Multiagent Learning of Coordination in Loosely Coupled Multiagent Systems

Chao Yu, Minjie Zhang, Senior Member, IEEE, Fenghui Ren, and Guozhen Tan

Abstract—Multiagent learning is a promising technique for agents to learn efficient coordinated behaviors in multiagent systems. In multiagent learning, concurrent multiple distributed learning processes can make the learning environment non-stationary for each individual learner. Developing an efficient learning approach to coordinate agents’ behaviors in this dynamic environment is a difficult problem, especially when agents do not know the domain structure and have only local observability of the environment. In this paper, a coordinated multiagent learning approach is proposed to enable agents to learn efficient coordinated behaviors by exploiting agent independence in loosely coupled multiagent systems. The main feature of the proposed approach is to explicitly quantify and dynamically adapt agent independence during learning so that agents can make a trade-off between a single-agent learning process and a coordinated learning process for an efficient decision making. The proposed approach is employed to solve two-robot navigation problems in different scales of domains. Experimental results show that agents using the proposed approach can learn to act in concert or independently in different areas of the environment, which results in great computational savings and near optimal performance.

Index Terms—Multiagent learning, reinforcement learning, agent independence, coordination, sparse interactions.

I. INTRODUCTION

Multi-Agent Learning (MAL) provides a promising paradigm for studying how agents can learn coordinated behaviors in Multi-Agent Systems (MASs) [1], and it is finding increasing applications in a wide range of real-world domains such as robotics [2], [3], distributed control [4], [5], resource management [6], [7] and automated trading [8], [9]. MAL uses techniques and concepts from disciplines such as artificial intelligence, game theory, psychology, cognition and sociology, and has attracted a great deal of interest in the research community in recent years [10], [11], [12], [13], [14]. The main challenge in MAL is that each learner must adapt its behavior concurrently in the context of other co-learners. This co-adaptation makes the learning environment non-stationary for each learner. Dynamics in such a non-stationary environment can cause the learning goal of a learner to change continuously, making MAL into a “moving-target learning” problem [1]. Moreover, agents in MAL need to use incomplete information to reason about other agents’ possible actions in order to make a reasonable decision. MAL of coordination has been studied intensively. The need for coordination arises because the effect of an agent’s action on the environment also depends on the actions of other agents in the same environment. Hence, the agents’ actions must be mutually consistent in order to achieve their intended effects. This paper focuses on how to exploit inherent agent independence relationships for efficient coordinated learning, without imposing on the agents the important assumptions that are required in most existing approaches. This research is driven by the fact that many real-world MASs can exhibit a large amount of context-specific independence so that a general decision-making problem can be decomposed efficiently into sub-problems which are easier to solve [15]. In this kind of MASs, different levels of independence between agents can confine coordinated behaviors to some specific parts of the environment. An agent thus only needs to consider other agents’ information for coordination when it is necessary.

Many studies have investigated MAL of coordination by exploiting agent independence under different emphasis and assumptions. For example, hierarchical MAL approaches have been developed to take advantage of agents’ structural dependence to speed up the skill acquisition in cooperative MASs [16], [17], [18]. Agent independence in these approaches is represented by a decomposition of a main task into several subtasks, each of which is relevant to some particular agents. Other examples include coordination-graph based learning approaches [19], [20] and the distributed value function approach [21], with the focus on studying how an optimal global joint policy can be achieved through local interactions among agents in a networked interaction structure. All these approaches focus on coordinated MAL when an explicit representation of agent independence is given beforehand. This explicit representation is indicated either by a predefined task decomposition or by a fixed interaction structure. In many cases, however, agents must learn coordinated behaviors when an explicit representation of agent independence is not available as a prior knowledge. As a result, agents need to build up such a representation from their own learning experiences. Some approaches to coordinated MAL, therefore, have been proposed to enable agents to learn their dependent relationships so that coordination is only considered when agents are really dependent on each other [20], [22], [23], [24]. These approaches, however, are based on prior knowledge of each agent’s individual optimal policy [23] or on the assumptions of agents’ full observability of the joint-state and/or joint-action of other agents [20], [22], [24]. These preconditions in existing approaches put great limits to the development of efficient coordinated MAL approaches in practical MASs.

Against this background, this paper proposes an approach to exploit agent independence for efficient coordinated learning, without requiring agents’ knowledge about the location of coordination or full observability of the environment. We firstly formalize agent independence with the aim of decomposing the decision-making process into sub-processes more efficiently than that in general decision-making models of MAL. We then propose a coordinated MAL approach which enables agents to learn a trade-off between a single agent learning process and a coordinated learning process through dynamic adaptation of the independence between agents. The single agent learning process can take advantage of the agents’ independence from one another in order to achieve computational savings, while the coordinated learning process can take advantage of each agent’s limited observability to mitigate the uncertainties of the learning environment in order to improve learning. Through the dynamic trade-off between these two processes, agents can
learn an efficient coordinated policy with minimum computational consumption. The proposed approach is employed to solve different scales of robot navigation problems. Experimental results show that agents using our approach can learn to act in concert or independently in different regions of the environment according to the approximated independence between the agents. A large amount of computational savings can be achieved and at the same time a near optimal performance can be guaranteed by using the proposed approach.

The rest of the paper is organized as follows. Section II introduces general decision-making models in MAL. Section III gives the description of robot navigation problems and formal definition of agent independence. Section IV describes the proposed MAL approach and Section V presents the experimental results. Section VI compares this study with related studies and finally Section VII draws some conclusions and suggests directions for future research.

II. GENERAL DECISION-MAKING MODELS IN MAL

This section briefly reviews the single-agent Markov decision process model and the extended multiagent model. Fundamental concepts are clarified and notations are established for further description.

A. Markov Decision Process

A Markov Decision Process (MDP) can be used to describe a single-agent sequential decision-making problem in which an agent must choose an action at every time step to maximize some reward-based functions [25]. Formally, an MDP can be defined by a 4-tuple \( M = (S, A, P, R) \), where \( S \) is a finite state space, \( A \) is a set of actions available to the agent, \( P(s, a, s') : S \times A \times S \rightarrow [0, 1] \) is a Markovian transition function when the agent transits from state \( s \) to \( s' \) after taking action \( a \), and \( R : S \times A \rightarrow R \) is a reward function that returns immediate reward \( R(s, a) \) to the agent after taking action \( a \) in state \( s \). An agent’s policy \( \pi : S \times A \rightarrow [0, 1] \) is a probability distribution that maps a state \( s \in S \) to an action \( a \in A \).

![Fig. 1. An example of MDP.](image)

Fig. 1 gives an example of MDP regarding running a company. In every state of the company (Famous or Unknown and Rich or Poor), the company has two actions of either Saving money (S) or Advertising (A). The company will transit to another state according to certain probability using either action and receive a reward (e.g., +10 in state Rich & Unknown). The goal in an MDP is to learn a policy \( \pi \) so as to maximize the expected discounted reward \( V^\pi(s) \) for each state \( s \in S \). The reward function is given by Eq. (1).

\[
V^\pi(s) = E_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) | s_0 = s \right]
\]  
(1)

where \( E_\pi \) is the expectation of policy \( \pi, s_t \) denotes the state at time \( t \), and \( \gamma \in [0, 1) \) is a discount factor.

For any finite MDP, there is at least one optimal policy \( \pi^* \), such that \( V^{\pi^*}(s) \geq V^\pi(s) \) for every policy \( \pi \) and every state \( s \in S \). \( \pi^* \) can be computed by using linear programming or dynamic programming techniques if an agent fully knows the reward and transition functions of the environment. When these functions are unknown to the agent, finding an optimal policy in MDP can be solved using Reinforcement Learning (RL) [26] methods, in which an agent learns through trial-and-error interactions with its environment. One of the most important and widely used RL approach is Q-learning [27], which is an off-policy model-free temporal difference control algorithm. In Q-learning, an agent makes a decision through the estimation of a set of Q values. The one step updating rule is given by Eq. (2), where \( \alpha \in (0, 1) \) is a learning rate.

\[
Q_{t+1}(s, a) = Q_t(s, a) + \alpha [R(s, a) + \gamma \max_{a'} Q_t(s', a') - Q_t(s, a)]
\]  
(2)

Every Q value of a state-action pair can be stored in a table for a discrete state-action space. It is proved that this tabular Q-learning method converges to the optimal \( Q'(s, a) \) w.p.1 when all state-action pairs are visited infinitely and an appropriate exploration strategy and learning rate are chosen [27].

