcal{D}^{(l)}_{\mathcal{I}^i_{\widehat{Q}}} = \{x^{(l)}_{\mathcal{I}^i_{\widehat{Q}}}(t), u^{(l)}_{\mathcal{I}^i_{\widehat{Q}}}(t), x^{(l)}_{\mathcal{I}^i_{\widehat{Q}}}(t+1)\}_{t=0}^T$ from $\mathcal{D}^{(l)}$

    $\hat{q}_i \leftarrow \text{LSTDQ}\left(\mathcal{D}^{(l)}_{\mathcal{I}^i_{\widehat{Q}}}, K^{(l)}_{\mathcal{I}^i_{\widehat{Q}}}\right)$ [(12)]; $\widehat{Q}_i = \text{psd\_proj}_\zeta\left(\text{smat}(\hat{q}_i)\right)$

    $K_i^{(l+1)} \leftarrow K_i^{(l)} - 2\alpha \mathbb{E}\left[\mathbb{J}_i \widehat{Q}_i \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}} \\ u_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} x^{\mathsf{T}}_{\mathcal{I}^i_O}\right]$ [(13)]

  end

end

Policy evaluation. We discuss the policy evaluation step in the direct case. The analysis extends to the indirect case in a straightforward manner. We assume that each agent $i \in \mathcal{V}$ has access to the evaluation policies of the agents in its $\mathcal{I}^i_{\widehat{Q}}$. Given $\bigcup_{j\in\mathcal{I}^i_{\widehat{Q}}} K_j$, agent $i \in \mathcal{V}$ estimates its corresponding Q-function using least squares temporal difference learning for Q-functions (LSTDQ) (Lagoudakis and Parr (2003)). To simplify the notation, define a projection operator $P^n_{\mathcal{S}_1,\mathcal{S}_2} \in \mathbb{R}^{n|\mathcal{S}_1|\times n|\mathcal{S}_2|}$ w.r.t. ordered subsets $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{V}$ such that $\forall\ i \in \mathcal{S}_1,\ j \in \mathcal{S}_2$, $P_{ij} \in \mathbb{R}^{n\times n}$ satisfies $P_{ij} = \begin{cases} \mathbb{I}_n, & \text{if } j \in \mathcal{S}_1 \\ \mathbf{0}_{n\times n}, & \text{otherwise} \end{cases}$. Since $\forall\ j \in \mathcal{I}^i_{\widehat{Q}}$, $\mathcal{I}^j_O \subseteq \mathcal{I}^i_{\widehat{Q}}$ and $\mathcal{I}^j_S \subseteq \mathcal{I}^i_{\widehat{Q}}$ by Lemma 3.1, let $K_{\mathcal{I}^i_{\widehat{Q}}} = [\cdots\ (K_j P^{n_x}_{\mathcal{I}^j_O,\mathcal{I}^i_{\widehat{Q}}})^\intercal\ \cdots]^\intercal$, $A_{\mathcal{I}^i_{\widehat{Q}}} = [\cdots\ (A_j P^{n_x}_{\mathcal{I}^j_S,\mathcal{I}^i_{\widehat{Q}}})^\intercal\ \cdots]^\intercal$, $B_{\mathcal{I}^i_{\widehat{Q}}} = [\cdots\ (B_j P^{n_u}_{\mathcal{I}^j_S,\mathcal{I}^i_{\widehat{Q}}})^\intercal\ \cdots]^\intercal$. It then follows that $\forall\ t \geq 0, \forall\ i \in \mathcal{V}, \forall j \in \mathcal{I}^i_{\widehat{Q}}$,

$$x_{\mathcal{I}^i_{\widehat{Q}}}(t+1) = A_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t) + B_{\mathcal{I}^i_{\widehat{Q}}} u_{\mathcal{I}^i_{\widehat{Q}}}(t) + w_{\mathcal{I}^i_{\widehat{Q}}}(t);\ \ u_{\mathcal{I}^i_{\widehat{Q}}}(t) = K_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t). \tag{8}$$

