Model-Free Optimal Control of Linear Multi-Agent Systems via Decomposition and Hierarchical Approximation

Gangshan Jing, He Bai, Jemin George and Aranya Chakrabortty

Abstract—Designing the optimal linear quadratic regulator (LQR) for a large-scale multi-agent system (MAS) is time-consuming since it involves solving a large-size matrix Riccati equation. The situation is further exasperated when the design needs to be done in a model-free way using schemes such as reinforcement learning (RL). To reduce this computational complexity, we decompose the large-scale LQR design problem into multiple sets of smaller-size LQR design problems. We consider the objective function to be specified over an undirected graph, and cast the decomposition as a graph clustering problem. The graph is decomposed into two parts, one consisting of multiple decoupled subgroups of connected components, and the other containing edges that connect the different subgroups. According to the resulting controller has a hierarchical structure, consisting of two components. The first component optimizes the performance of each decoupled subgroup by solving the smaller-size LQR design problem in a model-free way using an RL algorithm. The second component accounts for the objective coupling different subgroups, which is achieved by solving a least squares problem in one shot. Although suboptimal, the hierarchical controller adheres to a particular structure as specified by the inter-subgroup coupling in the objective function and by the decomposition strategy. Mathematical formulations are established to find a decomposition that minimizes required communication links or reduces the optimality gap. Numerical simulations are provided to highlight the pros and cons of the proposed designs.

Index Terms—Decomposition, model-free control, reinforcement learning, linear quadratic regulator, large-scale networks

I. INTRODUCTION

The general distributed control of multi-agent systems (MASs) has been approached from a variety of perspectives, including distributed optimization [1], [2], [3], game theory [4], [5], graph theory and derivative-free optimization methods [6], graphing methods [7], LASSO [8], GraSP [9], and PALM [10]. When applied to large-scale networks, however, these methods often result in poor numerical performance due to the computation of extremely large-dimensional gain matrices for implementing even simple state-feedback laws such as the linear quadratic regulator (LQR). In other words, they produce encouraging results when run offline, but their execution becomes time-consuming and numerically expensive when real-time control actions need to be taken. The problem becomes even more complex when the MAS model is unknown to the designer and model-free techniques such as reinforcement learning (RL) [11], [12] need to be used. Various off-the-shelf RL algorithms such as actor-critic methods, Q-learning, and adaptive dynamic programming have been proposed in the recent literature [13]-[17]. Though we propose to use the ADP technique of [17], naively applying such an approach would require global aggregation of agent inputs and state trajectories before solving the large-scale RL problem centrally. This drastically increases the computational complexity, communication overhead, and results in long learning time while introducing the single point of failure issue.

In this paper, we propose a hierarchical RL-based optimal control scheme where the hierarchy follows from decomposing a large-dimensional LQR control problem into several sets of smaller-sized LQR problems. We consider a system consisting of multiple agents, where each agent has its own decoupled linear dynamics. The control objective for the MAS is posed as a coupled integral quadratic cost. We use a connected undirected graph to represent the inter-agent coupling in this LQR objective, and pose the decomposition as a clustering problem for this graph. The graph is decomposed into two parts, one consisting of multiple decoupled subgroups of connected components, and the other containing edges that connect the different subgroups. Accordingly, the proposed RL-based optimal controller is composed of two components, one optimizing the performance of each decoupled subgroup by solving the smaller-size LQR design, the other accounting for the performance coupling the different subgroups, which is achieved by solving a least squares problem in one shot.

Because of the decomposition, our hierarchical RL controller is suboptimal. But the benefit, aside from drastic reduction of learning time, is that the controller exhibits a special structure as specified by the inter-subgroup coupling in the objective function and by the decomposition strategy due to which it needs lesser number of communication links than the optimal LQR. We show that one can find an optimal decomposition that minimizes the number of inter-agent communication links by solving a mixed-integer quadratic program (MIQP). Moreover, choosing the decomposition obtained by solving a minimum k-cut graph partitioning problem in fact helps reduce the optimality gap.

Model-based designs for hierarchical control have recently garnered a lot of attention [18]. The model-free versions of these designs, however, are much more challenging as both the learning mechanism and the controller implementation need to obey structural constraints [19]. Another important point to note here is that a common assumption that is often used for both model-based and model-free [20] hierarchical control is that the plant exhibits some kind of separation property in its
dynamics. The most common example is time-scale separation arising from clustering of the network nodes of the plant. Our problem setting, however, is different. In our case, the plant consists of decoupled sets of agents, and the clustering is imposed on the graph that defines the control objective. This clustering then defines the final structure of the hierarchical controller. Graph clustering has been used in control designs using a myriad of tools such as spectral factorization [21], convex optimization [22], singular perturbation theory [23], and trajectory sensitivity techniques [24]. In [25]-[28], graph clustering has been used to decompose a MAS with dynamics determined by the graph. The graph clustering formulation in our problem is different from these results as our goal is to find the best way to cluster the graph underlying the control objective such that it would minimize the number of inter-agent communication links required for the hierarchical controller or reduce the suboptimality gap. Our design is also fundamentally different from most cooperative control problems in the MAS literature [29], [30], [31], where the objective is to stabilize agents to a desired equilibrium set describing a collective behavior. Although a stabilizing controller may optimize some cost function [32], the cost function is not predefined and has to satisfy specific conditions.

The hierarchical control formulation in our work is inspired by the model-based approaches reported in [33] and [34], where in the latter the cost function is characterized based on a graph that couples the states of the agents. Similar cost functions can also be found in [35], [36], [37]. We extend these methods to a far more general formulation, and propose a decomposition approach for model-free optimal control of heterogeneous MAS. An application example on formation maneuver control is given to validate effectiveness of the hierarchical controller. Some preliminary results on the heterogeneous design was reported in [38]. Current results on stability proofs, analysis on the optimality gap, as well as the heterogeneous design was reported in [38]. Current results on stability proofs, analysis on the optimality gap, as well as the optimal decomposition are new additions.

The rest of the paper is organized as follows. Section II formulates the model-free optimal control problem for linear MAS. Section III proposes the hierarchical control framework via decomposition for heterogeneous MAS. Section IV analyzes the influence of the decomposition on the required number of communication links and the performance gap, and presents an MIQP problem to find an optimal decomposition strategy. Section V shows an application of the hierarchical control for heterogeneous MAS to formation maneuver control. Section VI concludes the paper. Stability proofs and other theoretical results are presented in the appendix.

**Notation:** Throughout the paper, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denotes an undirected graph with $N$ vertices, where $\mathcal{V} = \{1, \ldots, N\}$ is the set of vertices, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges; $0_{p \times q}$ denotes the $p \times q$ zero matrix; $I_d$ is the $d \times d$ identity matrix; $e_i \in \mathbb{R}^N$ is a $N$-dimensional unit vector with the $i$-th element being 1 and other elements being 0; $\otimes$ denotes the kronecker product. Define $\text{diag}\{A_1, \ldots, A_N\}$ as a block diagonal matrix with $A_i$'s on the diagonal. Given a matrix $A \in \mathbb{R}^{n \times n}$, $\lambda_i(A)$ denotes the $i$-th eigenvalue of $A$ such that $\lambda_1(A) \leq \cdots \leq \lambda_n(A)$, $\lambda_{\min}(A) = \lambda_1(A)$, $\lambda_{\max}(A) = \lambda_n(A)$, $A \succ 0$ implies that $A$ is positive definite, $A \succeq 0$ implies that $A$ is positive semi-definite, and $\text{tr}(A)$ denotes the trace of $A$; Similarly, $\sigma_1(A) \leq \cdots \leq \sigma_n(A)$ are singular values of $A$, $\sigma_{\min}(A) = \sigma_1(A)$, $\sigma_{\max}(A) = \sigma_n(A)$. We use $\text{cond}(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$ to denote the condition number of matrix $A$. For matrices $A, B \succeq 0$, $A \preceq B$ implies that $B - A \succeq 0$, $A \succeq B$ if $B \preceq A$.

