



Model-free reinforcement learning for motion planning of autonomous agents with complex tasks in partially observable environments

Junchao Li¹ · Mingyu Cai² · Zhen Kan³ · Shaoping Xiao¹

Accepted: 29 February 2024

© Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Motion planning of autonomous agents in partially known environments with incomplete information is a challenging problem, particularly for complex tasks. This paper proposes a model-free reinforcement learning approach to address this problem. We formulate motion planning as a probabilistic-labeled partially observable Markov decision process (PL-POMDP) problem and use linear temporal logic (LTL) to express the complex task. The LTL formula is then converted to a limit-deterministic generalized Büchi automaton (LDGBA). The problem is redefined as finding an optimal policy on the product of PL-POMDP with LDGBA based on model-checking techniques to satisfy the complex task. We implement deep Q learning with long short-term memory (LSTM) to process the observation history and task recognition. Our contributions include the proposed method, the utilization of LTL and LDGBA, and the LSTM-enhanced deep Q learning. We demonstrate the applicability of the proposed method by conducting simulations in various environments, including grid worlds, a virtual office, and a multi-agent warehouse. The simulation results demonstrate that our proposed method effectively addresses environment, action, and observation uncertainties. This indicates its potential for real-world applications, including the control of unmanned aerial vehicles.

Keywords Motion planning · Partially observable environments · Complex tasks · Linear temporal logic · Reinforcement learning · Recurrent neural networks

1 Introduction

The Partially Observable Markov Decision Process (POMDP) [1] provides a mathematical framework to model decision-making problems, including motion planning of autonomous agents, e.g., Unmanned Aerial Vehicles (UAVs). It differs from Markov Decision Processes (MDPs) [2] that have been widely applied in robotics and autonomous systems, assuming the environment is fully observable. POMDPs are more realistic for real-world applications in which the agent (e.g., the robot) may lack enough information from perception

Junchao Li, Mingyu Cai, Zhen Kan and Shaoping Xiao have contributed equally to this work.

Extended author information available on the last page of the article

and cannot completely identify the state of the environment. Similarly, another framework, Probabilistic-Labeled MDPs (PL-MDPs) [3, 4], introduces the labeling uncertainties to the traditional MDP model, capturing the probabilistic nature of the environments thus becoming a more precise model to representing the real-world complexities. On the other hand, simple go-to-goal motion planning tasks have been extensively studied by conventional pathfinding techniques [5] and Reinforcement Learning (RL) methods [6]. However, complex tasks like surveillance missions are more relevant to real-world applications. Therefore, it is challenging for the agent to learn how to plan its motions to accomplish complex tasks in partially observable environments, especially considering environment and transition uncertainties.

Researchers have utilized various model-based RL algorithms over the past decade to solve the POMDP problems, particularly for simple go-to-goal tasks. Many modern solvers can handle large spatial domains with thousands of states [7]. A commonly-used approach includes Point-Based Value Iteration (PBVI) [8–10] methods, consisting of model-based algorithms to approximately solve the POMDP problems by computing a value function over a finite subset of the belief space. It shall be noted that a belief state represents a probability distribution of the states where the agent can be. After each transition, the belief state needs to be updated by the transition and observation probabilities via the Bayesian approach. Indeed, such model-based approaches transform a POMDP problem into an equivalent MDP problem with the corresponding belief state space.

Another solution to POMDP problems is the model-free RL approach, in which the agent doesn't know transition and observation probabilities. Consequently, the policy maps a sequence of observations (i.e., the observation history) to the selected action. Mnih et al. [11, 12] first introduced a Deep Q-Learning (i.e., DQN) and tested it on several Atari 2600 games, in which the Q networks were trained to reach human-level performances. Particularly, their Q networks took the last four frames (in grayscale) as the input and utilized a Convolutional Neural Network (CNN) [13] to extract the image features for a fully-connected neural network to predict Q values (i.e., the state-action values). However, since the environment was modeled as MDPs in their work, this approach bypassed the issue of partial observability, and the last four frames allowed the agent to access limited past experiences.