According to the Bellman equation in an infinite horizon average cost MDP (Bertsekas (2007)), $\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}(t), u_{\mathcal{I}^i_{\widehat{Q}}}(t))$, $\forall i \in \mathcal{V}$, corresponding to the global policy π satisfies the fixed point equation

$$\lambda + \widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}(t), u_{\mathcal{I}^i_{\widehat{Q}}}(t)) = \sum_{j\in\mathcal{I}^i_{\text{GD}}} c_j(x_{\mathcal{I}^i_C}(t), u_{\mathcal{I}^i_C}(t)) + \mathbb{E}\big[\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}(t+1), K_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t+1))\big], \tag{9}$$

where $\lambda \in \mathbb{R}$ is a free parameter to satisfy the fixed point equation. Assuming a *linear architecture*, $\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})$ is parameterized as $\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}}) = \hat{q}_i^\intercal \phi_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})$, where $\phi_i(\cdot)$ is some known (possibly nonlinear) basis function of the state and control and $\hat{q}_i$ are unknown parameters.

Proposition 4.1 *Consider the cooperative LQR problem in (6). For any $i \in \mathcal{V}$, if $\hat{q}_i = \mathit{svec}\left(\widehat{Q}_i\right)$,* $\lambda = \left\langle \widehat{Q}_i, \sigma_w^2 \begin{bmatrix} \mathbb{I} \\ K_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ K_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix}^\intercal \right\rangle$, $\phi_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}}) = \mathit{svec}\left(\begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}} \\ u_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}} \\ u_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix}^\intercal \right)$, *and*

$\widehat{Q}_i = \begin{bmatrix} S_{\mathcal{I}^i_{\widehat{Q}}} & 0 \\ 0 & R_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} + \begin{bmatrix} A^\intercal_{\mathcal{I}^i_{\widehat{Q}}} \\ B^\intercal_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} \mathcal{L}\left(A_{\mathcal{I}^i_{\widehat{Q}}} + B_{\mathcal{I}^i_{\widehat{Q}}} K_{\mathcal{I}^i_{\widehat{Q}}}, S_{\mathcal{I}^i_{\widehat{Q}}} + K^\intercal_{\mathcal{I}^i_{\widehat{Q}}} R_{\mathcal{I}^i_{\widehat{Q}}} K_{\mathcal{I}^i_{\widehat{Q}}} \right) \begin{bmatrix} A_{\mathcal{I}^i_{\widehat{Q}}} & B_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix}$, *then the linear parameterization $\widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}}) = \hat{q}_i^\intercal \phi_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})$ satisfies (9).*

Denote $\phi_t = \text{svec}\left(\begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \\ u_{\mathcal{I}^i_{\widehat{Q}}}(t) \end{bmatrix} \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \\ u_{\mathcal{I}^i_{\widehat{Q}}}(t) \end{bmatrix}^\intercal \right)$, $\psi_t = \text{svec}\left(\begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \\ K_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \end{bmatrix} \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \\ K_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t) \end{bmatrix}^\intercal \right)$, $f = \text{svec}\left(\sigma_w^2 \begin{bmatrix} \mathbb{I} \\ K_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ K_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix}^\intercal \right)$, $\xi_t = \mathbb{E}\left[\text{svec}\left(\begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t+1) \\ u_{\mathcal{I}^i_{\widehat{Q}}}(t+1) \end{bmatrix} \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}}(t+1) \\ u_{\mathcal{I}^i_{\widehat{Q}}}(t+1) \end{bmatrix}^\intercal \right) \right]$, and $\hat{c}^i_t = \sum_{j\in\mathcal{I}^i_{\text{GD}}} c_j(x_{\mathcal{I}^i_C}(t), u_{\mathcal{I}^i_C}(t))$. With a slight abuse of notation, we use $\widehat{Q}_i$ to denote both the Q-function and the matrix parameterizing it. Then, (9) can be rewritten as