B. Decentralized MDPs

A Decentralized MDPs (Dec-MDPs) model is an extension of the aforementioned single agent MDP model to allow decentralized decision makings of multiple agents [28]. In Dec-MDPs, at each step, each agent has a local observation and subsequently chooses an action. An agent’s state transitions and rewards depend on the actions of all the agents. More formally, a Dec-MDPs model can be defined by a tuple \( Q = (n, S, \{A_i\}, P(s, a, s'), R(s, a), w) \), where \( n \) is the number of agents; \( S \) is a finite set of joint states of all agents; \( A_i (i \in [1, n]) \) is a finite set of actions available to agent \( i \); \( P(s, a, s') \), where \( s \in S \) is the joint state of all the agents and \( a \in A = \times_{i=1}^{n} A_i \) is the joint action of all the agents, represents the transition probability from state \( s \) to state \( s' \) when the joint action \( a \) is taken in state \( s \); and \( R(s, a) \) represents the reward received by the agents when joint action \( a \) is taken in state \( s \). For simplicity, we write \( a = \langle a_1, a_2, \ldots, a_n \rangle \) to denote the reduced action of agent \( i \), thus the joint action \( a = \langle a_1, \ldots, a_n \rangle \) can be simply represented as \( a = \langle a_i, \ldots, a_n \rangle \). This paper focuses on a more specialized version of the Dec-MDPs model, that is, factored Dec-MDPs [29], [15], in which the system state can be factored into \( n+1 \) distinct individual components so that \( S = S_0 \times S_1 \times \ldots \times S_n \), where \( S_0 \) denotes an agent-independent component of the state (i.e., assumed common knowledge among all agents), \( S_i (i \in [1, n]) \) is the state space of agent \( i \). For each agent, its local state \( s_i \) is defined as \( s_i = (s_0, s_i) (s_i \in S_i \times S_0) \), where \( s_i \in S_i \) represents the portion of global state specific to agent \( i \) and \( s_0 \in S_0 \) represents the portion of the global state shared among all agents.

The Dec-MDPs model is a particular case of the decentralized partial observable MDP model [30], where every agent has a partial observability so that the agent cannot determine its local state unambiguously through its local observation. In Dec-MDPs, however, an agent can have a local observability, which means that each agent can determine the corresponding local state from its local observations. In other words, agents in Dec-MDPs, altogether, have a joint full observability, or collective observability [31], which means that each agent in Dec-MDPs observes a part of the state and the combined observations of all agents can uniquely identify the overall state. This is in contrast with multiperiod MDPs [32], [33], where each agent, individually, already has full observability allowing the individual observation of an agent to identify the overall state unambiguously. Due to agents’ local observability of the environment, each agent in Dec-MDPs cannot access the global state/reward information to make a reasonable decision. The decentralized decision makings in Dec-MDPs thus bear an extremely high complexity (i.e., NEXP-complete).
to achieve a globally optimal policy, even in the 2-agent case, versus the P-completeness of a single MDP and multiagent MDPs [28].

III. PROBLEM DESCRIPTION AND DEFINITIONS

This section first gives a brief description of the robot navigation problems, which are a representative domain for loosely coupled MASs. The concept of agent independence is then formally defined for the purpose of inducing the proposed approach.

A. Robots Navigation Problems

Robot navigation problems (or referred to as grid world puzzles) are fundamental research issues in the AI community. Fig. 2 shows three domains of robot navigation problems. In each of these domains, two robots $R_1$ and $R_2$ are navigating in an environment, each trying to reach its own goal $G_1$ and $G_2$, respectively (In Fig. 2(b), $G_1$ and $G_2$ are in the same grid denoted by $G$). Each robot has four actions, i.e., “Move East”, “Move South”, “Move West” and “Move North”. The action space is therefore represented as all the joint actions of both robots. Each robot’s state is its location in the environment. The state space is therefore represented as all possible joint locations of both robots. The decision-making process can be modeled as Dec-MDPs, in which each robot has to make a decision independently to achieve an optimal global policy through its local observation (individual action and state). Each robot can only have a local observability of the environment, which means the robot cannot observe the other robot’s location to determine the overall state of the Dec-MDPs. Furthermore, both robots are coupled with a shared global transition and reward function, but neither robot can access such global transition and reward functions on its own. A robot thus cannot explicitly determine the effect of its individual decision-making on the environment in order to take a reasonable action.

As can be seen from Fig. 2, interactions between the robots do not occur very often in most states of the environment. This means that generally each robot can make its own decision without regard to the other robot’s state and/or action but in certain specific situations (e.g., near the doorway), the robots are tightly coupled, and thus need to coordinate with each other for better performance. The sparse interactions between the robots can cause varying levels of independence in the states so that coordinated behavior is only heavily required in some local parts of the environment. Intuitively, the robots have higher levels of independence in the states that are further away from the doorway/entrance than those in the states nearer to the doorway/entrance. This is because that robots are more prone to have collisions near the doorway/entrance so that a robot’s transition function and/or reward function may heavily depend on the other robot. By exploiting the independence between robots, it is thus possible to decompose the general decision-making process into two distinct sub-processes. One is to let each robot make its own decision individually and completely disregard the existence of the other robot when there is an independence between robots. The other is to let robots coordinate their behaviors with each other by considering the situation of the other robot when interactions (in terms of coordination) are inevitable. Through this decomposition, the original problem can be more easily solved as the uncertainties caused by incomplete information can be greatly reduced.

Many approaches have been proposed to solve robot navigation problems by exploiting the independence of robots [34], [35], [36], [23], [22]. These approaches simplify the problem by predefining the dangerous areas for the robots [35], [34], or by imposing some assumptions on each robot, either its individual optimal policy [23], or the full observability of the other robot’s state [22] and action [36]. This paper, however, proposes an approach which enables the robots to devise an efficient coordinated policy without prior knowledge of either the domain structure or the robot itself. Before introducing this approach, it is necessary to give a formal definition of agent independence in order to clarify how a general Dec-MDPs problem can be decomposed efficiently by exploiting agent independence.

B. Formalization of Agent Independence in MASs

As can be seen in Fig. 2, the agents (robots) can have certain levels of independence (in terms of state transitions and rewards) so as to allow a lower computational complexity for solving the Dec-MDPs problem. Much work has been done to exploit these kinds of independence between agents in order to achieve an efficient decision-making in Dec-MDPs. In [37], a specific Dec-MDPs model called transition-independent Dec-MDPs [37], [38] is identified, in which the overall transition function $P$ can be separated into $n+1$ distinct functions $P_0,...,P_n$ as $P = P_0 \prod_{i=1}^{n} P_i$, where each $P_i$ ($i \in [1,n]$) stands for an individual transition function of agent $i$ and $P_0$ is an initial transition component. For any next state $s'_i \in S_i$ of agent $i$, $P_i$ is given by $P_i(s,a,s') = P_i(s_1,a_1,s_1')$. In other words, the next local state of each agent is independent of the local states of all other agents, given this agent’s previous local state, local action and the external system component ($S_0$). It has been shown that the decision-making complexity in transition-independent Dec-MDPs is NP-complete, which is a significant reduction compared to that of general Dec-MDPs models [38]. It is argued, however, not all multiagent domains are fully transition-independent [15]. Further more, due to the shared reward component, it is still a non-trivial task to solve transition independent Dec-MDPs.

Similarly, a Dec-MDPs model can also be reward-independent where the joint reward function $R$ is represented as a function of individual reward functions $R_1,...,R_n$ by $R(s,a) = f(R_1(s_1,a_1),...,R_n(s_n,a_n))$ [5], [29]. It was recently shown that a reward independent Dec-MDPs problem retains NEXP-complete complexity [29], [39]. However, when associated with transition independence, reward independence implies that a Dec-MDPs model can be decomposed into $n$ independent MDPs, each of which can be solved separately. The complexity of reward-independent Dec-MDPs thus becomes the same P-complete as standard MDPs.

General transition/reward-independent Dec-MDPs decompose the overall function $P/R$ into each agent’s individual function $P_i/R_i$, which only depends on each agent’s local state $s_i$ and individual action $a_i$. Since an agent is potentially affected by other agents, the agent’s local state and individual action usually cannot fully determine its individual function $P_i/R_i$. Thus, in [35], an agent’s individual function is defined on the parameters of the overall state and action of all the agents. The individual function can then be decomposed into a local individual component based on the agent’s local information $s_i$ and $a_i$, and an interaction component based on all the agents’ information $s$ and $a$. The factorization in [35] is based on agents’ prior knowledge about the domain structure in terms of the location of interaction states where agents mutually affect
each other’s transitions/rewards. Outside these interaction states, a full independence is assumed so that all the agents can make decisions individually. In other words, a state is identified as one in which the agents are either completely independent or completely dependent on each other. This handling of independence is always intractable due to the complexity of Dec-MDPs and the lack of prior knowledge about the domain structure. To better reflect the uncertainties of agent independence, the transition/reward independent degree is introduced in the factorization of individual functions as given by Definition 1.