Based on the above-mentioned works, Hausknecht and Stone [14] proposed adding a Recurrent Neural Network (RNN) to the Q network architecture. They modeled the Atari 2600 games as POMDP problems and proposed a so-called Deep Recurrent Q-Networks (DRQNs) by replacing the first post-convolutional fully-connected layer of Q networks with a Long Short-Term Memory (LSTM) [15], which took a single image at each time step as the input feature. Hence, this approach was able to integrate the entire observation history instead of utilizing the observation sequences with a fixed length (e.g., four frames of images as the input in [11]). In addition, Foerster et al. [16] extended the DRQN to handle multi-agent RL problems in partially observable environments by proposing a Deep Distributed Recurrent Q-Networks (DDRQN), where the last action was fed as the input to the Q network. Zhu et al. [17] developed a new network architecture called Action-Specific Deep Recurrent Q-Network (ADRQN), in which an LSTM layer processed the action history and associated observations for the Q value computing. Some other similar works, including [18, 19], implemented RNNs for the control policy on POMDP problems with continuous state spaces.

It shall be noted that the works on POMDP problems mentioned above consider simple go-to-goal missions only. However, more complex tasks have been included in MDP problems via formal languages [20], such as Linear Temporal Logic (LTL). Generally,

user-defined high-level specifications can be expressed as an LTL formula, which is then converted to an ω -automaton over infinite words with a Büchi or a Rabin acceptance condition [21]. Consequently, robotic motion planning problems can be solved via control synthesis for a product of MDP and automaton. Recently, this formal approach was employed to verify the task objectives when solving POMDP problems with certain temporal logic constraints. Chatterjee et al. [22] studied the undecidability of the qualitative model checking in an infinite-horizon POMDP. Their approach relied on exploring the entire belief space and was most suitable to the problems with small state spaces. They also concluded that it might be unable to acquire the optimal policy, ensuring the maximum satisfaction probability of the specific logic formula in POMDPs. Icarte et al. [23, 24] proposed a type of finite state machine known as Reward Machine (RM) to specify the reward functions while exposing their structure. Reward Machine is capable of addressing POMDP problems using model-free RL. However, the complexity in constructing RMs for larger POMDPs may pose challenges.

Other works like [25–27] proposed solving such problems by converting LTL specifications to a Deterministic Rabin Automaton (DRA), then constructing a product of POMDP and DRA. Specifically, Sharan et al. [25] and Ahmadi et al. [27] employed Finite State Controllers (FSCs) to limit the policy search via the value iteration method. Also, Bouton et al. [26] utilized the approximate POMDP solver, SARSOP [9], to search for an optimal policy on the finite belief state space of the product POMDP. However, those approaches are model-based RL methods, which require the agent to know the transition and observation probabilities. Such a requirement limits the applications in unknown environments. It shall be noted that the most related work that employs RNNs and model-checking for POMDP problems satisfying the temporal logic constraints was carried out by Carr et al. [28, 29]. Their proposed method focused on constructing a policy-based procedure to iteratively improve the current policy by implementing RNN. However, their approach constructed an underlying MDP to map the original POMDP and utilized a model-based RL solver, SARSOP [9], to approximate the value function. Hence, as they stated, it was still a model-based approach.

On the other hand, a range of studies [30–33] have presented the examples to highlight the advantages of using a Limit-Deterministic Generalized Büchi Automaton (LDGBA) over a DRA and a Limit-Deterministic Büchi Automaton (LDBA) in RL problems. Differing from an LDBA, which has a single set of accepting states, an LDGBA features multiple sets of accepting states. This property enables an agent to visit each set of accepting states infinitely often, making more complex acceptance criteria suitable for diverse tasks. Consequently, when designing rewards based on the acceptance conditions, the conversion of LDGBAs mitigates the issue of sparsity of reward caused by LDBAs, which might slow down the convergence in RL, as discussed in [34]. Furthermore, translating LTL specifications into an LDGBA can potentially yield a smaller automaton state space compared to a DRA [32]. Due to their simpler acceptance conditions, LDGBAs are semantically easier than DRAs, making policy synthesis algorithms much simpler to implement [33].

The contributions of this paper are outlined as follows. The first contribution is proposing a novel framework that models the interactions between the agent and its partially observable surroundings as a Probabilistic-Labeled POMDP (PL-POMDP). The introduction of probabilistic labeling to the POMDP framework enables the capability to consider both static and dynamic events in the environment. In addition, this work converts the LTL formula into an LDGBA, which effectively represents LTL specifications for a complex task in the considered POMDP problem. Following these advancements, we construct a product of PL-POMDP and LDGBA so that the original problem of finding a policy in a

PL-POMDP that satisfies LTL specifications is reformulated into solving an optimal policy that maximizes the cumulative reward on the corresponding product POMDP. To the best of authors' knowledge, this innovative approach has not been reported in the existing literature. Another contribution is developing a model-free RL method to learn an optimal policy on the product POMDP. We implement an RNN into Q network architectures to process the sequential information the agent acquires: the observation history and the task recognition. The latter depends on whether the agent fully recognizes the LTL-induced automaton. Specifically, either the history of automation states or the history of the state labels is utilized, respectively.

