

# Bi-CL: A Reinforcement Learning Framework for Robots Coordination Through Bi-level Optimization

Zeichen Hu, Daigo Shishika, Xuesu Xiao, and Xuan Wang

**Abstract**—In multi-robot systems, achieving coordinated missions remains a significant challenge due to the coupled nature of coordination behaviors and the lack of global information for individual robots. To mitigate these challenges, this paper introduces a novel approach, Bi-level Coordination Learning (Bi-CL), that leverages a bi-level optimization structure within a centralized training and decentralized execution paradigm. Our bi-level reformulation decomposes the original problem into a reinforcement learning level with reduced action space, and an imitation learning level that gains demonstrations from a global optimizer. Both levels contribute to improved learning efficiency and scalability. We note that robots’ incomplete information leads to mismatches between the two levels of learning models. To address this, Bi-CL further integrates an alignment penalty mechanism, aiming to minimize the discrepancy between the two levels without degrading their training efficiency. We introduce a running example to conceptualize the problem formulation and apply Bi-CL to two variations of this example: route-based and graph-based scenarios. Simulation results demonstrate that Bi-CL can learn more efficiently and achieve comparable performance with traditional multi-agent reinforcement learning baselines for multi-robot coordination.

## I. INTRODUCTION

Multi-robot systems have extensive applications in various engineering fields, but their deployment relies on scalable algorithms that enable robots to make coordinated and sequential decisions using local observation [1]. To this end, Centralized Training with Decentralized Execution (CTDE) [2] emerges as a promising approach, offering a balanced framework for coordinating multiple robots through centralized learning processes that guide decentralized operational decisions. However, the effectiveness of CTDE faces significant challenges as the dimensionality of action spaces expands and each robot only has local observation of the system. These challenges complicate the training process and potentially impedes practical deployment.

In this context, we note that multi-robot cooperative missions often exhibit inherent hierarchical structures, allowing them to be decomposed into high-level and low-level tasks [3]. For instance, rescue missions utilize teams of robots to search large or hazardous areas [4]–[7], where a high-level planning task may focus on area coverage, and a low-level task may address navigation and obstacle avoidance. A similar scenario arises from warehouse automation [8], where high-level tasks involve managing robot fleets for picking, packing, and sorting, and low-level tasks ensure precise and safe navigation in densely populated environments. Nevertheless, in most cases, the decomposed

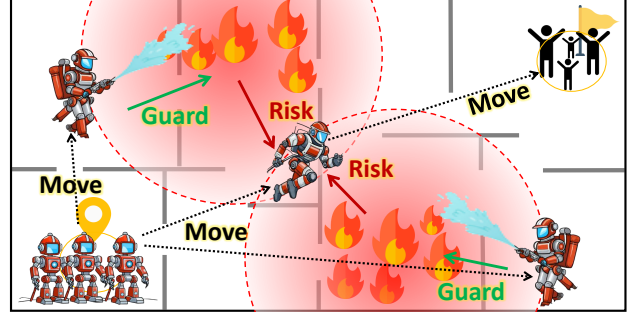


Fig. 1: A running example for a firefighting scenario. Robots can simultaneously perform two actions: move (to where) and guard (which adversary). Team reward depends on the risk of fire, which is a coupled function of both actions.

problems are internally coupled by state dependence, which can not be solved independently. Bi-level optimization [9] presents a solution to this issue, with the capability of enhancing learning efficiency and stability while maintaining the explicit connections between the two levels of problems. However, despite the abundance of research on addressing such coupling in static optimization problems for both single-agent [10] or multi-agent [11] cases, there remains a scarcity of studies tailored for multi-agent reinforcement learning (MARL) under CTDE with robots’ local observation. This gap motivates our research to develop new approaches that leverage the benefits of bi-level optimization for reinforcement learning applications in multi-robot systems, thereby enhancing their deployment efficiency across a variety of complex and dynamic environments.

**Statement of contribution:** The contributions of this paper include (i) the formulation of a bi-level Decentralized Markov Decision Process (Dec-MDP) for multi-robot coordination; (ii) a Bi-level Coordination Learning (Bi-CL) framework with a novel alignment mechanism to integrate multi-agent reinforcement learning and imitation learning under the CTDE scheme; (iii) simulated experiments that verify the effectiveness of the proposed Bi-CL algorithm and a comparison with traditional MARL algorithms to solve multi-robot coordination tasks.

**Running Example:** As visualized in Fig. 1, we introduce an example throughout this paper to conceptualize and verify the proposed bi-level approach for learning multi-robot coordination. Consider multiple robots traversing an area with multiple adversaries. During their traversal, each robot suffers risk or damage from adversaries accumulated over

time, if it enters their impact range. We use robots' states to represent their physical locations. Each robot has two decomposable actions: the action of *move* and the action of *guard*, which are performed simultaneously. The *move* action changes the robot's locations in the environment. In each step, the relative positions of robots to adversaries determine a base risk the robot accumulates. The *guard* is an additional action each robot can choose to perform against an adversary, which increases its own risk but reduces the risks that this adversary poses to other robots. Our goal is to minimize the accumulated risk and traveling time for all robots to arrive at a target position, given each robot only has local (partial) observation of the system.

Depending on the weights in performance evaluation, the team may face trade-offs between strategies. For example, each robot may move along the shortest path to reach the target as soon as possible v.s. spending more time to form guard behaviors to reduce team risk. These strategies boil down to robots' interdependent decision-making regarding move (to where) and guard (which adversary) at each time step, making the problem non-trivial. On the other hand, as a key idea this paper highlights, the move and guard actions of robots in this scenario exhibit a hierarchical structure. This structure can be leveraged through a bi-level reformulation, which simplifies the problem's complexity while still characterizing the interdependencies between the robots' two types of actions.

The same scenario discussed above applies to a handful of applications for multi-robot coordination, such as firefighting considering fire as adversaries; battlefields considering enemy robots as adversaries; and team car racing or football games considering non-cooperative players as adversaries.