Definition 1 (Transition/Reward Independent Degree): Transition/Reward Independent Degree $\mu/\nu \in [0,1]$ is a value to signify the extent of transition/reward independence of agent $k$ regarding to the remaining agents, such that agent $k$’s individual transition function $P_k((s_k, s_{-k}), (a_k, a_{-k}), (s'_k, s'_{-k}))$ can be decomposed by Eq. (3) and its individual reward function $R_k((s_k, s_{-k}), (a_k, a_{-k}))$ can be decomposed by Eq. (4).

$$P_k((s_k, s_{-k}), (a_k, a_{-k}), (s'_k, s'_{-k})) = \mu P^k((\hat{s}_k, a_k, \hat{s}'_k)) + (1 - \mu) P^k((s_k, s_{-k}), (a_k, a_{-k}), (s'_k, s'_{-k})) \quad (3)$$

$$R_k((s_k, s_{-k}), (a_k, a_{-k})) = \nu R^k((\hat{s}_k, a_k)) + (1 - \nu) R^k((s_k, s_{-k}), (a_k, a_{-k})) \quad (4)$$

In Definition 1, $P^k((\hat{s}_k, a_k, \hat{s}'_k)$ is agent $k$’s local individual transition function that depends only on agent $k$’s local information, and $P^k((s_k, s_{-k}), (a_k, a_{-k}), (s'_k, s'_{-k})$ is agent $k$’s interactive transition function that depends on the overall information of all agents. $R^k((\hat{s}_k, a_k))$ and $R^k((s_k, s_{-k}), (a_k, a_{-k}))$ are agent $k$’s local individual reward function and interactive reward function, respectively.

Transition/reward independent degree $\mu/\nu$ signifies the uncertainties of independence between agent $k$ and the remaining agents in terms of transition/reward function. It is assumed that if $P^h > P^l$, then $R^h > R^l$ holds, and vice versa. This means that if an agent is more likely to determine its transition function through its local information, then the agent is also more likely to determine its reward function through its local information. Agent $k$’s overall independence with the remaining agents, however, is signified by the value function $Q$ (see Eq. (1)). The independence of agents can thus be formally defined by Definition 2.

Definition 2 (Independent degree): An independent degree $\xi \in [\min(\mu, \nu), \max(\mu, \nu)]$ is a value that signifies the extent of independence of agent $k$ regarding to the remaining agents, so that the individual value function can be decomposed according to Eq. (5), where $Q^k$ and $Q^l$ are the expected local and interaction value function of agent $k$, respectively.

$$Q_k((s_k, s_{-k}), (a_k, a_{-k})) = \xi Q^k((\hat{s}_k, a_k)) + (1 - \xi) Q^l((s_k, s_{-k}), (a_k, a_{-k})) \quad (5)$$

Lemma 1: The independent degree $\xi$ is well-defined, that is, it always exists and is unique.

Proof: Let $a = (a_k, a_{-k}), s = (s_k, s_{-k})$. For each value function of agent $a_k$, $Q_k(a, s)$ depends on agent $a_k$’s transition function $P_k$ and reward function $R_k$. The other agents’ influence on agent $a_k$ is embodied in the common interaction components of transition function $P^l$ and reward function $R^l$. Thus,

$$Q_k(s, a) = R_k(s, a) + \gamma \sum_{s'} P_k(s, a, s') \max_{a'} Q_k(s', a') + \gamma \sum_{s'} [P_k(s, a, s') \max_{a'} Q_k(s', a') + (1 - \mu) P^l(s, a, s') \max_{a'} Q_k(s', a')]$$
A coordinated MAL approach is proposed through dynamic adaptation of the independent degrees during the learning process so that coordination can be achieved with low computational complexity. This approach estimates the optimal global policy for each agent when the model is not well specified. A model is not well specified when the transition/reward functions as well as independent degrees are unknown to the agents. Note that, although the transition and reward functions can usually be defined beforehand for a Dec-MDPs model, an explicit specification of independent degrees is not straightforward due to the complexities and uncertainties in Dec-MDPs. Thus, an estimation of these independent degrees during learning must be made in order to approximate the optimal policy.

### A. The Principle of the Learning Approach

As illustrated above, a hidden independent degree always exists in a state of a Dec-MDPs model, signifying the necessity of coordination between agents. A higher independent degree means that an agent is more independent from the remaining agents and can learn to build a model of the environment with higher certainty using only its individual information. In this situation, an agent can thus conduct an independent learning process by disregarding the existence of other agents. More specifically, when agent $i$ has an independent degree of $\xi_i^k$ in state $s_i^*$, it will apply independent learning with a probability of $\xi_i^k$. Otherwise coordinated learning can be conducted to combine other agents’ information for decision-making with a probability of $1-\xi_i^k$. The coordinated learning process can be carried out by assuming that each agent has limited observability of the environment, that is, $Distance(s_i,s_{-i}) \leq R : P[S(t) = s(O(t) = o_i) = 1$, where $S(t) = (s_i,s_{-i})$ is the joint state of all agents, $o_i$ is the individual observation of agent $i$, and $R$ is the agent’s perception distance. This means that only when other agents are in the perception distance of agent $i$ (i.e., $Distance(s_i,s_{-i}) \leq R$) can agent $i$ recover the joint state unambiguously through its individual observation. Equipping each agent with limited observability of the environment is reasonable because in many real-life applications agents can observe each other only when they are spatially close [35]. Here, the coordinated learning process is assumed to impose no cost on the agents, that is, whether or not to coordinate with other agents is completely determined by the value of independent degrees and this process has no probability to fail. The fact that an agent always succeed in activating the coordinated learning process, however, does not mean that the agent can successfully receive other agents’ information all the time since agents are only endowed with limited observability of the environment. An agent starts with an initial belief that there is no dependent relationship with the other agents. The agent then adapts the independent degrees dynamically during learning to achieve an efficient coordinated policy. This policy can make a trade-off between the independent learning process and the coordinated learning process to achieve computational savings and to mitigate the uncertainties of the learning environment. Through the dynamic trade-off between these two processes, agents can achieve efficient learning at the expense of minimum computational consumption. The sketch of this learning approach is given by Algorithm 1, where Line 4 indicates the dynamic trade-off process, which is determined by the explicit independent degrees, and Lines 5-8 show how to adapt the independent degrees to best reflect the real independence of the agents.

As agents are learning through trial-and-error interactions with the environment, the only available information to an agent is the agent’s learning experience and the immediate rewards from the environment. When a conflict occurs (i.e., penalized reward is received) in a state, the agent recognizes that its decision may depend on other agents’ decisions as well (i.e., coordination might be required) in the corresponding state. The estimated independence thus can be updated by considering both the potential contributions from other domain states and the most direct causes of the conflict based on the agents historical experiences.

More specifically, a penalized reward received in state $s$ is implicitly contributed to by all other states in the domain, with the extent of the effect determined by the difference between these domain states and state $s$. Let $\pi$ be the attribute vector that describes state $s$. The similarity $\rho(s,s')$ between two local states $s^*$ and $s'$ can be given by $\rho(s,s') = ||s - s'||$. In the robot navigation problem, which is modeled as Dec-MDPs, each robot has local observability of the environment so that the robot can accurately locate itself in the environment (e.g., through the robot’s sensor equipment to gauge its coordinates in the environment). The similarity between two states (locations) thus can be easily calculated. Based on the similarity of states, the cause of a penalized reward can be diffused to other states through a diffusion function as defined below.

**Definition 3 (Diffusion Function):** A diffusion function of reward $r$ in state $s^*$, $f_{s^*}(s) \in (0,1)$, is a Gaussian like function, and can be defined by Eq. (11), in which $s^*$ is the state where reward $r$ is received, $s$ is a local state of the agent in the domain and $\rho(s,s^*)$ is the similarity between state $s^*$ and state $s$.

$$f_{s^*}(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\rho(s,s^*)^2}$$ (11)

The diffusion function is a valid representation to reflect the contribution of each state $s$ in the domain to the penalized reward received in conflicting state $s^*$. When a reward $r$ is achieved in state $s^*$, $f_{s^*}(s)$ has the highest value signifying that reward $r$ is mainly caused by state $s^*$, and those states having the same similarity with the conflicting state $s^*$ are assumed to play the same role of causing the reward in state $s^*$. As the similarity between $s$ and $s^*$ decreases, state $s$ has a lower effect on causing reward $r$, which is reflected by the lower value of $f_{s^*}(s)$.