$$\hat{c}^i_t = \lambda + (\phi_t - \xi_t)\text{svec}(\widehat{Q}_i). \tag{10}$$

Since LSTDQ is an off-policy method, $\forall\ i \in \mathcal{V}$, we prescribe an arbitrary control law $u_{\mathcal{I}^i_{\widehat{Q}}}(t) = K^{\text{play}}_{\mathcal{I}^i_{\widehat{Q}}} x_{\mathcal{I}^i_{\widehat{Q}}}(t) + \eta_{\mathcal{I}^i_{\widehat{Q}}}(t)$, to generate a single trajectory $\{x_{\mathcal{I}^i_{\widehat{Q}}}(t), u_{\mathcal{I}^i_{\widehat{Q}}}(t), x_{\mathcal{I}^i_{\widehat{Q}}}(t+1)\}_{t=1}^T$, where $K^{\text{play}}_{\mathcal{I}^i_{\widehat{Q}}}$ may be different from $K_{\mathcal{I}^i_{\widehat{Q}}}$, and $\eta_{\mathcal{I}^i_{\widehat{Q}}}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_\eta^2 \mathbb{I}_{n_u|\mathcal{I}^i_{\widehat{Q}}|}\right)$ is a sufficiently exciting exploration noise for learning. We utilize the version of LSPI in Krauth et al. (2019), where new trajectory of samples are collected every iteration. Then (10) can be expressed in matrix form as

$$\hat{\mathbf{c}}_i = (\mathbf{\Phi} - \mathbf{\Xi} + \mathbf{F})\hat{q}_i^{\text{true}}, \tag{11}$$

where $\hat{q}_i^{\text{true}} = \text{svec}(\widehat{Q}_i^{\text{true}})$, $\mathbf{\Phi}^\intercal = [\phi_1, \phi_2, \cdots, \phi_T]$, $\mathbf{\Xi}^\intercal = [\xi_1, \xi_2, \cdots, \xi_T]$, $\hat{\mathbf{c}}_i^\intercal = [\hat{c}_1^i, \hat{c}_2^i, \cdots, \hat{c}_T^i]$, $\mathbf{F}^\intercal = [f_1, f_2, \cdots, f_T]$, and $\mathbf{\Psi}_+^\intercal = [\psi_2, \psi_3, \cdots, \psi_{T+1}]$. A least-squares solution to $\hat{q}_i^{\text{true}}$ in the error-in-variables problem (11) is given by

$$\hat{q}_i = (\mathbf{\Phi}^\intercal(\mathbf{\Phi} - \mathbf{\Psi}_+ + \mathbf{F}))^{-1}\mathbf{\Phi}^\intercal \hat{\mathbf{c}}_i. \tag{12}$$

To ensure that the least-squares estimate $\text{smat}(\hat{q}_i)$ is positive semi-definite, we perform a Euclidean projection onto the set of symmetric matrices lower bounded by $\zeta \cdot \mathbb{I}$ as in Krauth et al. (2019).

Policy improvement The policy improvement step in the LSPI algorithms in the unstructured single agent setting have a closed-form update. However, in our multi-agent setting, a closed-form update is not possible due to the observation constraints of the agents in $\mathcal{G}_O$. Instead, agent $i \in \mathcal{V}$ updates its policy K_i using (7) as

$$K_i \leftarrow K_i - \alpha \mathbb{E}\left[\nabla_{K_i} u_i \cdot \nabla_{u_i} \widehat{Q}_i(x_{\mathcal{I}^i_{\widehat{Q}}}, u_{\mathcal{I}^i_{\widehat{Q}}})\right] = K_i - 2\alpha \mathbb{E}\left[\mathbb{J}_i \widehat{Q}_i \begin{bmatrix} x_{\mathcal{I}^i_{\widehat{Q}}} \\ u_{\mathcal{I}^i_{\widehat{Q}}} \end{bmatrix} x_{\mathcal{I}^i_O}^\intercal\right],\ \forall\ i \in \mathcal{V}, \tag{13}$$

where α is the learning rate and $\mathbb{J}_i \in \mathbb{R}^{n_u \times (n_x + n_u)|\mathcal{I}^i_{\widehat{Q}}|}$ is a matrix where the sub-matrix corresponding to u_i is $\mathbb{I}_{n_u}$ and zero otherwise.