## II. LITERATURE REVIEW

### A. MARL for Multi-Robot Coordination

Reinforcement Learning (RL) has gained significant popularity for controlling robotic systems for complex tasks [12]. Using RL for multi-robot systems faces the challenge of increased system dimension and the need for robots to make decisions based on local observations. Addressing these challenges has led to the development of the CTDE [2] learning scheme, which enables robots to cooperatively learn coordination strategies through centralized training, while the learned policies are local and, thus, can be deployed for decentralized execution. Examples include MADDPG [2], which extends the centralized Deep Deterministic Policy Gradient (DDPG) framework to multi-robot systems. A similar case applies to Proximal Policy Optimization (PPO) algorithm and its generalization MAPPO [13]. While these approaches might be concerned with state dimensionality during centralized training, recent advances, such as Value-Decomposition Networks (VDN), can be leveraged to decompose the joint value function into individual value functions for each robot [14]. Building on this, QMIX further extends the framework, allowing for a more complex, state-dependent mixing of individual robot's value functions to learn coordinated behaviors [15].

These CTDE approaches often rely on end-to-end learning, which cannot leverage the hierarchical structures of robot actions considered in this paper. In addition, their performance and training efficiency may significantly degrade [16] if each robot's local observation covers only parts of the state space. Such local observation leads to a Decentralized Markov Decision Process (Dec-MDP), whose complexity is NEXP-complete [17]. In this case, even a slight reduction in the state/action space can have a huge impact on the overall computational complexity.

### B. Bi-level Optimization

Regarding state/action space reduction, bi-level optimization is a hierarchical mathematical formulation where the solution of one optimization task is restricted by the solution set mapping of another task [9]. In recent years, this technique has been incorporated with various machine learning methods for nested decision-making, including multi-task meta-learning [18], neural architecture search [19], adversarial learning [20], and deep reinforcement learning [21]. The applications of these bi-level learning approaches span economics, management, optimal control, systems engineering, and resource allocation problems [22]–[25], with a comprehensive review by Liu et al. [9]. Corresponding to the scope of this paper, when solving multi-robot coordination problems, bi-level decomposition has been introduced [11], [26] for actor-critic MARL to reduce the training burden of each level, thus improving the training efficiency and convergence. Le et al. [27] uses latent variables to introduce a mechanism similar to bi-level optimization that can significantly improve the performance of multi-agent imitation learning.

Nevertheless, a commonality among the mentioned works is the assumption that robots (agents) can access complete state information of the system. In cases where each robot has only partial information, the incomplete information can lead to a mismatch between the optimal solutions on the two levels. This challenge has motivated our work to incorporate an alignment penalty in order to generalize the bi-level optimization concept under CTDE, taking into account robots' local observations.

## III. PRELIMINARIES AND PROBLEM FORMULATION

### A. Formulation of a Bi-level Optimization

A regular centralized Markov Decision Process (MDP) can be defined by a tuple  $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$ , including state, action, state transition, discount factor, and reward. The goal is to learn a policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  to maximize the expected cumulative reward over the task horizon  $T$ , i.e.,

$$\max_{\pi} \mathbb{E}_{\mathbf{a}_t \sim \pi(\cdot | \mathbf{s}_t)} \left[ \sum_{t=0}^T \gamma^t R_t \right], \quad (1)$$

where  $\mathbf{a}_t \in \mathcal{A}$  and  $\mathbf{s}_t \in \mathcal{S}$  are the action and state of the system at each step.

Let's assume the actions of the system can be decomposed as  $\mathbf{a}_t = \{\mathbf{x}_t, \mathbf{y}_t\} \in \mathcal{X} \times \mathcal{Y}$ , and the system's state transition

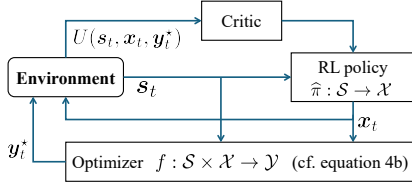


Fig. 2: A Centralized Bi-level Optimization for RL.

depends only on action  $x_t$  such that  $\mathcal{T} : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ :

$$s_{t+1} = \mathcal{T}(s_t, x_t). \quad (2)$$

However, the reward function depends on both actions:

$$R_t = U(s_t, x_t, y_t). \quad (3)$$

If the structure of problem (1) satisfies (2) and (3), it can be formulated into a bi-level problem:

$$\max_{\hat{\pi}} \mathbb{E}_{x_t \sim \hat{\pi}(\cdot | s_t)} \left[ \sum_{t=0}^T \gamma^t U(s_t, x_t, y_t^*(s_t, x_t)) \right], \quad (4a)$$

$$\text{s.t.} \quad y_t^*(s_t, x_t) = \arg \max_{y_t} U(s_t, x_t, y_t). \quad (4b)$$

In this paper, we make the following assumption:

**Assumption 1.** Problem (4a) is complex such that  $\hat{\pi}$  is solved using a reinforcement learning (e.g. actor-critic method), whereas problem (4b) can be solved explicitly and be described as a mapping  $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{Y}$ .

The assumption can be justified by the fact that (4a) is a planning problem involving state transitions and rewards for future  $T$  steps, while (4b) only involves a one-step reward.

Under Assumption 1, for the lower level problem (4b), the optimal  $y_t^*$  is solvable from  $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{Y}$ , based on the current  $s_t$  and the choice of  $x_t$ . Bringing this solution back to the upper level (4a), we only need to solve a reduced-dimension MDP defined by  $(\mathcal{S}, \mathcal{X}, \mathcal{T}, \gamma, R)$ , to obtain a policy  $\hat{\pi} : \mathcal{S} \rightarrow \mathcal{X}$  that generates action  $x_t$ . During implementation, the policy  $\pi$  in problem (1) can be a composition of  $\pi = \hat{\pi} \circ f$ . We learn the policy  $\hat{\pi}$  to generate  $x_t$  from  $s_t$ , then use  $f$  to generate  $y_t$  from  $\{s_t, x_t\}$ . Fig. 2 provides a visualization of this bi-level mechanism builds on actor-critic reinforcement learning.