In many cases, the roles of the states having the same similarity with a conflicting state can be different. For example, when an agent detects a conflict in state $s^*$, it wants to determine the roles of the neighboring states of state $s^*$, say $s_1$, $s_2$, $s_3$, and $s_4$, to cause this conflict. If the past statistical information indicates that the agent usually transits from states $s_1$ and $s_2$ to $s^*$, causing the conflict in state $s^*$, but in some cases it transits from state $s^*$ to states $s_3$ and $s_4$, it is obvious that states $s_3$ and $s_4$ are similar with states $s_1$ and $s_2$, but are not the causes of the conflict in $s^*$. Thus, agents should assign credit to those states that are really responsible for the resulting conflict. To solve this problem, Eligibility Trace (ET) is introduced in order to make a temporary record of the occurrence of an event, i.e., a penalized reward is received. ET is one of the basic mechanisms of reinforcement learning to speed up the learning.
process. For example, in temporal difference TD(λ) algorithms [26], λ refers to the use of an eligibility trace. Unlike TD(λ), where ET is updated at every step during learning to boost learning efficiency, in the proposed approach, ET is updated only when an event (i.e., a conflict) occurs. Let $S'$ be the state trajectory that causes an event and $q_i^t$ be the ET value of agent $i$ in state $s_i^t$, then for each state $s_i^t \in S_i$, the ET value can be updated by Eq. (12), where $\gamma \in [0, 1)$ is a discount rate and $\lambda \in [0, 1]$ is the trace-decay parameter.

$$
e_i^t(t + 1) = \begin{cases} \gamma \lambda q_i^t(t) + 1, & s_i^t \in S_i^t; \\ \gamma \lambda q_i^t(t), & \text{else.} \end{cases} \quad (12)$$

Based on the diffusion function and eligibility trace, a penalized reward $r_i(t)$ signifying the necessity of coordination can be diffused to all the states that are potentially eligible for this reward. The independent degree can thus be adjusted according to Eq. (13) and Eq. (14).

$$
x_i^t(t + 1) = G(x_i^t(t + 1)) \quad (13)$$

$$
\psi_i^t(t + 1) = \psi_i^t(t) + \beta (t) \phi_i^t(t)\psi_i^t(\psi_i^t(0) = 0) \quad (14)$$

In Eq. (13), $G(x)$ is a normalization function to map the value of $x$ to interval $[0, 1]$, with a lower value of $x$ corresponding to a higher value of $G(x)$. In Equation 14, $\psi_i^t(t + 1)$ is a value to signify the necessity of coordination, $s_i^t(t)$ is the state resulting in the penalized reward $r_i(t)$, and $q_i^t(t)$ is the ET value of agent $i$ in state $s_i^t$.

### B. An Explicit Learning Algorithm

An explicit Q-learning based algorithm derived from the above approach (refer to Algorithm 1) is proposed in Algorithm 2 to solve robot navigation problems from the perspective of agent (robot) $i$. Although other learning techniques, such as the on-policy TD control method Sarsa, Actor-Critic methods [26], are also suitable, Q-learning has been chosen as an illustration because of its robustness, relative simplicity and widely successful applications in a number of multiagent domains. The current study focuses on two-agent scenarios where agent $i$ and agent $j$ are learning concurrently in the same environment. Although two-agent scenarios greatly simplify general robot navigation problems, most current research into coordinated MAL in robot navigation problems still focuses on this setting because two-agent scenarios already encompass the main challenges encountered in general MAL problems, that is, concurrent learning dynamics, each agent’s local observability, and limited or no prior knowledge concerning the domain or the agents. The main contribution of this paper is thus the proposed idea that an explicit quantified representation of the agent independence can be built up for an efficient coordinated learning, without requiring the assumptions which are needed in most existing approaches.

In Algorithm 2, $Q_i(s_i, a_i)$ is a single-state Q-value table for agent $i$, $Q_j(s_j, a_j)$ is a joint-state Q-value table for both agent $i$ and agent $j$, $x_i^t \ (k = 1, ..., m)$ are the values of the independent degree and eligibility trace, respectively, when agent $i$ is in state $s_i^t$, and trajectory_list is a list to store the trajectory (i.e., the state transition history) of the agent. Agent $i$ decides whether to coordinate with agent $j$ based on the independent degree in a state (Line 4). If agent $i$ chooses action coordinate (Line 5), it will activate its perception process to determine the local state information of agent $j$ through its limited observability of the environment (Line 6-11). Otherwise, agent $i$ chooses its action based on single-state Q-value table $Q_i(s_i, a_i)$ (Line 12). The perception process can be carried out either through agent $i$’s limited observing capability or by using explicit communication with $j$. In robot navigation problems, a robot can rely on its sensor to locate the other robot or send a message to require the other robot to divulge its location [34]. Agent $i$ in the perception process, however, does not necessarily succeed in accessing agent $j$’s state information as this success is environment-dependent, which means that only when agent $j$ is really located within the perception distance of agent $i$ can agent $i$ receive agent $j$’s state information (Line 6). After the successful perception process, agent $i$ makes use of the local state information from agent $j$ to choose its action based on the joint-state Q-value table $Q_i(s_j, a_j)$ (Line 7-10). Otherwise, agent $i$ makes its decision based only on single-state Q-value table $Q_i(s_i, a_i)$ (Line 11). After each transition (Line 13), agent $i$ updates its Q values (signifying how to coordinate with agent $j$) (Lines 14-17) and independent degrees (signifying when coordination is beneficial) (Line 18), respectively.

### Algorithm 2: Coordinated learning algorithm for agent $i$

1. Initializes $Q_i(s_i, a_i), Q_j(s_j, a_j) \rightarrow 0, \ z_i^t(t) \rightarrow 1, \ z_j^t(t) \rightarrow 0$, trajectory_list $\leftarrow \emptyset$.
2. For each episode $n (n = 1, ...)$ do
   3. For each step $t (t = 1, ...)$ do
      4. Generates a random number $\tau, \tau \in [0, 1]$;
      5. If $\xi(s_i) \leq \tau$ then
         6. PerceptionFlag $\leftarrow$ True and read $s_j, a_j \leftarrow (s_j, a_j)$;
         7. If $s_j$ not in table $Q_j(s_j, a_j)$ then
            8. Adds $s_j$ to table $Q_j(s_j, a_j)$;
            9. Selects $a_i(t)$ from $Q_i(s_i, a_i)$;
         10. Else selects $a_i(t)$ from $Q_i(s_i, a_i)$;
         11. trajectory_list.add($s_i(t)$, transit to state $s_j$ and receive $r_i(t)$); 12. Else selects $a_i(t)$ from $Q_i(s_i, a_i)$;
         13. trajectory_list.add($s_i(t)$, transit to state $s_j$ and receive $r_i(t)$);
         14. If $\xi(s_i) \leq \tau$ and PerceptionFlag $= \text{True}$ then
            15. Updates $Q_i(s_j, a_j) \leftarrow Q_i(s_j, a_j) + \alpha[r_i(t) + \gamma \max_{a_i} Q_i(s_j^t, a_i^t)]$;
            16. Else
               17. Updates $Q_i(s_i, a_i) \leftarrow Q_i(s_i, a_i) + \alpha r_i(t) + \gamma \max_{a_i} Q_i(s_i^t, a_i^t)$;
               18. Call Algorithm 3 to adjust $z_i^t(t), s_i \leftarrow s_i^t$.
3. End For
4. End For
5. End For

### Algorithm 3: Adjusting the independent degree $z_i^t(t)$

1. Input: step time $t$, $z_i^t(t - 1), (k = 1, ..., m)$, trajectory_list;
2. If a collision occurs in state $s_i^t(t)$ then
   3. $\max_{x_{temp}} = 0$;
   4. For each $s_i^t \in S_i$ do
      5. If $s_i^t \in \text{trajectory_list}$ then
         6. $z_i^t(t) \leftarrow \gamma^t z_i^t(t - 1) + 1$;
         7. Else $z_i^t(t) \leftarrow \gamma^t z_i^t(t - 1)$;
         8. $\psi_i^t(t) \leftarrow \psi_i^t(t - 1) + \frac{e^{(\frac{1}{\gamma} - \frac{1}{\psi_i^t(t - 1)})}}{\psi_i^t(t)}$;
         9. If $\psi_i^t(t) \geq \max_{x_{temp}}$ then $\max_{x_{temp}} = \psi_i^t(t)$;
      10. End If
   11. End For
   12. End If