**II. Problem Formulation**

Consider a MAS composed of $N$ agents. Each agent $i$ is a linear time-invariant system described as

$$\dot{x}_i = A_i x_i + B_i u_i, \quad i = 1, \ldots, N$$

where $x_i \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^m$ are the state and control input of agent $i$. Throughout the paper, we will consider $A_i$ and $B_i$ as unknown, but their dimensions are known. Let $x = (x_1^T, \ldots, x_N^T)^T$, $u = (u_1^T, \ldots, u_N^T)^T$, $A = \text{diag}\{A_1, \ldots, A_N\}$, $B = \text{diag}\{B_1, \ldots, B_N\}$. The overall model of the system is written in a compact form as

$$\dot{x} = Ax + Bu.$$  

We assume that the agent dynamics are coupled through the cost function that needs to be minimized using a state-feedback LQR controller. Without loss of generality, we use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to describe the coupling relationship among the agents in the cost function. It is important to note that the graph $\mathcal{G}$ is not a communication or interaction graph between the agents. It is a graph that defines the coupling between the agent dynamics in the control objective. Similar settings have been considered in [33] and [35] with common applications in formation control, paper machine control and monitoring networks of cameras. The control objective is given as

$$J(x(0), u) = \int_0^\infty x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau) \, d\tau,$$

where $Q \succeq 0$ and $R \succ 0$ are in the following forms:

$$Q = \tilde{Q} + \tilde{G}, \quad R = \text{diag}\{R_1, \ldots, R_N\},$$

with $\tilde{Q} = \text{diag}\{\tilde{Q}_{11}, \ldots, \tilde{Q}_{NN}\}$ representing the subsystem-level objective, $Q_{ii} \in \mathbb{R}^{n \times n}$ and $R_i \succ 0$. $R_i \in \mathbb{R}^{n \times n}$ and $R_i \succ 0$ for $i = 1, \ldots, N$. $G = [G_{ij}] \in \mathbb{R}^{N \times N}$ is a Laplacian matrix corresponding to graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, i.e., $G_{ii} = \sum_{j=1, j \neq i}^N |G_{ij}|$ for $i = 1, \ldots, N$, and $G_{ij} < 0$ if $(i, j) \in \mathcal{E}$ and $i \neq j$; $\tilde{Q} \in \mathbb{R}^{n \times n}$ and $\tilde{Q} \succeq 0$. The second component in $Q$ represents the objective across the different subsystems. We observe that if $\mathcal{G}$ is not connected, then the optimal control problem can be decomposed immediately according to its independent connected components. Without loss of generality, we assume $\mathcal{G}$ is connected throughout the rest of the paper.

Given the MAS (1) with state $x$ measurable, we would ideally like to find the controller $u^*$ such that the performance index (3) with matrices $Q$ and $R$ defined in (4) is minimized despite $A$ and $B$ being unknown. We make the following assumption to guarantee existence and uniqueness of the optimal controller.

**Assumption 1:** The pair $(A, B)$ is controllable and $(Q^{1/2}, A)$ is observable.
For heterogeneous MAS with \( n \geq 2, Q > 0 \) is sufficient but not necessary for observability of \( (Q^{1/2}, A) \). However, if \( n = 1 \), then \((Q^{1/2}, A)\) is observable only if \( Q > 0 \). When \( A \) and \( B \) are known, the optimal controller for system \( (Q, A) \) is \[ u = -Kx = -R^{-1}B^TPx \] (see [39]), where \( P \) is the solution of the following algebraic Riccati equation:

\[
P A + A^T P + Q - P B R^{-1} B^T P = 0. \tag{5}
\]

For MAS \( (G, B) \), \( K \in \mathbb{R}^{MN \times nN} \) can be partitioned into \( N^2 \) blocks \( K(i, j) \in \mathbb{R}^{m \times n} \). Accordingly, the control input for agent \( i \) can be written as \( u_i = \sum_{j=1}^{N} K(i, j)x_j \). We define the communication graph associated with controller \( u \) as \( \mathcal{G_c}(u) = (V, \mathcal{E}_c(u)) \), where

\[
\mathcal{E}_c(u) = \{(i, j) \in V \times V : K(i, j) \neq 0_{m \times n}\}. \tag{6}
\]

Therefore, communication between agents \( i \) and \( j \) is not required if \( (i, j) \notin \mathcal{E}_c \). In the model-free case, where \( A \) and \( B \) are unknown, equation [5] can be transformed into an equation that is independent of \( A \) and \( B \), and is based on the measurement of the states and the control inputs [15]. [17]. The matrix \( P \) can then be obtained by implementing an RL algorithm, which solves a least squares problem at each iteration.

Both the model-based and model-free control methods, however, are time-consuming if the MAS is of large scale. The main purpose of this paper is to provide an alternative approach for synthesizing the controller based on a decomposition of the cost function and a clustering of the agent dynamics in the MAS, which reduces learning time significantly. This decomposition is described next.

III. HIERARCHICAL CONTROL FOR MAS

To reduce learning time, we propose a hierarchical control design that seeks a suboptimal controller for the general case when the agents have non-identical dynamics. As mentioned earlier, the underlying approach is to decompose the cost function to multiple decoupled cost functions with respect to multiple subgroups of the MAS. For ease of understanding the hierarchical control scheme, we assume that the decomposition strategy is given, which means that the agents have already been decomposed into multiple subgroups. Two optimal decomposition strategies are presented in Section IV.

A. Hierarchical Control Framework

Suppose that the vertex set \( V \) is decomposed into \( s \leq N \) disjoint vertex sets \( V_j, j = 1, \ldots, s \). The \( j \)-th subgroup is composed of \( N_j \) nodes \( \{j_1, \ldots, j_{N_j}\} \). Let \( n_j = nN_j \) and \( m_j = mN_j \). The dynamics of the \( j \)-th subgroup is written as

\[
\dot{x}_j = A_j x_j + B_j u_j, \quad j = 1, \ldots, s, \tag{7}
\]

where \( x_j = (x_{j,1}^T, \ldots, x_{j,N_j}^T)^T \in \mathbb{R}^{n_j}, u_j = (u_{j,1}^T, \ldots, u_{j,N_j}^T)^T \in \mathbb{R}^{m_j}, A_j = \text{diag}\{A_{j,1}, \ldots, A_{j,N_j}\} \in \mathbb{R}^{n_j \times n_j}, B_j = \text{diag}\{B_{j,1}, \ldots, B_{j,N_j}\} \in \mathbb{R}^{n_j \times m_j}. \]

Without loss of generality, we assume that the agents belonging to the same subgroup have contiguous indices, i.e., \( j_{i+1} = j_i + 1 \) and \( k_1 = j_{N_j} + 1 \) for \( k = j + 1 \). In the case when the indices of agents are not in such an order, one can relabel them to make this condition satisfied. The matrix \( G \) can then be decomposed as

\[
G = G_1 + G_2, \tag{8}
\]

where \( G_1 \) is a block-diagonal Laplacian matrix with \( s \) blocks and each block corresponds to a subgraph of \( \mathcal{G_c} \), and \( G_2 \) is a Laplacian matrix where \( G_2(i, j) = \sum_{j=1}^{N} |G_2(i, j)| \) that describes the couplings between the different subgroups. Then both \( G_1 \) and \( G_2 \) are positive semi-definite.

Given the decomposition \( V = V_1 \cup \ldots \cup V_s \), there are two steps to construct \( G_2 \):

1. Step 1. Remove each block on the diagonal of \( G \) corresponding to each subgroup \( V_j \). The remaining off-diagonal elements are viewed as the off-diagonal elements of \( G_2 \).

2. Step 2. Compute the diagonal elements of \( G_2 \) by using \( G_2(i, i) = \sum_{j=1}^{N} |G_2(i, j)| \).

Once \( G_2 \) is constructed, \( G_1 \) will be determined accordingly. In Fig. 1 we give two examples to demonstrate two different decompositions for the same \( G \).

Similar to \( G \), the four matrices \( Q, \tilde{Q}, \hat{Q} \) and \( R \) can also be transformed corresponding to the new indices of the agents. For simplicity of notation, we assume that \( Q, \tilde{Q}, \hat{Q} \) and \( R \) correspond directly to the relabelled agents. From [4], we have

\[
Q = \hat{Q} + (G_1 + G_2) \otimes \tilde{Q}. \tag{9}
\]

We define

\[
\hat{Q} = \tilde{Q} + G_1 \otimes \hat{Q}, \tag{10}
\]

and write \( \hat{R} = \text{diag}\{R_{j_1}, \ldots, R_{j_{N_j}}\} \in \mathbb{R}^{m_j \times m_j} \) for the \( j \)-th subgroup. Then [4] can be written as

\[
Q = \hat{Q} + G_2 \otimes \hat{Q}, \quad R = \text{diag}\{\hat{R}_1, \ldots, \hat{R}_s\} \tag{11}
\]

where \( \hat{Q} = \text{diag}\{\hat{Q}_1, \ldots, \hat{Q}_s\} \).