5. Theoretical Analysis

We present our main result on the sample complexity of the VD set decomposition for (6). Recall from Krauth et al. (2019) that a square matrix L is (τ, ρ)-stable if $\forall\ k \in \mathbb{Z}_{\geq 0}$, $\left\|L^k\right\| \leq \tau \rho^k$, where $\tau \geq 1$ and $\rho \in (0, 1)$. Let $n^i_{\hat{x}} = n_x|\mathcal{I}^i_{\widehat{Q}}|$, $n^i_{\hat{u}} = n_u|\mathcal{I}^i_{\widehat{Q}}|$. Theorem 5.1 below states the sample complexity and the estimation error of the Q-function parameter for the *direct* case.

Theorem 5.1 *Consider $\delta \in (0, 1)$. Let the initial global state and the global control (during sample generation) $\forall\ t$ satisfy $x(0) \sim \mathcal{N}(x_0, \Sigma_0)$, $u(t) = K^{play} x(t) + \eta_t$, $\eta(t) \sim \mathcal{N}(\mathbf{0}_{Nn_u}, \sigma_\eta^2 \mathbb{I}_{Nn_u})$, and $\sigma_\eta \leq \sigma_w$. For each $i \in \mathcal{V}$, let $K^{play}_{\mathcal{I}^i_{\widehat{Q}}}$, $K_{\mathcal{I}^i_{\widehat{Q}}}$ stabilize $(A_{\mathcal{I}^i_{\widehat{Q}}}, B_{\mathcal{I}^i_{\widehat{Q}}})$. Assume that $A_{\mathcal{I}^i_{\widehat{Q}}} + B_{\mathcal{I}^i_{\widehat{Q}}} K_{\mathcal{I}^i_{\widehat{Q}}}$ and $A_{\mathcal{I}^i_{\widehat{Q}}} + B_{\mathcal{I}^i_{\widehat{Q}}} K^{play}_{\mathcal{I}^i_{\widehat{Q}}}$ are (τ, ρ)-stable. Let $\mathfrak{P}_\infty = \mathcal{L}\left(A_{\mathcal{I}^i_{\widehat{Q}}} + B_{\mathcal{I}^i_{\widehat{Q}}} K_{\mathcal{I}^i_{\widehat{Q}}}, \sigma_w^2 \mathbb{I}_{n^i_{\hat{x}}} + \sigma_\eta^2 B_{\mathcal{I}^i_{\widehat{Q}}} B_{\mathcal{I}^i_{\widehat{Q}}}^\intercal\right)$ and $\bar{\sigma}_i = \sqrt{\tau^2\rho^4||\Sigma_0^{\hat{x}}|| + ||\mathfrak{P}_\infty|| + \sigma_w^2 + \sigma_\eta^2||B_{\mathcal{I}^i_{\widehat{Q}}}||^2}$. Suppose that T satisfies*

$$T \geq \tilde{O}(1) max\left\{(n^i_{\hat{x}} + n^i_{\hat{u}})^2, \frac{(n^i_{\hat{x}})^2(n^i_{\hat{x}} + n^i_{\hat{u}})^2||\widehat{K}^{play}_{\mathcal{I}^i_{\widehat{Q}}}||_+^4}{\sigma_\eta^4} \sigma_w^2 \bar{\sigma}_i^2 \frac{\tau^4||K_{\mathcal{I}^i_{\widehat{Q}}}||_+^8(||A_{\mathcal{I}^i_{\widehat{Q}}}||^2 + ||B_{\mathcal{I}^i_{\widehat{Q}}}||^2)^2}{\rho^4(1-\rho^2)^2}\right\}.$$