### B. Bi-level Formulation for Multi-robot Coordination Learning with Local Robot Observation

In this paper, we generalize the bi-level formulation into a multi-robot system with local observation. Consider a multi-robot system with a set of  $\mathbf{n} = \{1, \dots, n\}$  robots. Suppose the centralized actions and states in (1) are decomposed into each robot's state and action as  $s_t = \{s_t^1, \dots, s_t^n\} \in \mathcal{S}$  and  $a_t = \{a_t^1, \dots, a_t^n\} \in \mathcal{A}$ . Assume each robot's action can be further written as  $a_t^i = \{x_t^i, y_t^i\}$ , where the action  $x_t^i$  determines the robot's local state update  $s_{t+1}^i = \mathcal{T}^i(s_t^i, x_t^i)$ . The multi-robot system shares a global reward determined by all robots' states, and their two actions:  $R_t = U(s_t, x_t, y_t)$ .

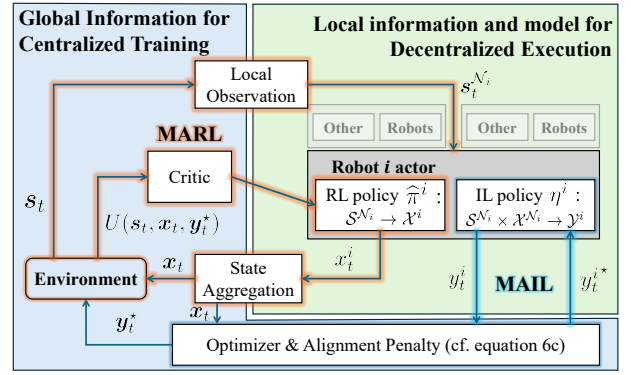


Fig. 3: A Bi-level Coordination Learning (Bi-CL) Algorithm: incorporating multi-agent reinforcement learning (MARL) and imitation learning (MAIL), guided by a global optimizer.

For many real-world applications, the deployment of a multi-robot system faces communication and sensing constraints, where each agent may only have access to the states of its neighbors through a network, denoted by set  $\mathcal{N}_i$ . The neighbors' states are a subset of the full system state, denoted by  $s_t^{\mathcal{N}_i} \subsetneq s_t$ . Such local observation turns the problem into a Dec-MDP. Under this constraint, the **problem of interest** in this paper is to learn a *local* policy  $\pi^i$  for each agent such that:

$$\max_{\pi^i} \mathbb{E}_{a_t^i \sim \pi^i(\cdot | s_t^{\mathcal{N}_i})} \left[ \sum_{t=0}^T \gamma^t R_t \right]. \quad (5)$$

While solving Dec-MDP problems is known to be difficult [17], in this paper, we will generalize the idea of bi-level optimization under the scheme of CTDE and use action space reduction to alleviate the challenges in high-dimensional parameter tuning for MARL. Note that such a generalization is non-trivial and the main challenge lies in the local observation and local decision-making of each robot using incomplete information. As a consequence, compared with the centralized case, the robots cannot solve an optimization problem like (4b) to directly determine their secondary action  $y_t^i$ .

**Running Example.** We demonstrate how formulation (5) aligns with the running example. Let state  $s_t^i$  represent robots' physical locations. The *move* action is associated with  $x_t^i$ , which impacts the robots' state update. The *guard* action is associated with  $y_t^i$ . The risk function (negative reward)  $U$  in each time step depends on the robots' current locations  $s_t^i$ , their move  $x_t^i$ , and how robots are guarding against adversaries, i.e.,  $y_t^i$ . Thus, the scenario satisfies (2)-(3) and can be reformulated into a multi-robot bi-level optimization problem. All robots only receive states from their neighbors  $s_t^{\mathcal{N}_i} \subsetneq s_t$ , through an underlying network.

## IV. MAIN APPROACH

As described in Sec. III-A, the centralized bi-level formulation leverages the optimizer (4b) to reduce the dimensionality of the action space, thereby enhancing training efficiency. However, such optimization requires global information,

which is not accessible to robots during decentralized execution. Thus, robots cannot locally generate  $y_t^*$ . To address this issue, in this section, we introduce a new reformulation to generalize the concept of bi-level optimization to a CTDE setup.

To present our approach, we reformulate (5) into the following Bi-level Coordination Learning (Bi-CL) problem,  $\forall i \in \{1, \dots, n\}$ ,

$$\max_{\hat{\pi}^i} \mathbb{E}_{x_t^i \sim \hat{\pi}^i(\cdot | s_t^{\mathcal{N}_i})} \left[ \sum_{t=0}^T \gamma^t U(s_t, x_t, y_t^*(s_t, x_t)) \right], \quad (6a)$$

$$\min_{\eta^i} \mathbb{E}_{y_t^i \sim \eta^i(\cdot | s_t^{\mathcal{N}_i}, x_t^{\mathcal{N}_i})} \|y_t^i - y_t^*(s_t, x_t)\|^2, \quad (6b)$$

$$\text{s.t. } y_t^*(s_t, x_t) = \arg \max_{y_t} \left[ U(s_t, x_t, y_t) - c_k \underbrace{\sum_{i=1}^n \mathcal{H}_i(y_t^i, \eta^i)}_{\text{alignment penalty}} \right] \quad (6c)$$

In equations (6a) and (6b), we aim to train two local policies for each robot: one policy,  $\hat{\pi}^i : \mathcal{S}^{\mathcal{N}_i} \rightarrow \mathcal{X}^i$ , generates action  $x_t^i$ ; another policy,  $\eta^i : \mathcal{S}^{\mathcal{N}_i} \times \mathcal{X}^{\mathcal{N}_i} \rightarrow \mathcal{Y}^i$ , generates action  $y_t^i$ . Note that the two policies are coupled since their inputs rely on the outputs of each other. If training two policies simultaneously, such coupling can pose a stability issue and reduce training efficiency. To address this, our idea is to introduce a global optimizer (6c) under the CTDE scheme to decouple and guide the training of both policies, as we explain next and visualized in Fig. 3.