Now the global network level control objective in [3] can be decomposed into group-level objectives \( J_j, j = 1, \ldots, s \) and network-level objective \( J_G \), i.e.,

\[
J(x(0), u) = \sum_{j=1}^{s} J_j(x_j(0), u_j) + J_G(x(0), u), \tag{12}
\]
where
\[
J_j(x_j(0), u_j) = \int_0^\infty x_j^T(\tau) \dot{Q}_j x_j(\tau) + u_j^T(\tau) \dot{R}_j u_j(\tau) \, d\tau,
\]
and
\[
J_G(x(0), u) = \int_0^\infty x^T(\tau) (G_2 \otimes \dot{Q}) x(\tau) \, d\tau.
\]
The algebraic matrix Riccati equation corresponding to the group-level integral quadratic cost in (13) is
\[
\mathcal{P}_j A_j + A_j^T \mathcal{P}_j + \dot{Q}_j - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j = 0, \quad i = 1, \ldots, s. \tag{15}
\]

From Assumption 1 \((A_j, B_j)\) is controllable. However, \((\dot{Q}_j^{1/2}, A_j)\) may not be observable for each \(j = 1, \ldots, s\). To guarantee existence and uniqueness of the solution to (15), we make the following assumption:

**Assumption 2:** \((\dot{Q}_j^{1/2}, A_j)\) is observable for \(j = 1, \ldots, s\).

**Remark 1:** For a general matrix \(A_j\), Assumption 2 may be satisfied even when \(\dot{Q}_j\) is singular. Observability of \((\dot{Q}_j^{1/2}, A)\) and \((\dot{Q}_j^{1/2}, A)\) can both be guaranteed without knowing the dynamics of each agent if \(\dot{Q}_j > 0\) for all \(j = 1, \ldots, s\).

The decomposition of the global control objective as given in (12) allows the individual groups to solve for their local optimal control gains (that minimizes (13)) in parallel. Next, we present an approximate control to account for the coupled objective given in (14).

Motivated by (13), we define
\[
\mathcal{R}^{-1} = R^{-1} + \dot{R}, \tag{16}
\]
where the expression for \(\dot{R}\) will be derived shortly. Now replacing \(R^{-1}\) with \(\mathcal{R}^{-1}\) in (5) yields
\[
\mathcal{P}_j A_j + A_j^T \mathcal{P}_j + Q - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j = \mathcal{P}_j A_j + A_j^T \mathcal{P}_j + Q - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j + G_2 \otimes \dot{Q} - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j
\]
developed part
\[
= \text{diag}\{P_j A_j, A_j^T P_j + \dot{Q}_j - P_j B_j \dot{R}_j B_j^T P_j\}
\]
with \(G_2 \otimes \dot{Q} - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j\), indicating that the Riccati equation can be decomposed into multiple decoupled smaller-sized Riccati equations if
\[
G_2 \otimes \dot{Q} - \mathcal{P}_j B_j \dot{R}_j B_j^T \mathcal{P}_j = 0. \tag{17}
\]

Therefore, we select \(\dot{R}\) as the solution of (17), where the block diagonal matrix \(\mathcal{P} = \text{diag}\{P_1, \ldots, P_s\}\) is obtained by solving the set of \(s\) decoupled Riccati equations given in (15). Now, the controller is designed hierarchically as
\[
u_h = -R^{-1}B^T \mathcal{P}_x = -R^{-1B^T \mathcal{P}_x - \dot{R}B^T \mathcal{P}_x}, \tag{18}
\]
where the first term is the local controller that can be obtained by solving multiple decoupled smaller-size Riccati equations, and the second term is the global controller based on \(\dot{R}\) solved from (17).

However, it may not always be possible to find a \(\dot{R}\) satisfying (17). If \(\dot{B}\) is square and non-singular (i.e., each agent is a fully actuated system), then \(\dot{R}\) follows simply as
\[
\dot{R} = (\mathcal{P}B)^{-1}(G_2 \otimes \dot{Q})(B^T \mathcal{P})^{-1}.
\]
However, this expression does not hold when \(\dot{B} \in \mathbb{R}^{nN \times mN}\) is rectangular and thus not invertible. In that case one may obtain \(\dot{R}\) by solving the following least squares semidefinite program (LSSDP):
\[
\min_{\dot{R}} \|\mathcal{P}\dot{B}\dot{R}B^T \mathcal{P} - G_2 \otimes \dot{Q}\|_F, \tag{19}
\]
s.t. \(\dot{R} \succeq 0\).

Let \((\mathcal{P}B)^+\) be the Moore-Penrose inverse of \(\mathcal{P}B\). In (40), it was shown that the solution \((\mathcal{P}B)^+ (G_2 \otimes \dot{Q})(\mathcal{P}B)^+\) has minimum norm among all solutions to (19). Therefore, one may compute \(\dot{R}\) as
\[
\dot{R} = (\mathcal{P}B)^+ (G_2 \otimes \dot{Q})(\mathcal{P}B)^+ . \tag{20}
\]
Specifically, when \(\mathcal{B} \in \mathbb{R}^{nN \times mN}\) is full column rank, the least square solution is uniquely determined by
\[
\dot{R} = ((\mathcal{P}B)^T \mathcal{P}B)^{-1} (\mathcal{P}B)^T (G_2 \otimes \dot{Q}) \mathcal{P}B ((\mathcal{P}B)^T \mathcal{P}B)^{-1} .
\]

Instead of minimizing the original objective function (3), the approximate controller (18) minimizes
\[
J(x(0), u) = \int_0^\infty x^T Q x + u^T R u \, dt, \tag{21}
\]
where \(Q = \dot{Q} + \dot{P} B \dot{R} B^T \mathcal{P}\), and \(R\) follows from (16).

The following theorem shows that the hierarchical controller (18) is stabilizing. The proof is given in Section VI.

**Theorem 1:** Under Assumptions 1 and 2, the hierarchical controller (18) guarantees that the MAS (1) is globally asymptotically stable.

**B. Hierarchical RL Algorithm**

We next use (17) Algorithm 3\footnote{Although (17) focuses on minimum-cost variance control, the proposed algorithms can be applied to conventional LQR problems in a specific case, see (17) Remark 1.} to find the decomposed LQR controllers in a model-free way. The matrix \(\mathcal{B}\) is needed for solving the least squares problem (17). For our design \(\mathcal{B}\) can be estimated by estimating \(\mathcal{B}_j\) for each subgroup \(j\) at the very first step of the RL algorithm as shown in (17). Suppose that the cost function is decomposed into \(s\) subgroups, and Assumption 2 is satisfied. Matrices \(G_1\) and \(G_2\) are determined accordingly. By combining the RL algorithm in (17) with our hierarchical control framework, we propose Algorithm 4 as the RL algorithm for heterogeneous MAS.

**Algorithm 1** RL Algorithm for Optimal Control of Heterogeneous MAS via Hierarchical Control

**Input:** \(\hat{Q}_j, j = 1, \ldots, s; G_2 \otimes \dot{Q}, R = \text{diag}\{\mathcal{R}_1, \ldots, \mathcal{R}_s\}\).

**Output:** Optimal controller \(u^*\)

1. Run (17) Algorithm 3\footnote{Although (17) focuses on minimum-cost variance control, the proposed algorithms can be applied to conventional LQR problems in a specific case, see (17) Remark 1.} to solve the LQR problem (13) for each subgroup \(j, j = 1, \ldots, s\), with group dynamics (7). Obtain \(\mathcal{P}_j\) and estimated \(\mathcal{B}_j\) for the \(j\)-th LQR problem.
2. Compute the Moore-Penrose inverse of \(\mathcal{P}_j B_j\) for \(j = 1, \ldots, s\). Then compute \(\mathcal{R}\) by (20).
3. Let \(\mathcal{P} = \text{diag}\{\mathcal{P}_1, \ldots, \mathcal{P}_s\}, \mathcal{B} = \text{diag}\{\mathcal{B}_1, \ldots, \mathcal{B}_s\}\). The hierarchical optimal controller is
\[
u_h = -(R^{-1}B^T \mathcal{P} + \dot{R}B^T \mathcal{P}) x.
\]
Remark 2: Algorithm 1 is presented from the viewpoint of centralized learning. It can also be implemented in a decentralized way under specific conditions. One such condition is as follows. Consider that a coordinator is chosen from the agents in each subgroup. The coordinator collects and transmits state information from and to all agents in that subgroup. Communication between the different subgroups happens through these coordinators. The interaction relationship between the different coordinators is determined by the structure of $R$ (details will be shown in Theorem 2). Through communication, the coordinator in each subgroup has access to $x_i$ and $x_j$ for any adjacent subgroup $i$, as well as to $P_i B_j$ and $P_j B_i$. Based on these information, each coordinator can use the steps of Algorithm 1 to learn the hierarchical optimal controller for subgroup $j$ in a decentralized fashion.

C. Numerical Comparison with Conventional RL

We end this section by citing a numerical example that compares the conventional RL algorithm of [14] with our hierarchical RL algorithm.