Then with probability at least $1-\delta$*, we have*

$$||\hat{q}_i^{true} - \hat{q}_i^{direct}|| \leq \frac{\tilde{O}(1)(n_{\widehat{x}}^i + n_{\widehat{u}}^i)||K_{\mathcal{I}_{\widehat{Q}}^i}^{play}||_+^2}{\sigma_\eta^2\sqrt{T}}\sigma_w\bar{\sigma}_i||\widehat{Q}_i^{true}||_F\frac{\tau^2||K_{\mathcal{I}_{\widehat{Q}}^i}||_+^4(||A_{\mathcal{I}_{\widehat{Q}}^i}||^2 + ||B_{\mathcal{I}_{\widehat{Q}}^i}||^2)}{\rho^2(1-\rho^2)},$$

where $\tilde{O}(1)$ *hides polylog* $\left(\frac{T}{\delta},\ \frac{1}{\sigma_\eta^4},\ \tau,\ n_{\widehat{x}}^i,\ ||\Sigma_0||,\ ||K_{\mathcal{I}_{\widehat{Q}}^i}^{play}||,\ ||\mathfrak{P}_\infty||\right)$.

To interpret the result, observe that achieving an $\epsilon-$close estimate of $\hat{q}_i^{\text{true}}$ requires at most

$$T \leq \tilde{O}(1)\max\left(\frac{W_i^2(n_{\widehat{x}}^i + n_{\widehat{u}}^i)^3}{\sigma_\eta^4\epsilon^2}||\widehat{Q}_i^{\text{true}}||^2, \frac{W_i^2(n_{\widehat{x}}^i)^2(n_{\widehat{x}}^i + n_{\widehat{u}}^i)^2}{\sigma_\eta^4}\right) \text{ samples,}$$

where $W_i = ||K_{\mathcal{I}_{\widehat{Q}}^i}^{\text{play}}||_+^2\sigma_w\bar{\sigma}_i\frac{\tau^2||K_{\mathcal{I}_{\widehat{Q}}^i}||_+^4(||A_{\mathcal{I}_{\widehat{Q}}^i}||^2 + ||B_{\mathcal{I}_{\widehat{Q}}^i}||^2)}{\rho^2(1-\rho^2)}$. We focus on the scaling w.r.t. the state and control dimensions $(n_{\widehat{x}}^i, n_{\widehat{u}}^i)$. For agent $i \in \mathcal{V}$, the *direct* decomposition based Algorithm 1 is more sample efficient than the centralized case (whose state dimension is Nn_x and control dimension is Nn_u) if $\mathcal{I}_{\widehat{Q}}^i \subset \mathcal{V}$, i.e., $|\mathcal{I}_{\widehat{Q}}^i| < N$. Under similar pre-conditions to Theorem 5.1, we prove in Appendix G that in the *indirect* case, achieving an $\epsilon-$close estimate of q_i^{true} requires at most

$$T \leq \tilde{O}(1)\max_{j\in\mathcal{I}_{\text{GD}}^i}\left(\max\left(\frac{W_j^2(n_x^j + n_u^j)^3}{\sigma_\eta^4(w_j^i)^2\epsilon^2}||Q_j||^2, \frac{W_j^2(n_x^j)^2(n_x^j + n_u^j)^2}{\sigma_\eta^4}\right)\right) \text{ samples,}$$