**Bi-level optimization:** First, we introduce optimization problem (6c) to solve  $y_t^*$ . If temporarily ignoring the alignment penalty term, the equations (6a) and (6c) together represent a CTDE version of the bi-level problem (4). During the centralized training process, we can utilize global information for all robots' states  $s_t$  and actions  $x_t$  to solve  $y_t^*$ . Based on  $y_t^*(s_t, x_t)$ , we train the local policy  $\hat{\pi}^i$  in (6a) for each agent's  $x_t^i$  using a *reinforcement learning* scheme.  $\hat{\pi}^i$  has a *reduced-dimension* action space because  $y_t^*$  is not learned.

On the other hand, while optimizer (6c) can be used in centralized training to obtain action  $y_t^*$  for all robots, it cannot be used by robots during decentralized execution due to the local observations. Thus, we introduce (6b) to train a local policy  $\eta^i$  to generate agent's  $y_t^i$ . The policy  $\eta^i$  can be trained conveniently using an imitation learning (IL) scheme that uses optimizer (6c) as the *expert demonstration*. Here, the imitation loss in (6b) measures the difference between the policy output  $y_t^i$  using local information and the optimal action solved from optimizer (6c) using global information.

**Alignment of the two policies:** We explain the alignment term in (6c). Based on the above explanation, the policy  $\hat{\pi}^i$  is trained assuming the actions  $y_t^* = \{y_t^{1*}, \dots, y_t^{n*}\}$  are optimal for all robots, and the policy  $\eta^i$  seeks to generate such optimal actions using only local information. Clearly, during decentralized execution, the optimality of  $\eta^i$  can hardly be guaranteed, thus creating a mismatch between the policies  $\hat{\pi}^i$  and  $\eta^i$ . To address this, we introduce an alignment

---

### Algorithm 1: The Bi-CL Algorithm

---

```

1 Initialize local RL models  $\hat{\pi}^i(s^{\mathcal{N}_i} | \theta_{\hat{\pi}^i})$  and IL
   models  $\eta^i(s^{\mathcal{N}_i}, x^{\mathcal{N}_i} | \theta_{\eta^i})$  for all robots. Initialize
   centralized critic  $Q(s, x | \theta_Q)$  if needed;
2 Initialize alignment penalty co-efficient  $c_k$ ;
3 for  $k = 0$  to  $T_k$  do
4   Update coefficient  $c_k = \frac{c}{1 + e^{-\beta(k-h)}}$ ;
5   for  $t = 0$  to  $T_t$  do
6     for  $i = 1$  to  $n$  do
7       Get action:  $x_t^i$  by adding random
         perturbation to  $\hat{\pi}^i(s_t^{\mathcal{N}_i} | \theta_{\hat{\pi}^i})$ ;
8       Get action:  $\hat{y}_t^i = \eta^i(s_t^{\mathcal{N}_i}, x_t^i | \theta_{\eta^i})$ ;
9       State update:  $s_{t+1}^i = \mathcal{T}^i(s_t^i, x_t^i)$ ;
10    end
11    Solve  $\arg \max_{y_t} [U(s_t, x_t, y_t) - c_k \sum_{i=1}^n \mathcal{H}_i]$ 
       to obtain  $y_t^*$ ;
12    Observe reward  $R_t = U(s_t, x_t, y_t^*)$ ;
13    Record  $(s_t, x_t, y_t^*, R_t, s_{t+1})$  to a fixed size
       buffer  $\mathcal{B}$  with first-in-first-out;
14    Sample  $(s_\tau, x_\tau, y_\tau^*, r_\tau, s_{\tau+1})$  from  $\mathcal{B}$  as a
       random minibatch of size  $w$ ;
15    Update critic network  $\theta_Q$  using the minibatch,
       if critic exist;
16    for  $i = 1$  to  $n$  do
17      Update IL learning model  $\theta_{\eta^i}$  using the
        minibatch;
18      Update RL learning model:  $\theta_{\hat{\pi}^i}$  using the
        minibatch;
19    end
20  end
21 end

```

---

penalty term  $c_k \sum_{i=1}^n \mathcal{H}_i(y_t^i, \eta^i)$  in (6c), where  $c_k > 0$  and

$$\mathcal{H}_i = \mathbb{E}_{y_t^i \sim \eta^i(\cdot | s_t^{\mathcal{N}_i}, x_t^{\mathcal{N}_i})} \|y_t^i - \hat{y}_t^i\|^2.$$

Here,  $\hat{y}_t^i$  is the output of the model  $\eta^i$ , then  $\mathcal{H}_i$  evaluates how well an action  $y_t^i$  aligns with the policy  $\eta^i$ . Minimizing this penalty helps to reduce the mismatch between the two local policies. Mathematically, the value of  $\mathcal{H}_i$  for each robot equals the loss in (6b). Hence,  $c_k$  is an important coefficient and we shall remark on its choice during the training. If  $c_k \rightarrow 0$ , the reformulation (6) disregards the policy mismatch, always opting for the solution that maximizes  $U$  to train  $\hat{\pi}^i$ . Conversely, when  $c_k \rightarrow \infty$ , the penalty forces the training of  $\hat{\pi}^i$  using the action chosen by  $\eta^i$ , making the two policies fully coupled, and the IL (6b) no longer updates. To strike a balance, our idea is to make sure the IL model is sufficiently trained before the penalty is applied. For this purpose, we define a modified logistic function:

$$c_k = \frac{c}{1 + e^{-\beta(k-h)}} \quad \text{with } c > 0, \beta > 0, h > 0 \quad (7)$$

where  $k$  is the training episode. We let  $c_k$  start with zero to 'jump start' the training of  $\hat{\pi}^i$  and  $\eta^i$  using the optimal  $y_t^*$  for maximizing  $U$ . Afterwards,  $c_k$  is gradually increased

to a sufficiently large value  $c$  to fine-tune  $\hat{\pi}^i$ , taking into account the output of  $\eta^i$  and ensuring alignment between the two policies,  $\hat{\pi}^i$  and  $\eta^i$ . The described training scheme is summarized in Algorithm 1.

**Remark 1.** In the presentation of Algorithm (1), we assume the use of actor-critic-based methods for (6a) and a quadratic mismatch function for the imitation loss (6b). However, the proposed algorithm is also applicable to general CTDE setups when using other reinforcement learning and imitation learning methods. Furthermore, in handling the local observations of robots, there exist more advanced methods [28, Sec. 2.2.2] in the literature on Dec-MDPs and POMDPs that leverage the memory (historic states) of robots to learn more powerful policies. In this paper, we only use simple memory-less policies, but these more advanced methods are also compatible with Algorithm (1).