This paper is organized as follows: Sect. 2 reviews the PL-POMDP definition and introduces deep Q-learning with RNN to solve a simple go-to-goal POMDP problem. Section 3 presents LTL, LDGBA, and product PL-POMDP. Then, the problem of PL-POMDP with LTL specifications is reformulated. Section 4 proposes the model-free approaches to solve product PL-POMDP problems and provides detailed algorithms. Finally, experiments and results are included in Sect. 5, followed by the discussions and conclusions, including future works.

2 Background

In this section, we first define the Probabilistic-Labeled POMDP (PL-POMDP) and then explain using DQN, a model-free RL method, to solve a POMDP problem with simple go-to-goal tasks. Finally, a simulation example is conducted for the demonstration. Our focus in this paper is extending the DQN for the POMDP problem with complex tasks, and the developed methodology will be detailed in Sect. 4.

2.1 PL-POMDP

The POMDP is usually adopted as a mathematical description of problems in which the agent cannot fully observe and completely identify its surroundings. We employ PL-POMDP with the consideration of static and dynamic events.

Definition 1 (PL-POMDP) Considering the transition and observation uncertainties and the probabilistic labels of states, we can denote a PL-POMDP by a tuple $\mathcal{P} = (S, A, T, s_0, R, O, \Omega, \Pi, L, P_L)$, which consists of:

- A finite set of states, $S = \{s_1, \dots, s_n\}$.
- A finite set of actions, $A = \{a_1, \dots, a_m\}$. Particularly, $A(s)$ is the set of actions available for the agent at the current state s .
- A transition probability function, $T : S \times A \times S \rightarrow [0, 1]$ when the agent moves from the current state $s \in S$ to the next state $s' \in S$ after executing an action $a \in A(s)$. There exists $\sum_{s' \in S} T(s, a, s') = 1$.
- An initial state, $s_0 \in S$.
- A reward function, $R : S \times A \times S \rightarrow R$. It shall be noted that reward function sometimes can be defined as $R(s')$, $R(s, a)$ or $R(a, s')$.
- A finite set of observations, $O = \{o_1, \dots, o_k\}$. At the current state s , $O(s)$ consists of possible observations the agent can perceive.

- An observation probability function, $\Omega : S \times A \times O \rightarrow [0, 1]$, represents the probability that the agent can perceive observation o at state $s' \in S$ after executing action $a \in A(s)$. This function follows $\sum_{o \in O(s')} \Omega(s', a, o) = 1$.
- A set of atomic propositions, Π .
- A labeling function, $L : S \rightarrow 2^\Pi$, outputs a set of all possible labels at state s . 2^Π is the power set of Π .
- A labeling probability function, $P_L(s, l)$, represents the probability of a label $l \in L(s)$ associated with a state $s \in S$. It satisfies $\sum_{l \in L(s)} P_L(s, l) = 1, \forall s \in S$.

When the agent interacts with its environment, it has a (transition) probability $T(s, a, s')$ to move from state $s \in S$ to state $s' \in S$ after choosing and executing an action $a \in A(s)$. At the next state s' , the agent has an (observation) probability $\Omega(s', a, o)$ to perceive an observation $o \in O(s')$. In addition, the agent receives a reward as feedback from the environment based on the reward function $R(s, a, s')$.

Unlike the observations, the states' labels are atomic propositions representing event occurrences at specific states, thereby indicating that the status of labels can be dynamic. The labeling probability function can characterize a static event l at state s if $L(s) = \{l\}$ and $P_L(s, l) = 1$ or a dynamic event otherwise. The labels indicate goal states to handle simple go-to-goal tasks. When considering a complex task, we can use formal language to formulate the task regarding labels and then convert the formula to a finite state automaton. The labels are also input symbols for the induced automaton. Therefore, the accomplishment of the complex task can be validated via model checking during the motion planning. We will provide the details of such an approach in Sect. 3.