where $W_j = ||K_{\mathcal{I}_Q^i}^{\text{play}}||_+^2\sigma_w\bar{\sigma}_j\frac{\tau^2||K_{\mathcal{I}_Q^i}||_+^4(||A_{\mathcal{I}_Q^i}||^2 + ||B_{\mathcal{I}_Q^i}||^2)}{\rho^2(1-\rho^2)}$, and $w_1^i, \cdots, w_{|\mathcal{I}_{\text{GD}}^i|}^i \in \mathbb{R}_+$ satisfy $\sum_{j=1}^{|\mathcal{I}_{\text{GD}}^i|} w_j^i = 1$. The user-defined weights w_j^i can be construed as the relative importance of the estimation accuracies of $q_j, \forall\ j \in \mathcal{I}_{GD}^i$. By a specific choice of w_j^i and the analysis in Appendix H, we show that the worst-case sample complexity of the *indirect* decomposition-based Algorithm 1 is equal to that of the *direct* case and strictly better if $\mathcal{I}_Q^j \subset \mathcal{I}_{\widehat{Q}}^i,\ \forall\ j \in \mathcal{I}_{\text{GD}}^i$. The necessary and sufficient graphical conditions to ensure strictly better sample efficiency of direct and indirect methods are derived in Lemma E.1 (Appendix E).

6. Simulations

Consider N agents, whose dynamics are given in (Krauth et al., 2019, Example 1). We investigate the performance of Algorithm 1 in two examples. **Example 1**: We prescribe the inter-agent couplings $\mathcal{E}_R = \{(i,i)|i \in \mathcal{V}\},\ \mathcal{E}_S = \mathcal{E}_O = \{(j,j-1),(j,j+1)|j = 2k+1 \leq N, k \in \mathbb{N}\} \cup \{(1,N)\} \cup \{(N,1),\ \text{if } N = 2k+1 \text{ for some } k \in \mathbb{N}\} \cup \{(i,i)|i \in \mathcal{V}\}$, which yield $\mathcal{E}_Q = \mathcal{E}_O$. **Example 2**: We prescribe the inter-agent couplings $\mathcal{E}_S = \{(i,i)|i \in \mathcal{V}\},\ \mathcal{E}_O = \mathcal{E}_R = \{(1,j)|j \in \mathcal{V}\} \cup \{(i,i)|i \in \mathcal{V}\}$, corresponding to a leader-follower network which yields $\mathcal{E}_Q = \mathcal{E}_R$. Example 1 demonstrates the case where $\mathcal{I}_Q^i \subset \mathcal{I}_{\widehat{Q}}^i \subset \mathcal{V}$ holds $\forall\ i \in \mathcal{V}$ with $\max_i(|\mathcal{I}_{\widehat{Q}}^i| - |\mathcal{I}_Q^i|) = 4$ for any N. In contrast, Example 2 highlights the case where for the leader $l \in \mathcal{V}$, it always holds that $\mathcal{I}_Q^l \subset \mathcal{I}_{\widehat{Q}}^l = \mathcal{V}$. We compare the direct and indirect decompositions with a 'centralized' (CTDE) MALSPI baseline in which each agent learns $Q_i(x(t), u(t))$ in the policy evaluation step while the

individual control is still subject to $\mathcal{G}_O$. In addition, we examine the 'undecomposed direct' method, in which each agent learns $\widehat{Q}_i(x(t), u(t))$ in the policy evaluation step, to study the effect of decomposition of the Q-function on the performance of the learned controller. The decomposition of the above baselines and the simulation parameters are summarized in Table 2, 3 (Appendix J) respectively. We simulate $N = 8$ agents with an evaluation trajectory length of $T_{\text{eval}} = 500$ steps. Fig. 1a and 1b show the comparison of the total average cost for 20 MC simulations using different Q-function architectures for Example 1 and 2, respectively. In both cases, we observe that the indirect method has the fastest rate of convergence and lowest average cost followed by the direct, undecomposed direct, and centralized methods. Given a sufficient number of samples, the centralized and undecomposed direct methods perform comparably to the direct and indirect methods. This empirically corroborates the strictly better sample efficiency of the Q-function decomposition as discussed in Section 5. The difference in the performance of the direct and indirect methods in the low-sample regime is more pronounced in Example 2 than in Example 1. This is attributed to the comparable sparsity of $\mathcal{I}^i_Q$ and $\mathcal{I}^i_{\widehat{Q}}$ (since $\max_i(|\mathcal{I}^i_{\widehat{Q}}| - |\mathcal{I}^i_Q|) = 4$) in Example 1, resulting in comparable sample efficiency for both methods. However, in Example 2, $\mathcal{I}^1_{\widehat{Q}} = \mathcal{V}$ for agent 1 (the leader), rendering the direct method to be less sample efficient compared to the indirect method. Table 1 (Appendix J) summarizes the average computational time per iteration of Algorithm 1 for $N = 8,\ 20,\ 40$; $T = 500$ steps in Example 1, 2. We observe that the computational time for the centralized and undecomposed direct methods scales exponentially with the number of agents whereas the time for the direct and indirect decomposition scales only with the number of agents in the $\mathcal{I}^i_Q$ and $\mathcal{I}^i_{\widehat{Q}}$, respectively. This corroborates the computational savings and scalability achieved by the Q-function decomposition proposed in Section 3.