## V. NUMERICAL RESULTS

In this section, we employ two variants of the running example to verify the proposed algorithm. We use simulated experiments to demonstrate: (i) the effectiveness of the proposed bi-level learning schemes, in particular, the alignment penalty term; (ii) the advantage of the proposed approach in terms of training efficiency compared with alternative MARL algorithms.

Through this section, we use the accumulated reward (5) to evaluate the proposed Bi-CL algorithm. To avoid ambiguity, we distinguish the following metrics:

**RL-Reward:** refers to the reward of the reinforcement learning policy  $\hat{\pi}^i$  when solving (6a) during centralized training.

**T-Reward:** refers to an average reward obtained by executing both the learned reinforcement learning policy  $\hat{\pi}^i$  and imitation learning policy  $\eta_i$  in the environment for thirty times.

**R-Gap:** refers to the subtraction between the RL-Reward and the T-Reward.

**Remark 2.** Note that RL-Reward and T-Reward are different because they use different ways to choose robots' action  $y_t^i$ . RL-Reward uses an 'optimal' solution  $\mathbf{y}_t^* = \{y_t^{1*}, \dots, y_t^{n*}\}$  of (6c); T-Reward uses output of the learned policy  $\eta_i$ . Only the T-Reward reflects the true performance of the policies because it is achievable under decentralized execution, and higher means better. Since R-Gap evaluates the difference, smaller means better.

### A. Coordinated Multi-robot Route Traversal

**Running Example (a):** We introduce a variant of the running example, by assuming all robots travel continuously through a route. Each robot can only observe a subset of full system states through an underlying communication network. Fig. 4 provides a visualization of the environment. The testing environment we use is a mathematical abstraction that may not exactly follow its physical layout and the number of adversaries may change.

- Each robot  $i$  has a continuous **move action** (velocity)  $x_t^i \in [-v_{\max}, v_{\max}]$ , and a discrete **guard action**  $y_t^i \in \mathcal{M}$  where  $\mathcal{M}$  is the set of all adversaries. The robot  $i$ 's **state**

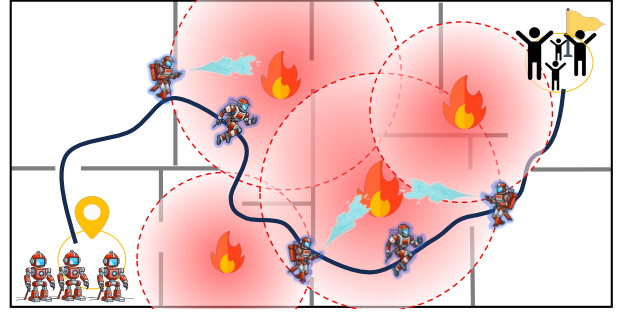


Fig. 4: Running Example (a): all robots travel along a route.

**transition** (position) follows dynamics  $s_{t+1}^i = s_t^i + x_t^i$ . These align with definition (2).

- Suppose adversary  $j \in \mathcal{M}$  has an impacted area  $\mathcal{B}_j$ . Each time-step, if  $s_t^i \in \mathcal{B}_j$ , robot  $i$  accumulates a cost  $c_j(s_t^i)$ . Besides, for any robot  $k$ , if  $s_t^k \in \mathcal{B}_j$ , it can perform guard against adversary  $j$ , i.e.,  $y_t^k = j$ . The guarding effect is characterized by a discount factor on the costs created by adversary  $j$ :

$$\alpha_t^{k,j}(x_t^k, y_t^k) = \begin{cases} 1 - \beta \frac{(v_{\max} - x_t^k)}{v_{\max}} & y_t^k = j, \\ 1 & \text{otherwise.} \end{cases}$$

Such a discount is more effective when the robot admits at a lower moving velocity  $x_t^i$ . The total **team cost** in each step is defined as:

$$R_t = U(\mathbf{s}_t, \mathbf{x}_t, \mathbf{y}_t) = - \sum_{i=1}^n \sum_{j=1}^m \left[ \prod_{k=1}^n \alpha_t^{k,j}(x_t^k, y_t^k) c_t^{i,j}(s_t^i) \right] - \delta,$$

where  $\delta$  is a constant time penalty. A one-time positive reward is added when all robots arrive at the target. The definition of  $U$  aligns with definition (3).

- The goal is to minimize the team accumulated cost in the form of (5) before all robots arrive at the target position.

The alignment of the problem with (2) and (3) allows it to be reformulated into (6) and then solved by our Bi-CL.

**Effectiveness of Alignment Penalty:** We first implement the proposed Bi-CL in an environment with 4 robots and 4 adversaries (fire). In Bi-CL, the reinforcement learning of  $\hat{\pi}^i$  uses an independent actor in each robot with centralized critic [29, Sec. 2.4], following the structure of MADDPG [2] but over reduced action space. Each robot only observes part of the global information, i.e., its own state and the states of a subset of other robots in the system. The cost function  $c_j$  is an affine function that depends on the closeness between the robot and the center of the adversary.