2.2 DQN for POMDP problems

To solve a POMDP problem with unknown transition probability, DQN is employed to map a sequence of observations to Q values for action selection. We consider observation histories up to j previous time steps, so the sequence of observations is denoted as $\mathbf{o}_t = (o_{t-j}, o_{t-j+1}, o_{t-j+2}, \dots, o_t)$ with a length of $j + 1$. Consequently, the policy is a function of observation sequence. The agent's objective in deciding a course of action is maximizing the expected return as below, representing the total discounted rewards the agent can collect from the current time under a policy ξ .

$$U^\xi(s_t) = \mathbb{E}^\xi \left[\sum_{\tau=0}^{\infty} \gamma^\tau R(s_{t+\tau}, a, s_{t+\tau+1}) \middle| s_t \right] \quad (1)$$

where s_t is the agent's state at time t , and $\gamma \in [0, 1]$ is the discount factor to balance the importance between immediate and future rewards. Then, the optimal policy can be found as $\xi^* = \text{argmax}_\xi U^\xi(s_0)$. It is worth mentioning that the optimal policy maps the sequence of observations \mathbf{o}_t to the action in POMDP.

We implement an LSTM to process the observation sequence in the Q network architecture as shown in Fig. 1. LSTM is one type of RNN that can model temporal dependencies between observations by introducing feedback loops in the network architecture. After processing the observation at each time step, the output is fed back into the network as input (together with the following observation) for the next time step. This allows the network to capture information about the order and timing of observations in the sequence. One-dimensional Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs)

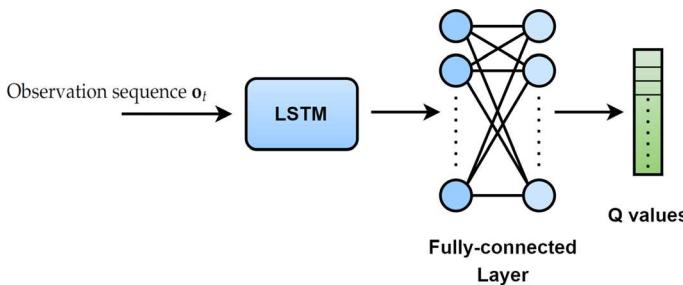


Fig. 1 Q network architecture

can also be used to process time-series data. However, they don't explicitly model the temporal dependencies between observations, and DNNs cannot process input sequences of variable length. In the next subsection, we use a simple go-to-goal simulation example to demonstrate the advantages of RNN in Q networks over CNN and DNN.

DQN usually has two Q networks: an evaluation Q-network $Q_E(\mathbf{o}_t, a_t; \theta_E)$ and a target Q-network $Q_T(\mathbf{o}_t, a_t; \theta_T)$, where θ_E and θ_T are the network weights, respectively. The evaluation Q-network Q_E is usually updated at each iteration by randomly selecting a batch of data samples from a so-called replay memory [35] during the learning process. At the same time, the target Q-network Q_T keeps fixed weights until copying from the evaluation Q-network Q_E once in a while, i.e., $\theta_T = \theta_E$.

At each step (e.g., time t) during the learning process, the target Q-network (Q_T) predicts Q values based on the sequence of observations \mathbf{o}_t so that the agent can choose an action a_t via the ϵ -greedy technique [6]. After executing the selected action and perceiving an observation o_{t+1} , a new sequence of observation $\mathbf{o}_{t+1} = (o_{t-j+1}, o_{t-j+2}, o_{t-j+3}, \dots, o_{t+1})$ is generated. Consequently, an experience is formed as $e_t = (\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1})$, where $r_t = R(s_t, a_t, s_{t+1})$. The experience is recorded as one data sample in the replay memory D , where the data samples are randomly selected to update the new Q values for Q network updating. The equation used to update the Q value associated with a_t is defined below.

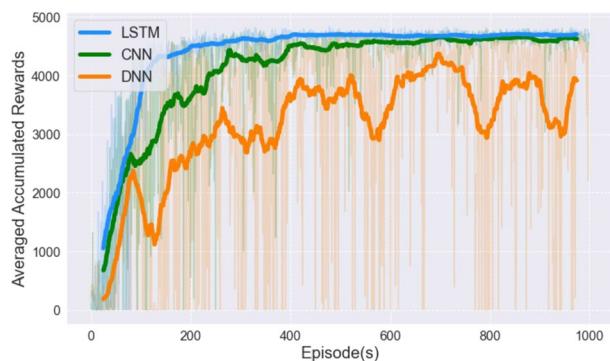
$$Q_{\text{new}}(\mathbf{o}_t, a_t) = Q_E(\mathbf{o}_t, a_t; \theta_E) + \alpha \left[r_t + \gamma \max_{a_{t+1}} Q_T(\mathbf{o}_{t+1}, a_{t+1}; \theta_T) - Q_E(\mathbf{o}_t, a_t; \theta_E) \right] \quad (2)$$

where α is the learning rate.