The dynamic adaptation process of independent degrees is given by Algorithm 3, where $s_i^t(x_k, y_k)$ is a local state with coordinate $(x_k, y_k)$, $x_u, y_u$ is the coordinate of the conflicting state $s_i(t)$, and $\psi$ is a value to signify the necessity of coordination (see Eq. (14)). As the similarity between $s_i^t$ and $s_i(t)$ decreases (i.e., the Euclidean distance of these two states increases, which is calculated as $(x_k - x_u)^2 + (y_k - y_u)^2$, the value of the diffusion function decreases, indicating that state $s_i^t$ has a lower impact on the cause of the conflict in state $s_i(t)$ (Line 7). The value of the diffusion function is combined with the eligible trace value $z_i^t(t)$ to indicate which previous states are really responsible for the conflict in state $s_i(t)$. As for the normalization function $G(x)$ in Eq. (13), the maximum $\psi$ (denoted as $\max_{x_{temp}}$) after each episode is chosen (Line 8) to let the
values of $\psi$ in all other states be compared with $\max_{i}\psi$ so as to confine $\psi$ to $[0, 1]$ (Line 10). This way of handling the normalization function $G(x)$ is used as an illustration here, but a number of other methods can be adopted to define a concrete normalization function.

![Fig. 3. Illustration of the relation between individual single-state Q-value $Q_i(s_i, a_i)$ and the joint-state Q-value $Q_j(s_j, a_j)$.](image)

Several discussions are laid out as follows.

(1) As can be seen from Algorithm 2 (Line 15), the update of joint-state Q-value $Q_j(s_j, a_j)$ utilizes the estimates of individual single-state Q-value $Q_i(s_i, a_i)$ in the next step. The relation between individual single-state Q-value $Q_i(s_i, a_i)$ and the joint-state Q-value $Q_j(s_j, a_j)$ is illustrated in Fig. 3. An agent chooses its action from individual Q-value $Q_i(s_i, a_i)$ and updates this value by using the maximum estimated Q-value of the state in the next step (say, state 7) with a probability of $\xi$. Otherwise, the agent chooses an action from joint-state Q-value $Q_j(s_j, a_j)$ with a probability of $1 - \xi$ and updates the joint-state Q-value (say, joint state 4-1) by using the maximum individual Q-value $Q_i(s_i, a_i)$ in the next step. From Fig. 3, we can see that the joint-state Q-value of different state/actions in table $Q_j(s_j, a_j)$ is independent from each other, and the joint-state Q-values $Q_j(s_j, a_j)$ only determines one-step coordinated behavior of an agent. This implies that if agents are loosely coupled (as in robot navigation problems) and conflicts between agents are only confined to some particular parts (usually accounting for a small proportion) of the environment, the joint-state Q-value table can be very sparse. Therefore, learning for an optimal global policy in this situation will be computationally inexpensive (almost the same with directly learning with individual single-state Q-values).

(2) In the robot navigation problem, there are multiple optimal policies to avoid an uncoordinated joint action in the conflicting states. For example, when both robots come to the doorway in the HG domain, either robot choosing to pass the doorway while the other makes a detour can constitute an optimal joint action. Both robots should decide separately which joint action to choose to avoid mis-coordination. This type of difficulty is called an “equilibrium selection problem” and it is inherent in any coordinated learning process. In the proposed approach, the equilibrium selection process is carried out by using only the agents’ joint-state information. This is in contrast with some other approaches which use joint-action information [40], or joint-state-action information [36], or simply a random action selection mechanism [22], to conduct the coordinated learning process. It is obvious that learning from the joint-action information of both agents can lead to a coordinated joint action. However, as shown by the experimental results, learning directly from the joint-state information of both agents can also produce such a coordinated joint action. Importantly, this implies that joint-state information alone can be enough for agents to learn coordinated behaviors. Learning directly from the joint-action information or requiring further such information will thus be redundant in an equilibrium selection process.

(3) In the proposed approach, agents adapt their independence every time an event (a conflict) occurs. This means the approach is fully event-driven, and might be potentially applied to other similar domains as long as a significant conflicting event can happen due to the dependence of the agents. Furthermore, an agent updates its Q values (signifying how to coordinate its behavior with the other agent) and independent degrees (signifying when coordination is beneficial) concurrently during the learning process, which means that the learning approach is fully on-line. This is significantly different from other existing approaches in which agents need a predefined period to learn the situations when coordination is beneficial so that coordinated learning can be carried out afterwards [36], [20].

(4) As stated above, this paper restricts discussions to two-agent scenarios. Two solutions are possible to extend the approach to scenarios involving more than two agents. One is to differentiate the identity of each agent and let an agent estimate the independent degrees with each of other agents individually. In this case, a joint Q-value table is built for each pair of agents during the coordinated learning process. The other solution is simply to disregard agents’ identities and build a joint Q-value tables for all the agents. Only an overall independent degrees is estimated during learning in this case. This solution is based on the idea that when an agent detects a conflict, this agent might not care who caused this conflict. The former solution seems to have the potential to achieve a better performance than the latter but costs a higher computational complexity.

V. EXPERIMENTAL STUDIES

Experimental studies have been carried out to test the effectiveness of the proposed approach, denoted as IDL (Independent Degrees Learning), to solve robot navigation problems. The aim of the experiment is to test whether an efficient coordinated policy can be learnt through dynamic adaptation of agent independence without the assumptions of agents’ prior knowledge about the domain structure or the agents’ full observability of the environment.

A. Experimental Setting

1) Benchmark Approaches: Because most existing approaches to coordinated MAL in loosely coupled MASs are based on certain pre-conditions (e.g., robots’ global observability or prior knowledge about the domain structure), a direct comparison of the proposed approach with these approaches is not applicable. Three other approaches are selected as benchmarks for comparison of learning performance.

IL (Independent Learning). In this approach, each agent treats the remaining agents simply as part of the environment, and learns its policy independently. The decision-making process is forcibly decomposed into $n$ separate MDPs. This results in a large reduction in the state-action representation. However, at the same time, a poor learning outcome might occur because of the lack of coordination. IL provides a valid perspective to study the so-called “moving target” effect complicating general MAL problems. Despite the lack of guaranteed optimal performance, this approach has been applied successfully in certain cases [41], [42].

JSAL (Joint State Action Learning). This approach provides another extreme scenario opposed to IL. Agents either communicate freely with a central controller and select their individual actions according to those indicated by the central controller, or have full observability of the environment to receive the joint-state-action information of all agents to control the learning process synchronously. With sufficient learning periods, JSAL is capable of achieving an optimal performance as the overall decision-making process is considered as a single MDP, in which agents learn in a static environment. JSAL faces some tricky issues in MAL, including (1) the curse of dimensionality: the search space grows rapidly with the complexity of agent behaviors, the number of agents involved
and the size of domains; (2) limited observability and restricted communication capability: agents might not have access to the needed information for the learning update because they are not able to observe the states, actions and rewards of all other agents; and (3) slow convergence: it may take many steps to explore all joint actions for every state, resulting in a slow convergence to the optimal policy.

### 2) Parameter Settings

In this experiment, the four approaches are first tested in three small domains, as shown in Fig. 2. In these domains, the state space is relatively small and in some regions of the domains, independent degrees might be quite low, indicating that coordination is heavily needed in order to avoid conflicts. These approaches are then applied to some larger domains, as shown in Fig. 4 [34]. There are fewer interactions between robots in these larger domains than in the small domains in Fig. 2.

When robots collide with a wall, they rebound and stay where they were. If they collide with each other, both robots break down and are transferred back to their original states. The exploration policy adopted is the fixed $\epsilon$ – greedy policy with $\epsilon = 0.1$ to indicate a small probability of exploration. Learning rate $\alpha = 0.05$, discount factor $\gamma = 0.95$, trace-decay parameter $\lambda = 0.8$, perception distance $R = 2$ and rewards $r$ are given as follows: +20 for reaching the goal state, $-1$ for colliding with a wall and $-10$ for colliding with the other robot. All approaches are run for $N = 10,000$ episodes and average the last $N_{avg} = 2000$ episodes to compute the overall performance. To use approach CL, $\delta$ is set to be 95%. Two cases of $\delta$ are studied. In the first case of CL, which is denoted as $CL_1$, the scanning distance $R_1$ is set to 2, while in the second case, which is denoted as $CL_2$, the scanning distance $R_2$ is set to 4. In both cases, the first 1000 (i.e., $N_{off} = 1000$) episodes are used to collect statistical information to determine the coordinated states in approach CL. All results are then averaged over $Run = 25$ runs. The values and meanings of parameter settings are listed in Table I for clarity.