Example 1: We consider $G$ as a graph composed of $s$ cliques, each clique has $c$ agents, and will be referred to as a subgroup. The graph between cliques is considered as an unweighted path graph. Moreover, there is only one link between any two connected cliques $i$ and $j$. See Fig. 2 as a demonstration of graph $G$ for $s = c = 3$. We set $G = L, Q = 0.5 I_{sN}$, where $L$ is the Laplacian matrix, $Q = G \otimes I_n + Q$ and $R = I_{mN}$. Then Assumption 2 always holds for arbitrary decomposition. For $n = 4$ and $m = 2, s = 3$, we set $A_i = \begin{pmatrix} -I_2 & I_2 \\ 0 & -I_2 \end{pmatrix}$ and $B_i = \begin{pmatrix} 0 \\ I_2 \end{pmatrix}$, $i = 1, ..., N$. For $n = 8$ and $m = 4, s = 3$, we set $A_i = A_i \otimes I_2$ and $B_i = B_i \otimes I_2$. The component of each agent’s initial state is randomly selected from $\{1, -1, 0\}$.

The computational time and the value of performance index $J$ for each case are shown in Table I. We see that the hierarchical RL algorithm saves a significant amount of time compared with the conventional RL algorithm. In Table I, “SOP” means “suboptimality”, which is defined as $\frac{J_{\text{HRL}} - J_{\text{OPT}}}{J_{\text{HRL}}}$, where $J_{\text{HRL}}$ and $J_{\text{OPT}}$ are the performances of the hierarchical controller and the optimal controller, respectively. We find that as the problem size increases, performance of the hierarchical controller becomes closer to the optimal one. This is because when there are more links in the overall graph, the links contained in $G_2$ become less important in the computation of the overall performance index. For the conventional RL algorithm, the SOP for each case is almost zero.

In Example 1 the agents in each clique are considered a subgroup. There exist other ways to decompose the graph $G$ in Fig. 2. Given the number of subgroups as three, the decomposition strategy used for this example is optimal in terms of the performance of the hierarchical controller. The reason will be shown in the next section. We will revisit this example in Subsection IV-D.

Remark 3: The proposed hierarchical design is also applicable to model-based optimal control. The strategy in that case will be to decompose the Riccati equation into multiple smaller-sized Riccati equations, leading to reduction of computation time.

IV. OPTIMAL DECOMPOSITION FOR MINIMIZING COMMUNICATION AND SUBOPTIMALITY

The decomposition presented in Section III is not unique. Given a MAS, there can be multiple ways to decompose $G$ that may result in different performances and different amounts of control energy required. In this section, we analyze how a decomposition affects the communication graph $G$, and the performance of the resulting hierarchical controller. Based on this analysis, we propose an approach to obtain an optimal decomposition.

A. Relationship Between Decomposition and Communication Graph

Let $\bar{R} \in \mathbb{R}^{mN \times mN}$ be the solution to (19). As $B^T P$ is block-diagonal according to the subgroups, the graph structure among different subgroups is preserved in $\bar{R}$ and $\bar{R} B^T P$. We partition $\bar{R} \bar{I}_m$ into $s^2$ blocks $\bar{R}(i, j) \in \mathbb{R}^{m \times m}$, and partition $G_2 \in \mathbb{R}^{N \times N}$ into $s^2$ blocks $G_2(i, j) \in \mathbb{R}^{N \times N}$, $i, j = 1, ..., s$, respectively, according to the $s$ subgroups. We state the following theorem.

Theorem 2: Consider the MAS (1) decomposed into $s$ subgroups. For any two distinct subgroups $i$ and $j$, if $G_2(i, j) = 0_{N \times N}$, then $\bar{R}(i, j) = 0_{m \times m}$, and $E_c(u_h)$ in (6) contains no edge between the agents in subgroup $i$ and in subgroup $j$, where $u_h$ is the hierarchical controller (18).

This theorem indicates that our hierarchical controller (18) inherits the structure of $G_2$ for different subgroups in the sense that any pair of agents from two different subgroups do not need to interact with each other through the feedback if there is no link between those two subgroups in graph $G$. Let $i \sim j$ and $i \sim j$ denote the adjacent and non-adjacent relationships between any two subgroups $i$ and $j$, respectively. We can show that the number of communication links required for our hierarchical controller is upper bounded by $N^2 - \kappa$, where

$$\kappa = \sum_{i \sim j} N_i N_j$$

is the number of pairs of agents that do not need to communicate with each other.

### Table I

| Dimension | Time (sec) | Performance |
|-----------|------------|-------------|
| s | c | n | m | RL | HRL | OPT | HRL | SOP |
| 3 | 2 | 4 | 2 | 0.57 | 0.07 | 28.76 | 40.57 | 41.08% |
| 3 | 3 | 4 | 2 | 7.04 | 0.08 | 52.75 | 62.24 | 18.00% |
| 3 | 4 | 4 | 2 | 29.69 | 0.24 | 81.03 | 87.57 | 8.08% |
| 4 | 4 | 8 | 4 | > 60 | 9.83 | 198.89 | 209.29 | 5.23% |

Fig. 2. Graph $G$ with 3 cliques, each clique contains 3 agents.
B. Relationship Between Decomposition and Performance

Let $u^*$ and $u_h^*$ be the optimal controllers corresponding to the original cost function $J(x(0), u)$ in (1) and the approximate cost function $J'(x(0), u)$ in (21), respectively. Here $J'(x(0), u)$ is a function of matrix $Q$, which depends on the decomposition strategy. It can be verified that

$$x^T(0)P x(0) = J(x(0), u^*),$$

$$x^T(0)P x(0) = J'(x(0), u_h^*),$$

where $P$ and $\mathcal{P}$ are the solutions to the following two Riccati equations, respectively:

$$PA + A^TP + Q - PBR^{-1}B^TP = 0, \quad (23)$$

$$\mathcal{P}A + A^T\mathcal{P} + Q - \mathcal{P}B(R^{-1} + R)^T\mathcal{P} = 0. \quad (24)$$

The following theorem shows the relationship between the optimal performance for the approximate problem, the optimal performance of the original problem, and the performance error.

**Theorem 3:** Given MAS (1) with initial state $x(0)$, the optimal value of the approximate cost (21), the optimal value of the original cost for our hierarchical controller have the following relationship:

$$J'(x(0), u_h^*) \leq J(x(0), u^*) \leq J(x(0), u_h^*). \quad (25)$$

**Proof:** See Appendix. \hfill $\blacksquare$

It is desirable to design a decomposition strategy such that the corresponding hierarchical controller $u_h^*$ makes $J(x(0), u_h^*)$ close to $J(x(0), u^*)$. To this end, we try to find the optimal decomposition such that

$$\Delta J_h = J(x(0), u_h^*) - J(x(0), u^*) \geq 0$$

is minimized. Since the performance error always depends on the initial state $x(0)$, we propose to analyze the average performance and the expectation $\mathbb{E}(\Delta J_h)$ given a random initial state vector $x(0)$. The following theorem shows an explicit form of $\mathbb{E}(\Delta J_h)$.

**Theorem 4:** Given MAS (1), let $K_h = R^{-1}B^TP$ be the hierarchical control gain matrix, and $K = R^{-1}B^TP$ be the optimal control gain matrix. Suppose that the initial state vector $x(0)$ is a random variable with zero mean, and $\sigma^2I_{nN}$ as its covariance matrix. Then,

$$\mathbb{E}(\Delta J_h) = \sigma^2 \text{tr}(V), \quad (26)$$

where $V$ is the solution of

$$A_h^TV + V A_h + W = 0, \quad (27)$$

where $A_h = A - B(R^{-1} + R)^TP$ and $W = (K_h - K)^T R (K_h - K)$.

Next, we present an upper bound for $\text{tr}(V)$ depending on $\mathcal{P}$ and $G_2$, as stated in the following lemma.

**Lemma 1:** For the MAS given in (1), the following statements hold for the matrices $V$ and $W$ defined in Theorem 4:

(i) If $\bar{Q} > 0$, then

$$\text{tr}(V) \leq \frac{\lambda_{\text{max}}(\mathcal{P}) \text{cond}(\mathcal{P}) \text{tr}(W)}{\lambda_{\text{min}}(\bar{Q})}; \quad (28)$$

(ii) $\text{tr}(W) \leq f_1(\text{tr}(G_2), \lambda_{\text{min}}(\mathcal{P})) + f_2(\text{tr}(G_2), \lambda_{\text{min}}(\mathcal{P})), \quad (29)$

with

$$f_1 = \frac{\text{tr}^2(G_2) \text{tr}^2(\bar{Q})}{\lambda_{\text{min}}(\mathcal{P}) \sigma^2(B)} \lambda_{\text{max}}(R), \quad (30)$$

$$f_2 = \left[\frac{\text{tr}(BR^{-1}B^T) + \sigma^2 \lambda_{\text{max}}(\mathcal{P}) \text{cond}(\mathcal{P})}{\lambda_{\text{min}}(\mathcal{P}) \sigma^2(B)}\right]. \quad (31)$$

where $\sigma(B)$ is the minimum nonzero singular value of $B$. \hfill $\blacksquare$

Lemma 1 implies that if there exists a decomposition such that all other possible decompositions, $tr(G_2)$ and $\lambda_{\text{max}}(\mathcal{P})$ are minimal, and $\lambda_{\text{min}}(\mathcal{P})$ is maximal, implying that $\text{cond}(\mathcal{P})$ is minimal as well, then the upper bound of $\mathbb{E}(\Delta J_h)$ corresponding to this decomposition is also minimal. However, there may not exist a decomposition such that all these indices are optimized simultaneously. Since minimizing $\text{cond}(\mathcal{P})$ is necessary for minimizing $\lambda_{\text{max}}(\mathcal{P})$ and maximizing $\lambda_{\text{min}}(\mathcal{P})$, we only consider $tr(G_2)$ and $\text{cond}(\mathcal{P})$ as the two most important indices for evaluating $\mathbb{E}(\Delta J_h)$.