2.3 A go-to-goal example

The term 'go-to-goal' refers to a fundamental behavior for an autonomous agent of navigating from its current location to a target location. In this work, we employ this term and define the concept of 'go-to-goal POMDP motion planning problem' as an agent planning a path from an initial state, A, to a designated goal state, B, within the POMDP state space.

We take a simple go-to-goal example in a grid world to demonstrate DQN, i.e., a model-free RL method, to solve a POMDP problem. Figure 2 illustrates a 10×10 grid world, in which the blue area labeled as 'a' and the green area labeled as 'b' are the initial and goal state, respectively. The block states, i.e., obstacles, are labeled with 'B'. The agent, e.g., a mobile robot in the grid world, must move from the initial state to the goal state. It can take four actions at each state: *up*, *left*, *down*, and *right*. Due to the transition uncertainty, the agent has a 0.9 probability of moving in the direction of the chosen action. Otherwise,

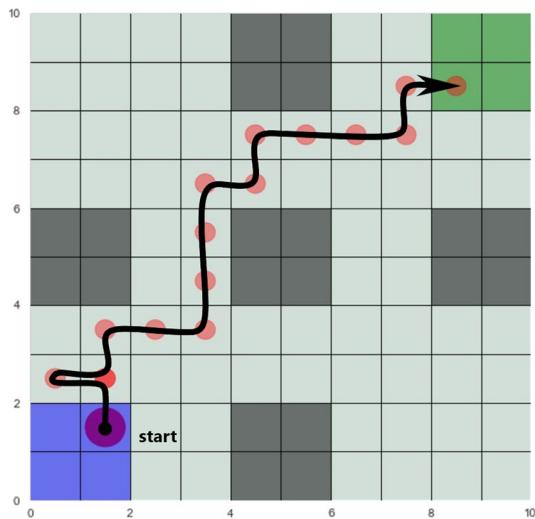
Fig. 2 A 10×10 grid world**Fig. 3** The comparison of accumulated rewards by using Q networks with LSTM, CNN, and DNN

it has an equal probability of moving to one of the two sideways directions. In addition, the agent will remain in the current state if the next state is outside the grid world or the obstacles. After reaching the next state, the agent can observe this state with a probability of 0.9 and adjacent states with a total probability of 0.1 uniformly distributed. The discount factor γ is set as 0.98.

It shall be noted that the observations are states in this grid world example. Therefore, the observation sequence is pre-processed by the technique of one-hot encoding since the row and column indices of a state are ordinal data. The generated 1-D vector as the input is fed into the LSTM layer. The extracted feature is then passed to two fully-connected layers with 16 neurons, respectively, to predict the corresponding Q values. The Rectified Linear Unit (ReLU) is utilized as the activation function in the Q networks. The learning process for this problem includes 500 steps per episode for 1000 episodes. Two Q networks, the evaluation network Q_E and the target network Q_T , are randomly initialized. The training batch size for the evaluation network is 32, and the target network is updated by copying the weight coefficients of Q_E every 50 steps.

We also test two other Q network architectures by implementing CNN and DNN. CNN-based Q networks use a 2D convolutional layer (filter=6, kernel size=(3,3), strides=1) followed by another 2D convolutional layer (filter=12, kernel size=(2,2), strides=1) before the fully connected layers. The DNN uses the fully connected layers only. Figure 3 compares the averaged accumulated rewards collected by the agent every ten episodes (the trend lines

Fig. 4 A path generated from the derived policy



represent the Simple Moving Average (SMA) of rewards every 50 episodes). The results illustrate that Q networks with LSTM can achieve faster convergence and a higher accumulated reward.

An optimal policy can be derived from the Q networks after convergence. Figure 4 displays a path generated from the derived policy for the agent moving on the grid world to accomplish this simple go-to-goal task. The start state is marked as the large purple solid circle, and the reached states are marked as light red dots. The brighter red dot and the bend of the black route indicate this state has been visited more than once. It shall be noted that even if the policy converges and approaches the optimal, the generated path may not. This is because of observation and transition uncertainties.