#### B. Results and Analysis

Fig. 5 (a-c) shows the learning dynamics of the four learning approaches in small domains. We can see that approach JSAL can obtain an optimal reward because robots learn their policies based on joint state-action information. The independent learning approach IL, instead, can only achieve a very low reward due to the lack of coordination. That is, when a robot is learning independently, the robots cannot take into consideration the non-stationarity of the learning environment. Thus, robots must continuously “catch up” with the dynamic and adaptive environment, causing a sub-optimal performance much lower than that in approach JSAL. Robots using approach CL conduct their learning independently during the first 1000 episodes and collect statistical information to determine the coordinated states where coordination is most required. After that, the average reward dramatically increases as robots can use the joint state-action information to lower the probability of conflicts when they are in the coordinated states. The performance in approach CL, however, still remains at a sub-optimal level due to an over estimation caused by the optimistic estimation mechanism (for more detail, see [36]), which always chooses the highest Q value based on the available information. Robots using the proposed approach IDL also conduct their learning independently at the beginning of the learning process because no dependent relationships between robots are assumed beforehand. The robots then adapt the dependent relationships (through adapting independent degrees) dynamically during the learning process. In this way, coordination can be achieved when there are low independent degrees in some states indicating that robots are mutually dependent on each other for decision-making. As can be seen in Fig. 5 (a-c), the learning curves of approach IDL converge more quickly than those of approach JSAL at the beginning of learning process, and then gradually converge to the optimal values of approach JSAL. Robots using approach CL deliver perform better than robots using approach $CL_2$, even in the latter, robots have a longer scanning distance of $R = 4$ and learn based on joint state-action information in the coordinated states.

### Table I Parameters Setting

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.05</td>
<td>The learning rate</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.95</td>
<td>The discount factor</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.8</td>
<td>The trace-decay parameter</td>
</tr>
<tr>
<td>$\delta$</td>
<td>95%</td>
<td>Threshold to determine the coordinate states in $e$-greedy</td>
</tr>
<tr>
<td>$r_{coll}$</td>
<td>1</td>
<td>The reward of colliding into a wall</td>
</tr>
<tr>
<td>$r_{wall}$</td>
<td>-10</td>
<td>The reward of colliding with each other</td>
</tr>
<tr>
<td>$N_{avg}$</td>
<td>2000</td>
<td>The number of learning episodes used to collect statistical information to determine the coordinated states in approach CL.</td>
</tr>
<tr>
<td>$N_{off}$</td>
<td>1000</td>
<td>The number of off-line learning episodes for average results</td>
</tr>
<tr>
<td>$W$</td>
<td>2</td>
<td>The perception distance in approach IDL.</td>
</tr>
<tr>
<td>$R_1$</td>
<td>2</td>
<td>The perception distance in approach CL,1.</td>
</tr>
<tr>
<td>$R_2$</td>
<td>4</td>
<td>The perception distance in approach CL,2.</td>
</tr>
<tr>
<td>$N_{run}$</td>
<td>25</td>
<td>The number of Monte Carlo runs.</td>
</tr>
</tbody>
</table>

### Table II Performance of Different Learning Approaches in Small-Scale Domains (With 95% Confidence Intervals)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Approach</th>
<th>States</th>
<th>Actions</th>
<th>Q Values</th>
<th>Reward</th>
<th>Confidence</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>CL_1</td>
<td>22.3</td>
<td>5.71</td>
<td>124.48</td>
<td>11.20</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>CL_2</td>
<td>25.5</td>
<td>7.44</td>
<td>143.12</td>
<td>15.77</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>JSAL</td>
<td>31.6</td>
<td>9.08</td>
<td>229.06</td>
<td>18.41</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>IL</td>
<td>2.29</td>
<td>0.43</td>
<td>0.22</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Fig. 4. Large-scale robot navigation domains used in our experiment. R1, R2 are two robots and the original state of one robot is the other’s goal.
The overall results in small domains are given by Table II. To show the different computational complexity of these approaches, the state/action size and corresponding number of estimated Q values are also laid out. As can be seen from Table II, approach IDL reduces the computational complexity slightly compared with approach CL and significantly compared with approach JSAL. The significant reduction from approach JSAL is more desirable in large-scale domains where the computational complexity of JSAL is too high to be implemented, which can be verified by the results in large domains given later. In all small domains, the proposed approach IDL can achieve a reward that is much higher than the reward in approach IL and is quite near to the optimal reward in approach JSAL. This is because that approach IDL can capture the different levels of independence between robots for an efficient learning so that coordination can be achieved to avoid conflicts in certain parts of the environment. The results of collision percentage show that approach IDL decreases the likelihood of collision compared with approaches CL and IL. In approach JSAL, the optimal reward is a bit lower than 20 and robots may also have a low probability of collision. This is because of the stochastic exploration during the last 2000 episodes, when average performance is calculated. The steps of both robots to reach their own goals are calculated for those episodes when both robots do not collide with each other and reach the goal successfully. The results show that robots in approach IL always find the shortest path to their goals, which, in turn, causes the high probability of collision because they do not coordinate with each other when both come to potentially dangerous states with low independent degrees. On the contrary, approach JSAL receives the joint state-action information of both robots so that a safe detour strategy will be learnt by both robots to reduce the probability of collision, thus increasing the steps to the goals. The 95% confidence intervals of step number to goals also imply that approach JSAL is the most unstable approach, that is, the policy in approach JSAL can be affected by the stochastic learning process. This means if robots are “lucky” enough at the early exportation stage, they can learn a short path to the goals. On the contrary, a much longer path is learnt if the learning process is deviated from a better trajectory. Approach IDL, however, combines the merits of both approach IL and JSAL, allowing robots to find the shortest path to the goals with a higher certainty while only making small detours around the states where coordination is most needed. This is why the step number to goals in approach IDL is a bit higher than that in approach IL but much lower than that in approach JSAL.

Fig. 5. Learning dynamics of the four different approaches (To avoid redundancy, in (d) CMU domain, the learning curve of approach CL2 is not depicted as this curve overlaps with the learning curves of approach IL and CL1.).

![Graphs showing learning dynamics of four different approaches: HG, TTG, TR, and CMU domains.](image)

Fig. 6. An illustration of robots’ transitions in the HG domain using different learning approaches. The solid arrows represent one step transition between two states around the conflicting area and the associate number n indicates the corresponding n-th step of transition.

![Illustration of robots’ transitions in the HG domain.](image)
(R = 2) can determine coordinated states (i.e., the three states near the doorway) off-line and combine each other’s state and action information for coordinated learning in these states. That is why robot R₂ can learn to make a tour by transitioning to a point beyond the doorway (i.e., transition 1 in Fig. 6 (b)). However, because the new state of robot R₂ is not in the learnt coordinated states, robot R₂ cannot receive robot R₁’s information for coordination in this new state. Although most of the time, robot R₂ can learn to avoid conflict with robot R₁ by transitioning to the right side, in some cases, robot R₂ will still transit back to the coordinated states (i.e., transition 2 in Fig. 6 (b)), causing a collision near the doorway. This probability of collision, however, is greatly reduced compared to that in approach IL. Fig. 6 (c) shows the most likely occurring transitions when robots use approach IDL. As robots can learn a value of independent degree in each state, they can combine each other’s state information for coordination during each transition (with a different probability indicated by the value of the corresponding independent degree). As can be seen from Fig. 6 (c), robot R₂ can successfully learn a big enough detour by transitioning from the right-up side of the doorway to the right side (i.e., transition 3 in Fig. 6 (c)), which implies that the robots can learn a collision-free strategy together. Comparing the transitions and corresponding performances between approach CL and IDL, we can also gain a better understanding of the benefit of taking into account the uncertainties of independence between agents (robots) in approach IDL (In approach CL, a fully dependence between robots is assumed in those learnt coordinated states, while in the remaining states robots are fully independent). Lastly, as stated above, robots using approach JSAL learn their policies in a centralized manner. Searching among a large number of state-action pairs causes robots to learn large-detour strategies to avoid potential collisions. A simple illustration is given by Fig. 6 (d).