Fig. 5 illustrates the RL-Reward curves and T-Reward for different  $c_t$  by setting  $c = 0, c = 1, c = 5, c = 10, c = 50$ , all with  $\beta = 2e - 3$  and  $h = 3000$ . This, together with (7), makes  $c_k$  gradually increase in an 'S' shape from 0 to  $c$ . Upon examining the results, it is evident that the R-Gap is large when  $c$  is small, and the gap diminishes as  $c$  increases. This verifies the effectiveness of the proposed alignment penalty term in (6c), as it motivates the policy  $\hat{\pi}_t^i$



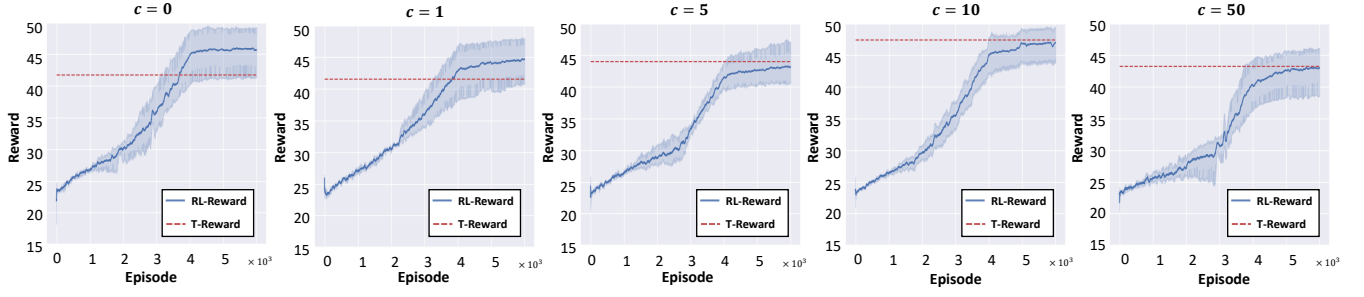


Fig. 5: Comparison of cumulative reward for different alignment penalties with four robots and four adversaries. The height of red dash lines determines implementation performance.

to be tuned using a  $y_t^{i*}$  that is closer to the output of  $\eta_t^i$ . It is also important to note that although the MARL training reward appears high when  $c$  is small, this is misleading and unattainable in decentralized execution due to the mismatch between  $\hat{\pi}_t^i$  and  $\eta_t^i$ . The attainable reward (T-Reward) is much lower with smaller  $c$ .

Finally, we observe in the last plot of Fig. 5 that a very large  $c = 50$  may negatively impact the training performance. This is due to two reasons. First, as we discussed in the algorithm development, larger  $c_k$  reduces the training efficiency of the imitation learning part of the algorithm. This is reflected by the lower values on both RL-Reward and T-Reward. Second, since  $c_k$  impacts the computation of  $y_t^{i*}$  during the training process, changing it too aggressively will lead to instability in training. This is evidenced by the middle part of the curve, where increased oscillation is observed.

The same test is performed under various environment setups and the results are shown in Table I. Here,  $Na\_Mb$  means the environment has  $a$  robots and  $b$  adversaries. Especially, in  $N5\_M4^*$ , each robot further reduces the number of other robots's states it can observe. The results in the table align with our above analysis. In all cases where  $c_k = c = 0$ , i.e., without alignment penalty, the performance is the worst. The R-Gap generally reduces when  $c$  grows large. Since T-Reward is the core metric for performance, the best result mostly occurs at  $c = 10$  with only one exception and the difference is small. Furthermore, to read the table column-wise, environments with the same number of adversaries, i.e.,

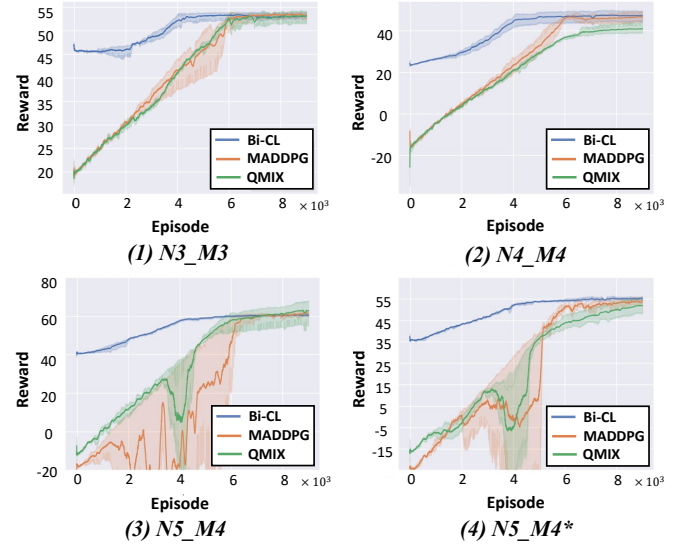


Fig. 6: Performance Comparisons in Different Scenarios.

$M4$  are comparable. From  $N4\_M4$  to  $N5\_M4$ , the T-Rewards generally increase because more coordination behaviors can be generated. When comparing  $N5\_M4^*$  and  $N5\_M4$ , the T-Rewards generally decrease due to the reduced sensing distances of the robots. This also leads to a larger R-Gap when  $c = 0$ , necessitating the introduced alignment penalty.

**Comparing Training Efficiency with Baselines:** We also compare the proposed Bi-CL algorithm with two well-established MARL algorithms, MADDPG [2] and QMIX [15], respectively. MADDPG and QMIX are implemented in complete action space  $\mathcal{A}^i = \mathcal{X}^i \times \mathcal{Y}^i$  of each robot. The same learning rates with Bi-CL are used. The comparison is visualized in Fig. 6 for the four cases in the above discussion. Here, we choose  $c = 10$  as the R-Gaps are small. Thus, the RL-Reward curve can represent the convergence of our algorithm and can approximate the true reward our algorithm can achieve. It can be observed that in all cases, Bi-CL can achieve similar final rewards compared with baselines, which justifies the effectiveness of the proposed algorithm. The efficiency of Bi-CL is evidenced by its faster convergence speed compared with baselines. Furthermore,

TABLE I: Average reward per episode of different  $c_k$  values

Environments	Metric	$c_k = 0$	$\beta = 2e^{-3}$			
			$c = 1$	$c = 5$	$c = 10$	$c = 50$
$N3\_M3$	T-Reward	51.83	53.01	<b>53.44</b>	53.17	52.13
	R-Gap	1.34	0.73	0.26	-0.45	<b>-0.23</b>
$N4\_M4$	T-Reward	42.10	42.03	44.05	<b>47.38</b>	43.56
	R-Gap	3.98	1.55	-0.87	-0.39	<b>-0.11</b>
$N5\_M4$	T-Reward	56.31	58.27	59.68	<b>60.08</b>	59.46
	R-Gap	4.79	4.45	1.17	<b>-0.04</b>	-0.15
$N5\_M4^*$	T-Reward	50.67	53.39	54.09	<b>54.60</b>	54.20
	R-Gap	6.54	3.84	1.82	-0.59	<b>0.27</b>

the starting reward of our algorithm is significantly better because it uses the optimization (6c) to boost and guide the policy training, while other methods are purely based on exploration. For a similar reason, we observe that both MADDPG and QMIX suffer from training stability issues for complex scenarios, while the proposed Bi-CL does not.