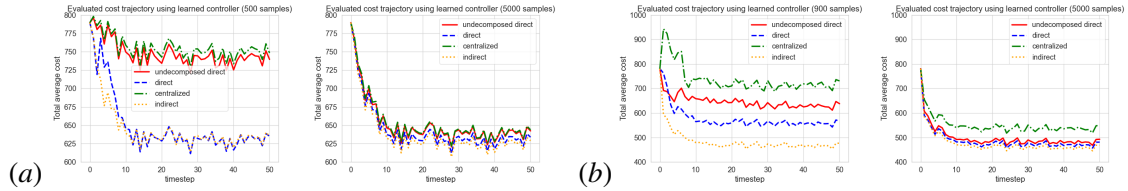

Figure 1: Comparison of the total average cost for Example 1 (a) and Example 2 (b) using direct, indirect, undecomposed direct, and centralized Q-function architectures in Algorithm 1.

7. Conclusion

We develop a systematic approach to leverage inter-agent coupling information and perform exact decompositions of the individual Q-function and the gradient of the global Q-function. Based on the decompositions, we introduce a cooperative deterministic policy gradient theorem and a cooperative MALSPI algorithm. We establish the theoretical sample and error guarantees for the obtained decomposition and provide necessary and sufficient graphical conditions for better sample efficiency of the proposed decomposition. We empirically validate the improved sample and computational efficiency using two numerical examples. Our future work will investigate the effect of approximate decompositions on the Q-function estimation and the convergence of the MALSPI algorithm.

Acknowledgments

This work was partly supported by the U.S. DEVCOM Army Research Laboratory (ARL) under Cooperative Agreement W911NF2120219 and National Science Foundation (NSF) award #2212582. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF or the U.S. Government.

References

Siavash Alemzadeh and Mehran Mesbahi. Distributed Q-learning for dynamically decoupled systems. In *2019 American Control Conference (ACC)*, pages 772–777. IEEE, 2019.

Bassam Bamieh, Fernando Paganini, and Munther A Dahleh. Distributed control of spatially invariant systems. *IEEE Transactions on automatic control*, 47(7):1091–1107, 2002.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume II*. Athena scientific, 2007.

Vincent D Blondel and John N Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926*, 2017.

David Gamarnik, David A Goldberg, and Theophane Weber. Correlation decay in random decision networks. *Mathematics of Operations Research*, 39(2):229–261, 2014.

Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored MDPs. *Advances in neural information processing systems*, 14, 2001.

Carlos Guestrin, Michail Lagoudakis, and Ronald Parr. Coordinated reinforcement learning. In *ICML*, volume 2, pages 227–234. Citeseer, 2002.

Yu-Chi Ho and K'ai-Ching Chu. Team decision theory and information structures in optimal control problems–part i. *IEEE Transactions on Automatic Control*, 17(1):15–22, 1972. doi: 10.1109/TAC.1972.1099850.

Gangshan Jing, He Bai, Jemin George, and Aranya Chakrabortty. Model-free optimal control of linear multiagent systems via decomposition and hierarchical approximation. *IEEE Transactions on Control of Network Systems*, 8(3):1069–1081, 2021a.