In conclusion, when $\bar{Q} > 0$, we can view $\text{tr}(G_2)$ and $\text{cond}(\mathcal{P})$ as two of the most important factors affecting the performance of our hierarchical controller corresponding to any given decomposition.

**Remark 4:** From [15], $\mathcal{P}$ depends not only on $\bar{Q}$, but also on $A$, $B$, and $R$. For a heterogeneous MAS, if different agents have largely different dynamics and $R_i$ differs largely for different $i$, then these differences will strongly influence the resulting $\mathcal{P}$ for different decomposition strategies. Since the dynamics of the agents are unknown, we cannot utilize this information for analyzing the relationship between the decomposition and the performance of the hierarchical controller. The performance is mainly determined by $\text{tr}(G_2)$ and $\text{cond}(\mathcal{P})$ only when the differences between the agent dynamics are small, and $R_i$ for different $i$ is similar.

C. Approaches for Optimizing Decomposition

Subsection IV-A shows that maximizing $\kappa = \sum_{i,j} N_i N_j$ helps reduce the number of required communication links in the hierarchical controller $u_h$. On the other hand, Subsection IV-B shows that minimizing $\text{tr}(G_2)$ and $\text{cond}(\mathcal{P})$ help optimize the performance of the closed-loop system. The index cond($\mathcal{P}$) is always influenced by $A$, $B$ and $R$, so it is impractical to optimize cond($\mathcal{P}$) by choosing the decomposition strategy without explicit knowledge on $A$ and $B$. However, minimizing tr($G_2$) is independent of the system dynamics, thus is tractable. Therefore, we seek to maximize $\kappa$ for minimizing the number of communication links, and minimize tr($G_2$) for optimizing the performance of our hierarchical controller.

Given the desired number of subgroups, minimizing tr($G_2$) is equivalent to minimizing the total number of edges crossing any two different subgroups, which is actually a minimum s-cut problem. In [44], an algorithm is proposed to solve the minimum s-cut problem in polynomial time for a specified s. Maximizing $\kappa$, in comparison, is more complicated because...
the structure of the graph is important in computing $\kappa$. For example, in the first example shown in Fig. 1, $\kappa = 2$ and $\text{tr}(G_2) = 6$, and the interaction between subgroup 2 and subgroup 3 is not required in the hierarchical controller. If we use another decomposition, say, $V_1 = \{1\}$, $V_2 = \{2,3\}$ and $V_3 = \{4,5\}$, then $\kappa = 0$ and $\text{tr}(G_2) = 6$. Although the number of edges crossing the subgroups remains the same, the resulting hierarchical control may require communications between two different subgroups.

Given $s$ as the desired number of subgroups, we next find the decomposition for maximizing $\kappa$. Let $\eta_i \in \mathbb{R}^N$ be a vector where each component is either 0 or 1, $i = 1, ..., s$. We use $\eta_i$ to determine the members of subgroup $i$. Let $\eta_i(k)$ be the $k$th element of $\eta_i$ and define

$$\eta_i(k) = \begin{cases} 1, & \text{if } k \in V_i; \\ 0, & \text{otherwise}, \end{cases}$$

where $V_i$ is the set of agents in the $i$th subgroup. Then the number of agents in subgroup $i$ is $N_i = |V_i| = \eta_i^\top 1_N$. Given $s$ as the number of subgroups, we use matrix $E = (\eta_1, ..., \eta_s) \in \mathbb{R}^{N \times s}$ to denote a decomposition. Accordingly, $G_1(E)$ and $G_2(E)$ become the two decomposed matrices in $G$ following this decomposition $E$, i.e.,

$$G = G_1(E) + G_2(E).$$

Note that $G_1(E)$ may not be block-diagonal but can be transformed to be so by relabelling the agents.

Now we formulate a MIQP problem for seeking the decomposition maximizing $\kappa$. From the definition of $E$, we have $\kappa = \sum_{i=1}^{s} N_i N_j = \sum_{i=1}^{s} \sum_{j \neq i} \eta_i^\top 1_N \eta_j^\top 1_N$. Let $\tilde{G} = |G|$ where $\tilde{G}_{ij} = |G_{ij}|$. For any two subgroups $i$ and $j$, the total amount of interaction weights between them is $\langle \eta_i^\top \tilde{G} \eta_j \rangle = \eta_i^\top \tilde{G} \eta_j$. Therefore, $i \sim j$ if and only if $\eta_i^\top \tilde{G} \eta_j = 0$. Let $l_{ij} \leq 1 - \eta_i^\top \tilde{G} \eta_j / T$ and $\tau_{ij} = \eta_i^\top \tilde{G} \eta_j / T$, where $T$ is an integer greater than or equal to the sum of the weights of $G$. Then $\tau_{ij} = \tilde{G} \eta_j / T = 0$ if $i \sim j$, and $\tau_{ij} = \tilde{G} \eta_j / T \leq 1$ otherwise.

If we consider $l_{ij}$ as a binary variable, then the maximum value of $l_{ij}$ is 1 if $i \sim j$, and is 0 otherwise. Therefore, the decomposition maximizing $\kappa$ is given by the solution to the following MIQP problem:

$$\begin{align*}
\min_{\eta_1, ..., \eta_s} & - \sum_{i=1}^{s} \sum_{j=1}^{s} \eta_i \eta_j^\top \tau_{ij} \\
\text{s.t.} & \sum_{i=1}^{s} \eta_i = 1_N, \quad \eta_i^\top 1_N \geq 1, \\
& \tau_{ij} = \eta_i^\top \tilde{G} \eta_j, \quad l_{ij} \leq 1 - \eta_i^\top \tilde{G} \eta_j / T, \\
& l_{ij}, \eta_i(k) \in \{0, 1\}, \quad i, j = 1, ..., s.
\end{align*}$$

The MIQP problem (33) can be solved by branch and bound algorithms. Commercial software such as Matlab, Lindo and Gurobi can be used for this purpose, as will be shown in our simulations. Let $\kappa^*(s)$ be the maximum $\kappa$, given $s$ as the number of subgroups, and $z_0$ as the number of zero elements in $G$. We next present two properties for $\kappa^*(s)$.

**Theorem 5**: Given matrix $G \in \mathbb{R}^{N \times N}$ in (4), $\kappa^*(s)$ is non-decreasing in $s$ and $\kappa^*(s) \leq z_0$. ■

D. Numerical Verifications

As we discussed in Subsection IV-B when $Q > 0$, and $\text{cond}(P)$ is similar for different decompositions, $\text{tr}(G_2)$ is the key factor determining the performance of the hierarchical controller. In what follows, we will present two numerical examples where $\text{cond}(P)$ is similar for different decompositions.