### B. Coordinated Multi-robot Graph Traversal

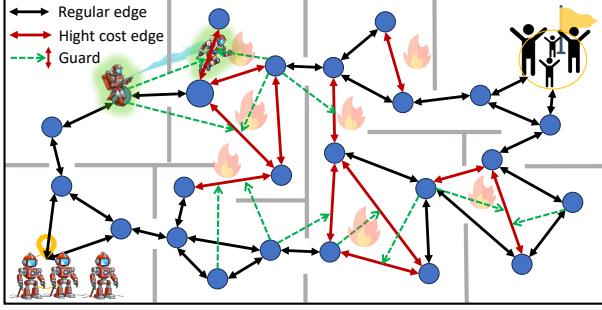


Fig. 7: Running Example (b): all robots travel through a graph.

**Running Example (b):** In the second scenario, we assume all robots travel through a graph environment, which highlights the coordination of robots from a spatial aspect. Fig. 7 provides a visualization of the environment; however, the testing environment we use includes randomly generated graphs that may not exactly follow the physical layout. Each robot can only observe a subset of full system states. The environment is characterized by an environment graph describing robot traversability from one node to another. It also features a coordination mechanism (green arrows) describing some robots, when located on certain nodes, can *guard* other robots' edge traversal to reduce their cost, similar to the work by Limbu et al. [30], [31].

- Each robot  $i$  has a discrete **move action**  $x_t^i \in \mathcal{N}(s_t^i)$  that moves the robot to a new node in the graph based on its current position  $s_t^i$ . Here,  $\mathcal{N}(s_t^i)$  is the set of adjacent neighboring nodes of robot position  $s_t^i$  in the environment graph. The robot  $i$ 's **state transition** (position) follows  $s_{t+1}^i = x_t^i$  subject to  $x_t^i \in \mathcal{A}(s_t^i)$ . The robot also has a discrete **guard action**  $y_t^i \in \mathcal{G}(x_t^i)$ , where  $\mathcal{G}(x_t^i)$  is the set of edges that can be guarded by a robot at its new position  $s_{t+1}^i = x_t^i$  (visualized by green arrows in Fig. 7). The robots' actions and state transitions align with definition (2).

- When robot  $i$  travels an edge  $\{s_t^i, x_t^i\}$ , i.e., from  $s_t^i$  to  $s_{t+1}^i = x_t^i$ , there is an associated edge cost  $c(\{s_t^i, x_t^i\})$ . Besides, based on robot coordination, if another robot  $k$  can and chooses to guard this edge, i.e.,  $y_t^k = \{s_t^i, x_t^i\} \in \mathcal{G}(x_t^k)$ , then this edge cost will be discounted by a factor

$$\alpha_t^{i,k}(s_t^i, x_t^i, y_t^k) = \begin{cases} \alpha^* & y_t^k = \{s_t^i, x_t^i\} \in \mathcal{G}(x_t^k), \\ 1 & \text{otherwise.} \end{cases}$$

Based on this, the total **team cost** in each step is defined as:

$$R_t = U(s_t, \mathbf{x}_t, \mathbf{y}_t) = -\sum_{i=1}^n \left[ \prod_{k=1}^n \alpha_t^{i,k}(s_t^i, x_t^i, y_t^k) c(\{s_t^i, x_t^i\}) \right] - \delta$$

where  $\delta$  is a constant time penalty. A one-time positive reward is added when all robots arrive at the target. The definition of  $U$  aligns with definition (3).

- The goal is to minimize the team accumulated cost in the form of (5) before all robots arrive at the target position.

The alignment of the problem with (2) and (3) allows it to be reformulated into (6) and then solved by our Bi-CL.

TABLE II: Comparison of Average Reward and Convergence time for Running Example (b)

Graph	Robots/ Nodes	T-Reward			Converge ( $\times 10^3$ )		
		Bi-CL	MAPPO	QMIX	Bi-CL	MAPPO	QMIX
Sparse	3/5	<b>83.16</b>	82.08	78.82	<b>3.58</b>	4.89	5.85
	3/10	64.86	<b>65.39</b>	64.97	<b>3.69</b>	4.87	5.95
	5/10	54.91	50.33	<b>55.21</b>	<b>3.75</b>	4.96	6.14
	5/15	<b>51.37</b>	33.14	49.67	<b>3.71</b>	5.24	6.03
Dense	3/5	93.21	<b>96.90</b>	94.23	<b>3.63</b>	4.95	6.19
	3/10	<b>73.41</b>	60.89	71.08	<b>3.79</b>	5.09	6.38
	5/10	66.32	54.14	<b>68.08</b>	<b>3.80</b>	4.94	6.29
	5/15	<b>60.08</b>	47.82	58.16	<b>3.85</b>	4.94	6.44

**Comparing Training efficiency with Baselines:** In this scenario, we compare the proposed Bi-CL algorithm with two MARL algorithms MAPPO [2] and QMIX [15], respectively. Here, we use MAPPO instead of MADDPG because centralized PPO's suitability for this task has been validated by the study in Limbu et al. [31]. In Bi-CL, the reinforcement learning part is implemented by QMIX in action space  $\mathcal{X}_i$  for each robot. While the MAPPO and original QMIX are implemented in complete action space  $\mathcal{A}^i = \mathcal{X}^i \times \mathcal{Y}^i$  of each robot. The simulations are configured with the following setups. We experiment with 5, 10, 15 nodes in the environment graph with 3 or 5 robots in two types of graph connectivity, i.e., sparse and dense, for edges and the number of edges that can be guarded. We randomly create 15 environment graphs as our test set. For all approaches, training is terminated when the cumulative reward no longer changes more than 1% for 10 steps.