![Fig. 7](image-url)

Fig. 7. Values of 1 – ξ in small-scale domains (robot R₁).

Fig. 7 shows the estimated values of 1 – ξ in three small domains from the perspective of robot R₁. These values signify the dependent degrees between the robots. A higher value of 1 – ξ means that a robot is more dependent on the other robot for decision-making. As expected, the values of 1 – ξ in the conflicting states, that is, the areas near the entrance or doorway where robots are more inclined to collide with each other, are much higher than the values of those “safe” states where robot R₁ can generally choose its action without regard to robot R₂. It is also interesting to note that the dependent degrees in domain TR are comparatively lower than those in the other two domains. This can be easily explained by the fact that robots in domain TR are not so tightly coupled because there is more than one route for the robots to reach their own goals. This is the opposite of domain HG and TTG, where only one route is available, so that the robots are heavily dependent on each other for decision-making, especially in the conflicting states, consequently causing high values of 1 – ξ near these states. From Fig. 7, we can see that the independent degrees can indeed be learnt to reflect the real situations of independence between robots.

Table III gives the overall performance of these four learning approaches in large domains. The fundamental difference between these larger domains and the smaller domains in Fig. 2 is that interactions in these large domains occur much less frequently than in the smaller domains. Robots have high independent degrees in most states, so explicit coordination actions are not often required. This explains why in some domains such as MIT, CIT, SUNY and CMU, the independent learning approach IL can already achieve a very good performance. Even in these domains, approach IDL can still improve the performance further with the exception of domain SUNY where coordination is not necessary at all. It is noted that the minor difference between the reward 18.80 in approach IDL and 19.10 in approach IL in domain SUNY is caused by the extra exploration introduced by the coordinated learning process in approach IDL. In all other domains where coordination is more necessary, the benefit that this coordinated learning process brings outweighs the uncertainties it causes. Another interesting finding in Table III is that as the state and action sizes grow, the performance of JSAL decreases. This is because approach JSAL searches the whole state-action space, thus robots cannot learn an optimal policy in 10,000 learning episodes. For example, in the CMU domain, there are 17, 689 × 16 = 283, 024 Q values to be estimated. Fig. 5 (d) plots the learning process of the learning approaches in the CMU domain, from which we can see that approach JSAL converges too slowly to reach an optimal value. The proposed learning approach, however, only needs to consider the joint-state information when the independent degrees signify that it is necessary. In this way, the search space is reduced substantially compared with approach JSAL.

In summary, the proposed approach IDL outperforms the uncoordinated approach IL by considering coordination only when the learnt robot independence indicates that this coordination is necessary. On the other hand, approach IDL reduces the computational complexity considerably and enables robots to learn a shorter path to the goal with a higher certainty than approach JSAL. Furthermore, approach IDL performs better than the state-of-the-art approach CL in considering the uncertainties of robot independence and without acquiring joint-action information in coordinated learning. Experimental results showed that approach IDL could efficiently learn the independent degrees in different states in order to quantify agent independence so that an efficient coordination policy could be devised with low computational complexity.
of which is restricted to the states and actions relevant to some particular agents. Coordination-graph based approaches \[19\], \[24\] took advantage of an additive decomposition of the joint reward function to local value functions, each of which is related only to a small number of neighboring agents, to efficiently construct jointly optimal policies. Distributed value functions based learning approach \[21\] exploited the dependence between networked agents so that each agent can learn a value function to estimate a weighted sum of future rewards of all the agents in a network. All the above approaches, however, focused on solving coordinated MAL when an explicit representation of agent independence was given beforehand. The agent independence is represented either through a predefined decomposition of the main task or a fixed interaction network topology. Our work differs from these approaches in that no explicit representation of the agent independence is assumed so that agents should build such a representation from experience.

Many approaches have been proposed with an assumption of agents’ full observability of the domain state for efficient coordinated learning. An approach called sparse tabular Q-learning \[45\] was proposed to learn joint action values on those coordinated states where coordination is beneficial. The action space is reduced significantly because agents can learn individually without taking into account the other agents in most situations, but only need to conduct coordinated learning in the joint action space when dependence between agents exists. These coordinated states where agents are mutually dependent, however, need to be specified beforehand and the agents are assumed to have prior knowledge of these states. Kok \textit{et al.} further extended their approach to enable agents to coordinate their actions when there are more complicated dependencies between agents \[24\], and used statistical information about the obtained rewards to learn these dependencies \[20\]. All these studies are based on the framework of collaborative multiagent MDP \[46\], in which agents make decisions by using the joint state information. This full observability of world state for each individual agent is not assumed in our work as the problems studied here are framed as Dec-MDPs, where each agent has only local observability so that the joint state cannot always be determined through an agent’s local information.

Some approaches have been developed in recent years with the aim of learning when coordination is beneficial in a loosely coupled MAS (particularly in robot navigation problems). De \textit{et al.} \[22\] proposed a solution to coordinated learning problems called Second. This approach decouples a MAL process into two separate layers. One layer learns where it is necessary to observe the other agent and the other layer adopts a corresponding technique to avoid conflicts. This approach, however, assumed that agents have full observability of the environment, which is undesirable in many systems. In our approach, agents only have limited observability of the environment and a joint-state learning process is conducted during the coordinated learning process. Spaan \textit{et al.} \[35\] introduced a model for solving the coordination problem in loosely coupled MASs called Interaction-driven Markov Games (IDMG). In IDMG, the states where agents should coordinate with each other are specified in advance and a fully cooperative Markov Game is defined in these coordinated states such that agents can compute the game structure and Nash equilibria to choose their actions accordingly. IDMG is based on the game-theory solution to resolve the planning problem that requires the computation of multiple equilibria, which is computation-demanding. Later, Melo \textit{et al.} \[47\] proposed a two-layer extension of the Q-learning algorithm to enable agents to learn where coordination is beneficial by augmenting the action space with a pseudo-coordination action. In their approach, agents are able to learn a trade-off between the benefits arising from good coordination and the cost of choosing the pseudo-coordination action. As a result, agents can learn to use the pseudo-coordination action in states only when it is necessary. The learning performance in \[47\], however, can be affected by the cost of choosing the pseudo-coordination action. In our approach, the coordination action is activated by the independent degrees, which carry the historical information indicating that such a coordination action will potentially bring more benefits. Our approach is fully event-driven, and the learning performance is not affected by the coordination action. De \textit{et al.} \[23\] proposed an algorithm called CQ-learning to enable agents to adapt the state representation in order to coordinate with each other. CQ-learning, however, depends on the assumption that each agent has already had an optimal individual policy so that every agent can have a model of its expected rewards. In this way, those states where the expected rewards differed significantly from the observed rewards are marked as dangerous states in which the other agent’s state information should be considered for decision-making. The assumption of individually learnt optimal policy is not required in our approach. Recently, an off-line learning approach was proposed to solve the robot navigation problems \[36\], \[43\]. This off-line learning first adopts a statistical method to detect those states where coordination is most necessary. An \textit{optimistic estimation} mechanism is then applied to coordinate agents’ behaviors based on their limited observability of the environment. The approach proposed in this paper, instead, considers the uncertainties of agent independence and enables agents to learn the necessity of coordination and coordinated behaviors concurrently. Moreover, our approach does not use any joint-action information in the coordinated learning process.

Achieving coordination in multiagent systems has also been extensively studied from the aspects of decentralized adaptive control systems. For example, a decentralized control strategy has been proposed in \[48\] that enables a group of robots to dynamically adapt their coverage so as to spread out over an environment while aggregating in areas of high sensory interests. A fully adaptive decentralized controller of robot manipulators for trajectory tracking was proposed in \[49\]. Many other studies (e.g., \[50\], \[51\], \[52\]) focused on decentralized adaptive control of synchronization (i.e., coordination) in complex multiagent systems with stochastic dynamics or interconnection uncertainties. Our work differs from all these work because the decentralized coordination problem in this paper is modeled as Dec-MDPs, which is against the work in decentralized adaptive control systems in which agents’ state equation is represented in the form of a differential/difference equation. In \[53\], a MARL approach was proposed to facilitate the coordinated learning process in grid-world domains when value functions are not shared among agents. This work, however, did not address learning where to coordinate by exploiting agent independence.