**Example 2**: We reconsider the example with graph $G$ in Fig. 2 matrices $Q$, $R$ and agents’ dynamics are still the same as those considered in Example 1. By generating the initial states of agents 1000 times such that each time every component of agents’ states is a random number from the normal distribution with mean 0 and variance 0.5, the average performance values for different decompositions are shown in Table II. By solving a mixed integer linear program (MILP) formulation for the minimum s-cut problem, we obtain the second decomposition, which maximizes $\kappa$ and induces a hierarchical controller with the best performance. By solving MIQP (33) with Gurobi via Matlab (takes 7.8843s), we obtain the third decomposition, which maximizes $\kappa$ and requires fewest communication links in the controller.

| Decomposition | $\kappa$ | $\text{tr}(G_2)$ | $\text{cond}(P)$ | $J$ | $n_c$ | SOP | $\%$ |
|---------------|---------|-----------------|-----------------|-----|------|-----|------|
| $\{1,2,3\},\{4,5\},\{6,7\}$ | 4 | 8 | 16.4 | 15.9 | 32 | 75.3% |
| $\{1,2,3\},\{4,5\},\{6,7\}$ | 4 | 8 | 16.4 | 15.9 | 32 | 75.3% |
| $\{1,2,3\},\{4,5\},\{6,7\}$ | 4 | 8 | 16.4 | 15.9 | 32 | 75.3% |
| $\{1,2,3\},\{4\},\{5,6,7\}$ | 15 | 6 | 17.0 | 15.8 | 21 | 22.15% |
| Undecomposed | n/a | n/a | n/a | 12.9 | 36 | 0 |

**Remark 5**: When $\hat{Q}$ is well-conditioned (i.e., when $\text{cond}(\hat{Q})$ is small), and $\lambda_{\max}(G)$ is not large, $\hat{Q}$ will be well-conditioned because $\text{cond}(\hat{Q}) \leq (\lambda_{\max}(G) + \lambda_{\max}(\hat{Q}))/\lambda_{\min}(\hat{Q})$. Moreover, when all agents have the same dynamics and equal values of $R_i$, $P$ will be well-conditioned as well. In that case, $\text{tr}(G_2)$ is the most important factor determining the performance of the hierarchical controller, as can be seen clearly from the above examples. On the other hand, when $\hat{Q}$ is ill-conditioned, different decompositions may cause largely different $\text{cond}(\hat{Q})$ and $\text{cond}(P)$.

V. APPLICATION TO MULTI-AGENT FORMATION MANEUVER CONTROL

In this section, we show an application of the proposed hierarchical RL control framework in a multi-agent formation control problem.
A. Problem Formulation

Consider a group of agents, denoted by the vertex set $\mathcal{V}$, traveling between multiple waypoints in a two-dimensional plane. The set $\mathcal{V}$ is categorized into two subsets, $\mathcal{V} = \mathcal{L} \cup \mathcal{F}$, where $\mathcal{L}$ is the set of leaders tracking specific positions at each waypoint, and $\mathcal{F}$ is the set of followers adjusting their positions according to measured information from other agents.

1) Agent Dynamics: Each agent is considered to have the following dynamics

$$m_i\ddot{q}_i + c_i\dot{q}_i = u_i, \quad i = 1, \ldots, N$$  (34)

where $q_i \in \mathbb{R}^2$, $\dot{q}_i \in \mathbb{R}^2$, $m_i > 0$, $c_i > 0$ and $u_i \in \mathbb{R}^2$ are the position, the velocity, the mass, the damping coefficient and the control input of agent $i$, respectively. The parameters $m_i$ and $c_i$ are considered to be unknown.

2) Control Objective: The general control objective is to design a control law without knowing $m_i$ and $c_i$ for each agent $i$, such that the agents maintain a desired formation shape at each waypoint, achieve a flocking behavior during the travel between each of two neighboring waypoints, and take minimum control efforts in transformation between two formation shapes.

3) Formation Graph and Communication Graph: We consider the formation graph $\mathcal{G}_f = (\mathcal{E}_f, \mathcal{V})$ and the communication graph $\mathcal{G}_c = (\mathcal{E}_c, \mathcal{V})$ as two different graphs. Graph $\mathcal{G}_f$, together with the target position $h = (h_1^T, \ldots, h_N^T)^T \in \mathbb{R}^{2N}$, form a framework $(\mathcal{G}_f, h)$ describing the desired formation. Here $h_i$ is the target position of agent $i$ at the next waypoint. The edges determined by $\mathcal{E}_f$ specify the agent pairs with constrained relative positions, while the edges determined by $\mathcal{E}_c$ specify those agent pairs who communicate with each other. Usually $\mathcal{E}_c$ is determined by the wireless communication ranges of the agents. Therefore, staying cohesive will require more number of communication links.

4) LQR Problem Formulation: Let $x_i = (q_i^T - h_i^T, \dot{q}_i^T)^T$ for $i = 1, \ldots, N$. Then the system equation can be rewritten as

$$\dot{x}_i = A_ix_i + B_iu_i$$  (35)

where $A_i = \left( \begin{array}{cc} 0 & I_2 \\ 0 & -\frac{c_i}{m_i}I_2 \end{array} \right)$, and $B_i = \left( \begin{array}{c} 0 \\ \frac{1}{m_i}I_2 \end{array} \right)$. Let $S_1 = (I_2, 0_{2 \times 2})$, $S_2 = (0_{2 \times 2}, I_2)$, then $q_i - h_i = S_1x_i$, and $\dot{q}_i = S_2x_i$.

In the literature, asymptotic convergence of formation stabilization has been widely studied [29], [31]. However, the performance of the group trajectory during transience is usually not guaranteed. To capture an optimal performance of the trajectory of the agents during travel between the waypoints, we make the whole group minimize the following performance index:

$$J_1 = \int_0^\infty \sum_{(i,j) \in \mathcal{E}_f} ||q_i - q_j - (h_i - h_j)||^2 + \sum_{i \in \mathcal{L}} ||q_i - h_i||^2 dt$$

$$= \int_0^\infty x^T(L \otimes S_1^T S_1)x + x^T(\Lambda \otimes I_2)x dt$$  (36)

where $L \in \mathbb{R}^{N \times N}$ is the Laplacian matrix corresponding to the formation graph $\mathcal{G}_f$, $\Lambda = \text{diag}\{\Lambda_1, \ldots, \Lambda_N\}$, $\Lambda_i = 1$ if $i \in \mathcal{L}$ and $\Lambda_i = 0$ otherwise.

A flocking behavior requires agents to stay cohesive and have a common velocity, therefore we additionally define the following performance index:

$$J_2 = \int_0^\infty \dot{q}^T(L \otimes I_2)\ddot{q}dt = \int_0^\infty x^T(L \otimes S_2^T S_2)x dt.$$  (37)

The overall goal is to minimize

$$J = J_1 + J_2 = \int_0^\infty [x^T((L + \Lambda) \otimes I_4)x + u^T u] dt$$  (38)

subject to system model (35).

Remark 6: Different from formation control problems, where the steady states of all agents depend on their initial states, in the maneuver control problem each agent has a specific target position at each waypoint. In this case, it is easy for each agent to obtain its and others’ target steady states by either (i) using available target relative positions from its neighbors and communicating with them, or (ii) achieving the target position through a centralized task assignment.

B. Feasible Decompositions for RL-Based Hierarchical Control

To use our hierarchical control design, Assumptions 1 and 2 should be satisfied. Let $A = \text{diag}\{A_1, \ldots, A_N\}$ and $B = \text{diag}\{B_1, \ldots, B_N\}$, where $A_i$ and $B_i$ are shown in (35). Since $m_i > 0$ and $c_i > 0$ for each agent $i$, $(A, B)$ must be controllable. Moreover, when graph $\mathcal{G}_f$ is connected and there exists at least one leader in the whole group, it holds that $Q = (L + \Lambda) \otimes I_4 > 0$, implying that $(Q^{1/2}, A)$ is observable. Then Assumption 1 is satisfied. To ensure validity of Assumption 2 the decomposition must satisfy the following result.

Theorem 6: A decomposition makes Assumption 2 satisfied if and only if each subgroup contains at least one leader and its formation graph is connected.

An immediate consequence of Theorem 6 is that the number of subgroups cannot exceed the number of leaders in the MAS. To find the decomposition maximizing $\kappa$, we modify the MIP [33] to include the following two constraints:

(i) $\xi \eta_i \geq 1$ for $i = 1, \ldots, s$, where let $\xi \in \mathbb{R}^N$ with $\xi(i) = 1$ if $i \in \mathcal{L}$ and $\xi(i) = 0$ otherwise;

(ii) $G_{kk} - \sum_{i=1}^s \eta_i^T g_k e_k^T \eta_i \geq \epsilon (\sum_{i=1}^s \eta_i^T 1_N e_k^T \eta_i - 1)/N$, $k = 1, \ldots, N$, where $g_k \in \mathbb{R}^N$ is the $k$-th column of $G$, and $\epsilon$ is the minimum nonzero entry of $G$.

The constraint (i) implies that each subgroup contains at least one leader, while constraint (ii) means that each agent $k$ either forms a subgroup $(\sum_{i=1}^s \eta_i^T 1_N e_k^T \eta_i = 1)$ or has a nonempty neighbor set in the subgroup it belongs to.