3 Problem definition

In this study, we introduce a framework for solving a PL-POMDP problem with LTL specifications by transforming the LTL formula into an LDGBA that represents the task variables and safety constraints of the POMDP and generating a product of POMDP and LDGBA. The problem of satisfying the given LTL constraints in a POMDP is equivalent to the problem of reaching (Büchi) accepting states in the product POMDP.

3.1 Linear temporal logic (LTL)

Linear temporal logic [20] is a logical formalism for linear-time properties, representing the relation between state labels in sequential executions. In addition to the Boolean connectors, LTL extends propositional logic by adding some temporal operators, including two basic ones \bigcirc (pronounced “next”) and \mathcal{U} (pronounced “until”). This study assumes that $a \in \Pi$ is an atomic proposition, and ϕ, ϕ_1 and ϕ_2 are single LTL formulas. Then, LTL formulas can be formed according to the following grammar [36]:

$$\phi ::= \text{True} \mid a \mid \phi_1 \wedge \phi_2 \mid \neg\phi \mid \text{O}\phi \mid \phi_1 \mathcal{U} \phi_2 \quad (3)$$

where negation (\neg) and conjunction (\wedge) are the Boolean operators. Formula $\text{O}\phi$ is true at the current time if ϕ is true the next time. In addition, Formula $\phi_1 \mathcal{U} \phi_2$ is true at the current time if ϕ_2 is true for some future time and ϕ_1 is true at all times until that future time.

Other commonly-used temporal operators are \diamond (pronounced “eventually”) and \square (pronounced “always”). Formula $\diamond\phi$ ensures that ϕ will be true eventually in the future, while $\square\phi$ is true from now on forever. They can be derived as follows:

$$\begin{aligned} \text{eventually} : \quad \diamond\phi &\equiv \text{True} \mathcal{U} \phi \\ \text{always} : \quad \square\phi &\equiv \neg(\diamond\neg\phi) \end{aligned} \quad (4)$$

When using \models to represent the satisfaction relationship, we can interpret the semantics of an LTL formula over words as below. A word is an infinite sequence $w = w_0 w_1 \dots$ with $w_i \in 2^\Pi$ where Π is a set of atomic propositions for all $i \geq 0$.

$$\begin{aligned} w \models \text{True} \\ w \models \alpha &\Leftrightarrow \alpha \in L(w[0]) \\ w \models \phi_1 \wedge \phi_2 &\Leftrightarrow w \models \phi_1 \text{ and } w \models \phi_2 \\ w \models \neg\phi &\Leftrightarrow w \not\models \phi \\ w \models \text{O}\phi &\Leftrightarrow w[1 :] \models \phi \\ w \models \phi_1 \mathcal{U} \phi_2 &\Leftrightarrow \exists t \text{ s.t. } w[t :] \models \phi_2, \forall t' \in [0, t), w[t' :] \models \phi_1 \end{aligned} \quad (5)$$

3.2 Limit-deterministic generalized Büchi automaton (LDGBA)

The previous subsection describes that a user-specified complex task can be formulated via LTL. Then, we can convert the LTL formula into an automaton, including LDGBA [37], to evaluate task satisfaction via model checking [20].

Definition 2 (LDGBA) An LDGBA $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$ consists of a finite set of states Q , a finite alphabet (i.e., a finite set of input symbols) $\Sigma = 2^\Pi$ where Π is a set of atomic propositions, a transition function $\delta : Q \times (\Sigma \cup \{\epsilon\}) \rightarrow 2^Q$, an initial state $q_0 \in Q$, and a set of accepting sets $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_f\}$ where $\mathcal{F}_i \subseteq Q$, $\forall i \in \{1, \dots, f\}$. Furthermore, the state set Q can be decomposed into deterministic and non-deterministic sets, i.e., Q_D and Q_N , respectively. They satisfy the following requirements.

- $Q_D \cup Q_N = Q$ and $Q_D \cap Q_N = \emptyset$.
- The state transitions in Q_D are total, i.e., $|\delta(q, \alpha)| = 1$.
- The transitions in Q_D are restricted within it, i.e., $\delta(q, \alpha) \subseteq Q_D$ for every state $q \in Q_D$ and $\alpha \in \Sigma$.
- The ϵ -transitions do not take the input symbols and are only valid from $q \in Q_N$ to $q' \in Q_D$.
- The accepting sets, consisting of accepting states, are only defined in the deterministic set. In other words, $\mathcal{F}_i \subseteq Q_D$ for every $\mathcal{F}_i \in \mathcal{F}$.