The comparison results are presented in Table II, which verifies the effectiveness of the Bi-CL algorithm for learning coordinated actions among robots within graph-based coordination tasks. Despite variations in the number of robots, graph nodes, or densities of the graph edges, the proposed Bi-CL algorithm delivers comparable results to the baseline algorithms while requiring fewer training episodes to converge. By observing the column of MAPPO, the results tend to fall below those of other methods under conditions of increased environmental complexity. This suggests that policy gradient approaches may only learn a sub-optimal policy solution in these cases. The QMIX method is more robust in terms of optimality, however, it generally requires more training episodes to converge. In comparison, the proposed Bi-CL method enjoys much more efficient convergence properties. This improvement is mainly due to (i) the reduction of the action space for reinforcement learning, and (ii) the use of optimization (6c) to jump-start the training.

## VI. CONCLUSION

We presented a bi-level formulation for multi-robot coordination learning with local observation, wherein robots' state transitions and their cooperative behaviors are abstracted and optimized on different levels, significantly enhancing learning efficiency. A key enabler of our Bi-CL algorithm was an alignment penalty that enables upper-level learning to account for potential discrepancies arising from local observations in lower-level optimization. We validated our algorithm through a running example, designing two distinct environments: route-based and graph-based. Experimental results demonstrated that our algorithm can effectively learn in both environments, underscoring its versatility and applicability across a diverse set of coordination tasks. We evaluated the performance enhancement of the Bi-CL using different alignment penalty parameters. Comparative analysis with baselines verified the efficiency of our algorithm.

For future work, we aim to explore the scalability of our Bi-CL to accommodate larger multi-robot systems and more complex environments, further refining the alignment penalty mechanism to enhance its adaptability and efficiency. Moreover, we intend to extend our way of handling robots' information loss to effectively manage dynamic, stochastic, and noisy scenarios, thereby enhancing its resilience and performance in unpredictably evolving multi-robot coordination environments.

## REFERENCES

- [1] Y. Rizk, M. Awad, and E. W. Tunstel, "Cooperative heterogeneous multi-robot systems: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–31, 2019.
- [2] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.
- [3] Z. Yan, N. Jouandeau, and A. A. Cherif, "A survey and analysis of multi-robot coordination," *International Journal of Advanced Robotic Systems*, vol. 10, no. 12, p. 399, 2013.
- [4] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *Ieee Access*, vol. 8, pp. 191617–191643, 2020.
- [5] X. Xiao, J. Dufek, T. Woodbury, and R. Murphy, "Uav assisted usv visual navigation for marine mass casualty incident response," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6105–6110, IEEE, 2017.
- [6] X. Xiao, J. Dufek, and R. R. Murphy, "Autonomous visual assistance for robot operations using a tethered uav," in *Field and Service Robotics: Results of the 12th International Conference*, pp. 15–29, Springer, 2021.
- [7] B. Liu, X. Xiao, and P. Stone, "Team orienteering coverage planning with uncertain reward," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9728–9733, IEEE, 2021.
- [8] Y. Lian, Q. Yang, Y. Liu, and W. Xie, "A spatio-temporal constrained hierarchical scheduling strategy for multiple warehouse mobile robots under industrial cyber-physical system," *Advanced Engineering Informatics*, vol. 52, p. 101572, 2022.
- [9] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10045–10067, 2021.
- [10] A. Biswas and C. Hoyle, "A literature review: solving constrained non-linear bi-level optimization problems with classical methods," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 59193, p. V02BT03A025, American Society of Mechanical Engineers, 2019.
- [11] H. Zhang, W. Chen, Z. Huang, M. Li, Y. Yang, W. Zhang, and J. Wang, "Bi-level actor-critic for multi-agent coordination," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7325–7332, 2020.
- [12] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.
- [13] C. Yu, A. Velu, E. Vinitisky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24611–24624, 2022.
- [14] P. Sunehag, G. Lever, A. Grusl, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, et al., "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.
- [15] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7234–7284, 2020.
- [16] P. J. Gmytrasiewicz and P. Doshi, "A framework for sequential planning in multi-agent settings," *Journal of Artificial Intelligence Research*, vol. 24, pp. 49–79, 2005.
- [17] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.
- [18] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," in *International Conference on Learning Representations*, 2018.
- [19] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *International Conference on Learning Representations*, 2018.
- [20] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv preprint arXiv:1610.01945*, 2016.
- [21] M. Hong, H. Wai, Z. Wang, and Z. Yang, "A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic," *arxiv e-prints*, art, *arXiv preprint arXiv:2007.05170*, 2020.
- [22] J. Fortuny-Amat and B. McCarl, "A representation and economic interpretation of a two-level programming problem," *Journal of the operational Research Society*, vol. 32, pp. 783–792, 1981.
- [23] A. Sinha, P. Malo, and K. Deb, "A review on bilevel optimization: From classical to evolutionary approaches and applications," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 276–295, 2017.
- [24] Y. Mintz, J. A. Cabrera, J. R. Pedrasa, and A. Aswani, "Control synthesis for bilevel linear model predictive control," in *2018 Annual American Control Conference (ACC)*, pp. 2338–2343, IEEE, 2018.
- [25] T. Nisha, D. T. Nguyen, and V. K. Bhargava, "A bilevel programming framework for joint edge resource management and pricing," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17280–17291, 2022.
- [26] L. Zheng, T. Fiez, Z. Alumbaugh, B. Chasnov, and L. J. Ratliff, "Stack-elberg actor-critic: Game-theoretic reinforcement learning algorithms," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 9217–9224, 2022.
- [27] H. M. Le, Y. Yue, P. Carr, and P. Lucey, "Coordinated multi-agent imitation learning," in *International Conference on Machine Learning*, pp. 1995–2003, PMLR, 2017.
- [28] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.
- [29] Y. Xiao, W. Tan, and C. Amato, "Asynchronous actor-critic for multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4385–4400, 2022.
- [30] M. Limbu, Z. Hu, S. Oughourli, X. Wang, X. Xiao, and D. Shishika, "Team coordination on graphs with state-dependent edge costs," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 679–684, IEEE, 2023.
- [31] M. Limbu, Z. Hu, X. Wang, D. Shishika, and X. Xiao, "Team coordination on graphs with reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024.