C. Numerical Experiments

We consider 12 agents governed by (35) with the formation graph $\mathcal{G}_f$ shown in Fig. 3. The black nodes are leaders, and the other nodes are followers. The mission of these agents is to transform their formation shape from Fig. 3 (a) to Fig. 3 (b) before entering into a narrow space as shown towards the right in the figure. The leader set is $\mathcal{L} = \{1, 8, 12\}$. 
Set mass $m_i = i/2$ and damping parameter $c_i = i/5$ for $i = 1, ..., N$. Fig. 4 and Fig. 5 show the solution to the optimal control problem formulated in Subsection V-A and the solution provided by the formation stabilization law in [29], respectively. We firstly find the optimal control law via the conventional model-based approach in [39]. By implementing the optimal control law, the value of the performance index is $J = 1112.64$, the control effort $J_u = \int_0^\infty u^t u dt = 359.11$, the number of communication links is 66, which implies that the communication graph $\mathcal{G}_c = (\mathcal{V}, \mathcal{E}_c)$ is fully connected. Under the same communication graph $\mathcal{G}_c$, by using the formation stabilization law in [29] we get

$$u_i = -\sum_{(i,j) \in \mathcal{E}_c} (q_i - q_j - (h_i - h_j)) - k_i(q_i - h_i)$$

(39)

where $k_i = 1$ if $i \in \mathcal{L}$ and $k = 0$ otherwise. The performance indices are computed as $J = 2011.21$ and $J_u = 945.56$. From Fig. 4 and Fig. 5 we observe that compared with the stabilization law (39), by implementing the optimal controller, the trajectories of agents are much shorter. Therefore a significant amount of control effort is saved by using an optimal controller.

Next, we compute the optimal controller in the model-free scenario via the conventional RL algorithm and our hierarchical RL algorithm, respectively. By implementing the RL algorithm in [14], the obtained optimal controller is almost the same as the one obtained in the previous model-based case. The learning time is 29.74s.

Now, we apply our proposed hierarchical approximation. For validity of Assumption 2 according to Theorem 3 we decompose the group into $\kappa = 3$ subgroups such that each subgroup has a connected formation graph and contains at least one leader. In Fig. 6 the decomposition strategies maximizing $\kappa$ and minimizing $\text{tr}(\mathcal{G}_2)$ are shown, respectively, where the nodes with the same color belong to the same subgroup. The first decomposition is obtained by solving (33) with additional constraints presented in Subsection V-B which costs 22.7734s for Gurobi via Matlab. Fig. 7 shows the trajectories of the agents using the hierarchical control law with the decomposition maximizing $\kappa$. The second decomposition is obtained by solving a MILP formulation for the minimum $s$-cut problem with additional constraints in Subsection V-B.

Suppose we choose the first decomposition strategy in Table III. During both the RL algorithm implementation and the optimal controller implementation, communications between agents $\{1, ..., 6\}$ and $\{10, 11, 12\}$ are not required. The communications between the rest of agents can be achieved through 3 coordinators in these 3 subgroups. The RL learning process can be done distributively as described in Remark 2.

We implement Algorithm 1 based on all the three decompositions in Table III. The learning time of Algorithm 1 and the performance of the hierarchical controller based on each decomposition are listed in Table III. For simplicity, each decomposition is done according to the order of the agents. As an example, for the first decomposition in Table III the indices of agents in the three subgroups are $\{1, ..., 6\}$, $\{7, 8, 9\}$ and $\{10, 11, 12\}$. It is seen from Table III that the first decomposition maximizes $\kappa$, and induces the fewest communication links in the hierarchical controller. The second decomposition minimizes $\text{tr}(\mathcal{G}_2)$, but does not induce the best performance. Instead, the first decomposition, which induces the minimum condition numbers on $\mathcal{P}$ and $Q$, yields a better performance. This is because $\text{cond}(Q) = \text{cond}(\Lambda) = \infty$, which implies that $\tilde{Q}$ is ill-conditioned. In this case, different decompositions will lead to largely different $\text{cond}(\mathcal{P})$ and $\text{cond}(Q)$, and thus $\text{tr}(\mathcal{G}_2)$ will be a less important metric to characterize the suboptimality.

VI. CONCLUSION

We presented a model-free hierarchical RL algorithm for optimal control of linear MAS with heterogeneous agents, based
From LaSalle’s Invariance Principle, we know that the system is globally asymptotically stable. From (18), the hierarchical controller of the $i$-th subgroup is the trivial solution $\hat{x}_i(0) = 0$, i.e., $\hat{Q} \leq 0$. As a result, the only solution that stays in $E$ is the trivial solution $x = 0$. Then we conclude that the only solution that stays in $E$ is globally asymptotically stable.

Next, we partition $\Phi$ into $s^2$ blocks $\Phi_{ij} \in \mathbb{R}^{n_i \times n_j}$, $i,j = 1,...,s$, according to the $s$ subgroups. For different $i$ and $j$ such that $G_2(i,j) = 0_{N_i \times N_j}$, we have $\Phi_{ij} = 0_{n_i \times n_j}$. As a result,

$$\hat{R}(i,j) = \Xi_i \Phi_{ij} \Xi_j^\top = 0_{m_j \times m_i}.$$  

From (18), the hierarchical controller of the $j$-th subgroup is in the following form

$$u_j = -\hat{R}_j^{-1} \Pi_j \hat{x}_j - \sum_{i=1}^{s} \hat{R}(i,j) \Pi_i \hat{x}_i,$$  

which implies that communication between subgroup $i$ and subgroup $j$ is not required if $\hat{R}(i,j) = 0_{m_j \times m_i}$.

**Proof of Theorem 1**. The second inequality $J(x(0), u^*) \leq J(x(0), u_0^*)$ is valid because $u^*$ is the optimal controller corresponding to $J(x(0), u)$. Next, we prove the first inequality. From (15), we observe that $P$ is also the solution to

$$P A + A^\top P + Q - PBB^{-1}B^\top P = 0.$$  

Note that $Q - \hat{Q} = G_2 \otimes \hat{Q} \geq 0$, i.e., $\hat{Q} \leq Q$. Using (23) and (41) Lemma 3, we obtain $P \leq \hat{P}$. It follows that $J(x(0), u_0^*) \leq J(x(0), u^*)$.

**Proof of Theorem 2**. We first study $\mathbb{E}(J(x(0), u_0^*))$. We can write

$$J(x(0), u_0^*) = \int_0^\infty x^\top Q x dt + \int_0^\infty u_0^\top R u_0^* dt,$$

where $U = e^{(A-BK_0)t}x(0)$. It follows that

$$\mathbb{E}(J(x(0), u_0^*)) = \text{tr}(U \mathbb{E}(x(0) x^\top(0))) = \sigma^2 \text{tr}(U),$$

where $U$ is the solution of

$$A_s^\top U + UA_s + \frac{Q + K_h R K_h}{2} = 0,$$

with $A_s = A - BK_h$.

Next, we analyze $\mathbb{E}(J(x(0), u^*))$. It can be easily verified that

$$\mathbb{E}(J(x(0), u^*)) = \mathbb{E}(x^\top(0) P x(0)) = \sigma^2 \text{tr}(P),$$

where $P$ is the solution of (23). Let $V = U - \hat{P}$. Then

$$\mathbb{E}(\Delta J_0) = \mathbb{E}(J(x(0), u_0^*)) - \mathbb{E}(J(x(0), u^*)) = \sigma^2 \text{tr}(V).$$

Note that (23) can be rewritten as

$$PA_s + A_s^\top P + Q + K^\top R K_h + \frac{K_h^\top R K_h}{2} = 0.$$  

The subtraction of (51) and (52) yields (27).

**Proof of Lemma 7**. Using (15), the positive definiteness of $\hat{Q}$ completes the proof.

**TABLE III**

| Decomposition | Performance Indices |
|---------------|---------------------|
| $N_1$ | $N_2$ | $N_3$ | $\kappa$ | $tr(G_2)$ | $\text{cond}(P)$ | $\text{cond}(\hat{Q})$ | $J$ | $J_0$ | $n_c$ | Time(sec) | SOP |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | 3 | 3 | 18 | 12 | 234.43 | 46.13 | 1129.26 | 356.29 | 48 | 0.67 | 1.49% |
| 1 | 10 | 1 | 0 | 2 | 309.26 | 38.75 | 1228.51 | 353.08 | 65 | 9.78 | 10.41% |
| 7 | 2 | 3 | 0 | 12 | 265.56 | 55.35 | 1130.10 | 359.59 | 66 | 1.40 | 1.57% |