\section{VI. Related Work}

Much attention has been paid to the problem of coordinated MAL in recent multiagent research. Numerous approaches have been proposed to exploit the structural or networked dependence of agents for efficient coordinated learning. Hierarchical MAL approaches \[16\], \[44\] have been proposed to solve agent coordination problem where an overall task can be subdivided into a hierarchy of subtasks, each of which is restricted to the states and actions relevant to some particular agents. Coordination-graph based approaches \[19\], \[24\] took advantage of an additive decomposition of the joint reward function to local value functions, each of which is related only to a small number of neighboring agents, to efficiently construct jointly optimal policies. Distributed value functions based learning approach \[21\] exploited the dependence between networked agents so that each agent can learn a value function to estimate a weighted sum of future rewards of all the agents in a network. All the above approaches, however, focused on solving coordinated MAL when an explicit representation of agent independence was given beforehand. The agent independence is represented either through a predefined decomposition of the main task or a fixed interaction network topology. Our work differs from these approaches in that no explicit representation of the agent independence is assumed so that agents should build such a representation from experience.

Many approaches have been proposed with an assumption of agents’ full observability of the domain state for efficient coordinated learning. An approach called sparse tabular Q-learning \[45\] was proposed to learn joint action values on those coordinated states where coordination is beneficial. The action space is reduced significantly because agents can learn individually without taking into account the other agents in most situations, but only need to conduct coordinated learning in the joint action space when dependence between agents exists. These coordinated states where agents are mutually dependent, however, need to be specified beforehand and the agents are assumed to have prior knowledge of these states. Kok \textit{et al.} further extended their approach to enable agents to coordinate their actions when there are more complicated dependencies between agents \[24\], and used statistical information about the obtained rewards to learn these dependencies \[20\]. All these studies are based on the framework of collaborative multiagent MDP \[46\], in which agents make decisions by using the joint state information. This full observability of world state for each individual agent is not assumed in our work as the problems studied here are framed as Dec-MDPs, where each agent has only local observability so that the joint state cannot always be determined through an agent’s local information.

Some approaches have been developed in recent years with the aim of learning when coordination is beneficial in a loosely coupled MAS (particularly in robot navigation problems). De \textit{et al.} \[22\] proposed a solution to coordinated learning problems called Second. This approach decouples a MAL process into two separate layers. One layer learns where it is necessary to observe the other agent and the other layer adopts a corresponding technique to avoid conflicts. This approach, however, assumed that agents have full observability of the environment, which is undesirable in many systems. In our approach, agents only have limited observability of the environment and a joint-state learning process is conducted during the coordinated learning process. Spaan \textit{et al.} \[35\] introduced a model for solving the coordination problem in loosely coupled MASs called Interaction-driven Markov Games (IDMG). In IDMG, the states where agents should coordinate with each other are specified in advance and a fully cooperative Markov Game is defined in these coordinated states such that agents can compute the game structure and Nash equilibria to choose their actions accordingly. IDMG is based on the game-theory solution to resolve the planning problem that requires the computation of multiple equilibria, which is computation-demanding. Later, Melo \textit{et al.} \[47\] proposed a two-layer extension of the Q-learning algorithm to enable agents to learn where coordination is beneficial by augmenting the action space with a pseudo-coordination action. In their approach, agents are able to learn a trade-off between the benefits arising from good coordination and the cost of choosing the pseudo-coordination action. As a result, agents can learn to use the pseudo-coordination action in states only when it is necessary. The learning performance in \[47\], however, can be affected by the cost of choosing the pseudo-coordination action. In our approach, the coordination action is activated by the independent degrees, which carry the historical information indicating that such a coordination action will potentially bring more benefits. Our approach is fully event-driven, and the learning performance is not affected by the coordination action. De \textit{et al.} \[23\] proposed an algorithm called CQ-learning to enable agents to adapt the state representation in order to coordinate with each other. CQ-learning, however, depends on the assumption that each agent has already had an optimal individual policy so that every agent can have a model of its expected rewards. In this way, those states where the expected rewards differed significantly from the observed rewards are marked as dangerous states in which the other agent’s state information should be considered for decision-making. The assumption of individually learnt optimal policy is not required in our approach. Recently, an off-line learning approach was proposed to solve the robot navigation problems \[36\], \[43\]. This off-line learning first adopts a statistical method to detect those states where coordination is most necessary. An \textit{optimistic estimation} mechanism is then applied to coordinate agents’ behaviors based on their limited observability of the environment. The approach proposed in this paper, instead, considers the uncertainties of agent independence and enables agents to learn the necessity of coordination and coordinated behaviors concurrently. Moreover, our approach does not use any joint-action information in the coordinated learning process.

Achieving coordination in multiagent systems has also been extensively studied from the aspects of decentralized adaptive control systems. For example, a decentralized control strategy has been proposed in \[48\] that enables a group of robots to dynamically adapt their coverage so as to spread out over an environment while aggregating in areas of high sensory interests. A fully adaptive decentralized controller of robot manipulators for trajectory tracking was proposed in \[49\]. Many other studies (e.g., \[50\], \[51\], \[52\]) focused on decentralized adaptive control of synchronization (i.e., coordination) in complex multiagent systems with stochastic dynamics or interconnection uncertainties. Our work differs from all these work because the decentralized coordination problem in this paper is modeled as Dec-MDPs, which is against the work in decentralized adaptive control systems in which agents’ state equation is represented in the form of a differential/difference equation. In \[53\], a MARL approach was proposed to facilitate the coordinated learning process in grid-world domains when value functions are not shared among agents. This work, however, did not address learning where to coordinate by exploiting agent independence.

\section{VII. Conclusion}

This paper proposed an MAL approach to solve coordination problems by exploiting agent independence in loosely coupled MASs. The approach enables agents to learn an efficient coordinated policy through dynamic adaptation of the estimation of agent independence. The proposed approach requires neither prior knowledge about the structure of the domain (e.g., the location of coordination) nor assumptions about the learning agents (e.g., full observability or optimal individual policies). These features set our approach apart from most
existing approaches in related work and render it potentially suitable for wider practical applications. Experimental results showed that the proposed approach could guarantee a near-optimal performance in different scales of domains with low computational complexity.

In this paper, independent degrees are adjusted immediately when a conflict occurs. This mechanism can be improved by using a delayed updating rule to adjust the independent degrees based on a specified period of learning experiences. This issue is left for future research.

REFERENCES


IEEE TRANSACTIONS ON CYBERNETICS, VOL. XX, NO. XX, 2014


Chao Yu received his BSc degree from Huazhong University of Science and Technology, China, in 2007, MSc degree from Huazhong Normal University, China, in 2010, and PhD degree in computer science from University of Wollongong, Australia, in December 2013. He is now a lecturer in the School of Computer Science and Technology, Dalian University of Technology, China. His research interests include multiagent systems and multiagent learning, with their wide applications in modeling and solving various real-world problems.

Minjie Zhang received the BSc and the MSc degrees from Fudan University, China and the PhD degree in computer science from the University of New England, Australia. Now, she is a professor of computer science at the University of Wollongong, Australia. Prof. Minjie Zhang is an active researcher and has published more than 100 papers in the past 10 years. Her research interests include multiagent systems and agent-based modeling in complex domains. She is a senior member of IEEE.

Fenghui Ren received his MSc and PhD from the University of Wollongong in 2006 and 2010, respectively. He received his BSc from the Xidian University in 2003. Currently, he is an ARC DECRA Fellow and a lecturer in the School of Computer Science and Software Engineering at the University of Wollongong, Australia. Dr. Ren is an active researcher and published over fifty research papers. His major research areas include agent-based concept modeling of complex systems, data mining and pattern discovery in complex domains, agent-based learning, smart grid systems, and self-organizations in distributed and complex systems.

Guozhen Tan is now a Professor and the Dean of the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. He is also the Director of Engineering and Technology Research Center for the Internet of things and Collaborative Sensing, Liaoning province. He received his B.S. degree from Shenyang University of Technology, Shenyang, M.S. degree from Harbin Institute of Technology, Harbin, and Ph.D. degree from Dalian University of Technology, Dalian, China. He was a visiting scholar with the Department of Electrical and Computer Engineering of University of Illinois at Urbana-Champaign, IL, U.S., from Jan. 2007 to Jan. 2008. His research interests include the Internet of things (IOT), cyber-physical systems (CPS) VANET, Intelligent transportation systems (ITS), network optimization algorithms, etc. He was the recipient of the 2006 National Science & Technology Progress Award (second class) for his work in vehicle position and navigation, location-based service, traffic signal control, rapid responsiveness and processing for traffic emergency. He has been an Editor for the Journal of Chinese Computer Systems. He has been member of China Computer Federation (CCF), committee member of Internet Professional Committee of CCF committee member of Professional Committee of Software Engineering of CCF, and committee member of Professional Committee of High Performance Computing of CCF.