Gangshan Jing, He Bai, Jemin George, Aranya Chakrabortty, and Piyush K Sharma. Learning distributed stabilizing controllers for multi-agent systems. *IEEE Control Systems Letters*, 6:301–306, 2021b.

Gangshan Jing, He Bai, Jemin George, Aranya Chakrabortty, and Piyush K Sharma. Distributed multi-agent reinforcement learning based on graph-induced local value functions. *IEEE Transactions on Automatic Control*, 2024.

Soummya Kar, José M. F. Moura, and H. Vincent Poor. $\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through $\mathrm{consensus} + \mathrm{innovations}$. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013. doi: 10.1109/TSP.2013.2241057.

Mruganka Kashyap and Laurent Lessard. Guaranteed stability margins for decentralized linear quadratic regulators. *IEEE Control Systems Letters*, 7:1778–1782, 2023.

Jelle R. Kok and Nikos Vlassis. *Using the max-plus algorithm for multiagent decision making in coordination graphs*, page 1–12. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3540354379.

Daphne Koller and Ronald Parr. Computing factored value functions for policies in structured mdps. In *IJCAI*, volume 99, pages 1332–1339, 1999.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Karl Krauth, Stephen Tu, and Benjamin Recht. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 32, 2019.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

Wenhao Li, Bo Jin, Xiangfeng Wang, Junchi Yan, and Hongyuan Zha. F2a2: Flexible fully-decentralized approximate actor-critic for cooperative multi-agent reinforcement learning. *Journal of Machine Learning Research*, 24(178):1–75, 2023.

Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67(12):6429–6444, 2021.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Sergio Valcarcel Macua, Aleksi Tukiainen, Daniel García-Ocaña Hernández, David Baldazo, Enrique Munoz de Cote, and Santiago Zazo. Diff-DAC: Distributed actor-critic for average multitask deep reinforcement learning. In *Adaptive Learning Agents (ALA) Conference*, 2018.

Nader Motee and Ali Jadbabaie. Optimal control of spatially distributed systems. *IEEE Transactions on Automatic Control*, 53(7):1616–1629, 2008.

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. IEEE, 2024.

Christos H Papadimitriou and John Tsitsiklis. Intractable problems in control theory. *SIAM journal on control and optimization*, 24(4):639–654, 1986.

Youngsuk Park, Ryan Rossi, Zheng Wen, Gang Wu, and Handong Zhao. Structured policy iteration for linear quadratic regulator. In *International Conference on Machine Learning*, pages 7521–7531. PMLR, 2020.

Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, and Junwu Xiong. Value propagation for decentralized networked deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, pages 256–266. PMLR, 2020.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*, 2018.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, 2019.

Michael Rotkowitz and Sanjay Lall. A characterization of convex problems in decentralized control. *IEEE transactions on Automatic Control*, 50(12):1984–1996, 2005.

Parikshit Shah and Pablo A. Parrilo. $\mathcal{H}_2$-optimal decentralized control over posets: A state-space solution for state-feedback. *IEEE Transactions on Automatic Control*, 58(12):3084–3096, 2013. doi: 10.1109/TAC.2013.2281881.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pages 387–395. PMLR, 2014.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Hans S Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

Lintao Ye, Hao Zhu, and Vijay Gupta. On the sample complexity of decentralized linear quadratic regulator with partially nested information structure. *IEEE Transactions on Automatic Control*, 68(8):4841–4856, 2022.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International conference on machine learning*, pages 5872–5881. PMLR, 2018.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940, 2021.

Runyu Cathy Zhang, Weiyu Li, and Na Li. On the optimal control of network LQR with spatially-exponential decaying structure. In *2023 American Control Conference (ACC)*, pages 1775–1780. IEEE, 2023.

Yan Zhang and Michael M Zavlanos. Cooperative multi-agent reinforcement learning with partial observations. *IEEE Transactions on Automatic Control*, 2023. doi: 10.1109/TAC.2023.3288025.