Fig. 7. The trajectory for maneuver control via hierarchical approximation with the decomposition maximizing $\kappa$. on a decomposition approach to reduce learning time. The derived controller is suboptimal but has a specific structure. Two indices of the hierarchical controller, i.e., required interagent communication links and performance of the MAS, are analyzed and shown to be dependent on the decomposition strategy. Optimizing these two indices is formulated as a MIQP and a minimum $s$-cut problem, respectively. The hierarchical controller is applied to a formation maneuver control problem. Simulation experiments illustrate the effectiveness of our hierarchical RL control strategies.
of $\tilde{Q}$, and [43] Corollary 4.5.11], the following holds:

\[-\lambda_{\max}(PA\Sigma P^{-1} + A_{s}^T) = -\lambda_{\max}(PA\Sigma P^{-1} - \tilde{P}B\tilde{R}^{-1}B^T + A^T - \tilde{P}B\tilde{R}^{-1}B^T) = -\lambda_{\max}(-\tilde{P}B\tilde{R}^{-1}B^T - Q\tilde{P}^{-1}) = -\lambda_{\min}(\tilde{P}^{-1/2}B\tilde{R}^{-1}B^T\tilde{P}^{-1/2} - \tilde{P}^{-1/2}Q\tilde{P}^{-1/2}) = \min(\tilde{P}^{1/2}B\tilde{R}^{-1}B^T\tilde{P}^{1/2} + \tilde{P}^{-1/2}Q\tilde{P}^{-1/2}) = \min(\tilde{P}^{1/2}B\tilde{R}^{-1}B^T\tilde{P}^{1/2} + \tilde{P}^{-1/2}Q\tilde{P}^{-1/2}) \geq \min(\tilde{P})\lambda_{\min}(B\tilde{R}^{-1}B^T) + \min(\tilde{P})\max(\tilde{Q})/\lambda_{\max}(\tilde{P}) > 0.\]

(53)

In most applications, usually $\tilde{B}$ is not square, implying that $\lambda_{\min}(B\tilde{R}^{-1}B^T) = 0$. Therefore, we further omit the term associated with $\lambda_{\min}(B\tilde{R}^{-1}B^T)$, i.e.,

\[-\lambda_{\max}(PA\Sigma P^{-1} + A_{s}^T) \geq \min(\tilde{Q})/\lambda_{\max}(\tilde{P}).\]

(54)

From [42] Corollary 3.2, it holds that:

\[\text{tr}(V) \leq -\lambda_{\max}(P)\text{tr}(P^{-1}W)/\lambda_{\max}(PA\Sigma P^{-1} + A_{s}^T).\]

(55)

Reusing [43] Corollary 4.5.11], we have

\[\text{tr}(P^{-1}W) = \text{tr}(P^{-1/2}WP^{-1/2}) \leq \lambda_{\max}(P^{-1/2})\text{tr}(W)/\lambda_{\min}(P).\]

(56)

Combining (54), (55) and (56), the bound on $\text{tr}(U)$ stated in (28) is obtained.

(ii) It can be verified that

\[W = P\tilde{M}P + PM\tilde{P} + \Delta \tilde{P}M\Delta P - PM\tilde{P},\]

(57)

where $M = B\tilde{R}^{-1}B^T$, $\tilde{M} = B\tilde{R}^T\tilde{B}$, $\Delta P = P - \tilde{P}$, and $\Delta P = \tilde{P}$. We have shown in Theorem 3 that $P \leq \tilde{P}$. It follows that $\text{tr}(P^2 - \tilde{P}^2) = \text{tr}((P + \tilde{P})P^{-1/2}(P - \tilde{P})P^{-1/2}) \geq 0$. As a result, $\text{tr}(PM\tilde{P} - \tilde{P}M) = \text{tr}(P - P^2)M = -\text{tr}(\tilde{M}^{-1/2}(P - \tilde{P})\tilde{M}^{-1/2} \leq -\lambda_{\min}(\tilde{M})\text{tr}(P^2 - \tilde{P}^2) \leq 0.

This implies that $\text{tr}(W) \leq \text{tr}(PM\tilde{P}) + \text{tr}(\Delta \tilde{P}M\Delta P)$. To prove the statement (ii), we will prove $\text{tr}(PM\tilde{P}) \leq f_1$ and $\text{tr}(\Delta \tilde{P}M\Delta P) \leq f_2$ successively.

For the first inequality, by [43] Corollary 4.5.11], the following holds:

\[\text{tr}(PBR\tilde{R}B^T) \leq \lambda_{\max}(PB)\lambda_{\max}(PB)^+ \times \text{tr}((G_2 \otimes \tilde{Q})(PB)^T + R(PB) + (G_2 \otimes \tilde{Q})) = \text{tr}(R^{1/2}(PB)^+ (G_2 \otimes \tilde{Q})^2 (PB)^T R^{1/2}) \leq \text{tr}^2(G_2 \otimes \tilde{Q})\lambda_{\max}[(PB)^+ P + R(PB)^+] = \text{tr}^2(G_2 \otimes \tilde{Q})\lambda_{\max}(PBR\tilde{R}B^T),\]

(58)

where $S^2 = (PB)^+ (PB)^+ T$, i.e., $S = TA^{1/2}T^T$ in which $T$ and $\Lambda$ follow from the eigen-decomposition of $(PB)^+ (PB)^+ T$ satisfying $TA^{1/2}T = (PB)^+ (PB)^+ T$. Since $S$ is square and symmetric, we can apply [43] Corollary 4.5.11] to analyze the eigenvalues of $SRS$.

The singular values of $S$ are given by the diagonal entries of $\Lambda^{1/2}$. The minimum singular value is zero while the maximum singular value is less than $\sigma_{\max}((PB)^+)$). Therefore, we have

\[\lambda_{\max}(SRS) \leq \sigma_{\max}^2((PB)^+)\lambda_{\max}(R).\]

(59)

Using the fact that $\sigma_{\max}((PB)^+) = 1/\sigma_1(PB)$, where $\sigma_1(PB)$ is the minimum nonzero singular value of $PB$, we next look for a lower bound on $\sigma_1(PB)$. We have

\[\sigma_1(PB) = \sigma_1(B^TP)\]

(60)

Thus, we have

\[\lambda_{\max}(SRS) \leq \lambda_{\max}(P)\sigma_1(B)\lambda_{\max}(P)\sigma_{\max}(B).\]

(61)

Together with (58), we have $\text{tr}(PM\tilde{P}) \leq f_1$.

In the following derivations for $f_2$, some steps that are similar to the approach stated above will be omitted for brevity. We use the following two inequalities:

\[\text{tr}(\Delta PBR\tilde{R}B^T) \leq \lambda_{\max}(\Delta P)\lambda_{\max}(PB)\lambda_{\max}(PB)^+,\]

(62)

\[\text{tr}(\Delta PBR^2) \leq \lambda_{\max}(\Delta P)\sigma_{\max}(PB)\sigma_{\max}(PB)^+ \text{tr}(R) \leq \lambda_{\max}(\Delta P)\sigma_{\max}(PB)\sigma_{\max}(PB)^+ \text{tr}(G_2 \otimes \tilde{Q})\]

(63)

Combining the two inequalities, we get $\text{tr}(\Delta P\tilde{M}R\Delta P) \leq f_2$.

Proof of Theorem 3 Without loss of generality, given $s$ as the number of subgroups, and the decomposition $E = (\eta_1, \ldots, \eta_s)$ corresponding to the maximum $\kappa = \kappa^s(s) = \sum_{i=1}^{s} N_i N_i$. Next we show that for $s + 1$, there exists a decomposition $E = (\eta_1, \ldots, \eta_{s+1})$ such that the corresponding $\kappa^\prime \geq \kappa^s(s)$. For any subgroup with at least two agents, if we decompose it into two subgroups $k$ and $l$, the resulting $\kappa$ will be $\kappa^\prime < \kappa^s(s) + N_k N_l > \kappa^s(s)$ if $k \approx l$, and $\kappa^\prime < \kappa^s(s)$ otherwise. This shows the nondecreasing property of $\kappa(s)$. Since $s \leq N$, and $\kappa(N) = z_0$, we conclude that $\kappa$ is bounded by $z_0$.

Proof of Theorem 6 For each subgroup $j$, the performance index can be written as $J = \int_0^{\infty} \sum_{j=1}^{s} (L_j^+ + \kappa^s(s) L_j) + u^T u) dt$, where $L_j$ is the Laplacian matrix corresponding to the subgraph $G_j^f$ of $G_f$ involving agents in $V_j$, $\kappa$ is a diagonal matrix.

Sufficiency: When there is a leader in subgroup $j$, there is at least one positive value on diagonal of $\kappa$. Since $G_j^f$ is connected, using [46] Lemma 3, we have $Q_j = (L_j^+ + \kappa) I_j > 0$. As a result, $(Q_j^{1/2}, A_j)$ is observable.

Necessity: Suppose that no leaders exist in subgroup $j$ or $G_j^f$ is not connected. Then $Q_j = L_j^+ I_4$ and $L_j$ has at least one zero eigenvalue. Let $\zeta$ be the eigenvector associated with eigenvalue $0$ of $L_j$. We observe that $\zeta$ and $A_j$ have a common eigenvector corresponding to eigenvalue $0$, which is $\zeta \in \{1, 1, 0, 0\}^T$. Therefore, $(Q_j^{1/2}, A_j)$ can never be detectable.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*.