A run of an LDGBA, subject to an input word $w = w_0 w_1 \dots$, can be expressed as $q = q_0 q_1 \dots$, according to the transition function $\delta(q_i, w_i) = q_{i+1}$. Let $\text{inf}(q)$ represent the infinite portion of q . Theoretically, q satisfies the LDGBA acceptance condition, i.e., the

LDGBA accepts the word w , if there exists $\inf(\mathbf{q}) \cap \mathcal{F}_i \neq \emptyset$, $\forall i \in \{1, \dots, f\}$. We suggest readers check Owl [21] for more details about automaton generation. This study aims to solve the POMDP problems with LTL constraints as defined below.

Problem 1 Given a PL-POMDP \mathcal{P} and a complex task expressed via an LTL formula. The objective is to find a policy $\xi^*(\mathbf{o}_t)$, where \mathbf{o}_t denotes a sequence of observations on \mathcal{P} , that can complete the task by satisfying the acceptance condition of the LTL-induced LDGBA.

3.3 Product POMDP

Definition 3 (Product POMDP) Given a PL-POMDP $\mathcal{P} = (S, A, T, s_0, R, O, \Omega, \Pi, L, P_L)$ and an LDGBA $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$, the product PL-POMDP (simply named as the product POMDP) is defined by $\mathcal{P}^* = \mathcal{P} \times \mathcal{A} = (S^*, A^*, T^*, s_0^*, R^*, O, \Omega^*, \mathcal{F}^*)$, consisting of the following components.

- A finite set of labeled states, $S^* = S \times Q$ or $s^* = \langle s, q \rangle \in S^*$ where $s \in S$ and $q \in Q$.
- A finite set of actions, $A^* = A \cup \{\epsilon\}$.
- A transition function, $T^* = S^* \times A^* \times S^* \rightarrow [0, 1]$, and

$$T^*(s^*, a^*, s') = \begin{cases} T(s, a^*, s') & \text{if } q' = \delta(q, l), l \in L(s') \text{ and } a^* \in A \\ 1 & \text{if } a^* \in \{\epsilon\} \text{ and } q' \in \delta(q, \epsilon) \text{ and } s' = s, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

where $s^* = \langle s', q' \rangle$.

- An initial state $s_0^* = \langle s_0, q_0 \rangle \in S^*$ where $s_0 \in S$ and $q_0 \in Q$.
- A reward function $R^* : S^* \times A^* \times S^* \rightarrow R$, and

$$R^*(s^*, a^*, s') = \begin{cases} R(s, a^*, s') & \text{if } a^* \in A, l \in L(s'), q' = \delta(q, l) \in \mathcal{F}_i, \text{ and } \mathcal{F}_i \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

- An observation function $\Omega^* = S^* \times A^* \times O \rightarrow [0, 1]$ is

$$\Omega^*(s^*, a^*, o) = \Omega(s', a^*, o) \quad (8)$$

If $a^* \in A$. Otherwise, if $a^* \in \{\epsilon\}$, the agent stays at the same state s , i.e., $s' = s$, but $q' = \delta(q, \epsilon)$, and no observation is perceived.

- A set of accepting sets $\mathcal{F}^* = \{\mathcal{F}_1^*, \mathcal{F}_2^*, \dots, \mathcal{F}_f^*\}$ where $\mathcal{F}_i^* = \{\langle s, q \rangle | s \in S; q \in \mathcal{F}_i\}$ and $i = 1, \dots, f$.

A random path on the product POMDP, represented by $(s_0, q_0)(s_1, q_1) \dots$, is an integration of a path $s_0s_1 \dots$ on the PL-POMDP and a path $q_0q_1 \dots$ on the LDGBA. Similar to (1), the expected return starting from the initial state under a policy ξ^* on the product POMDP can be written as

$$U^{\xi^*}(s_0^*) = \mathbb{E}^{\xi^*} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t^*, a_t^*, s_{t+1}^*) \middle| s_{t=0}^* = s_0^* \right] \quad (9)$$

It shall be noted that the product POMDP \mathcal{P}^X can be viewed as a PL-POMDP \mathcal{P} with the augmented state space, which includes automaton state space. Therefore, the product POMDP accounts for the temporal logic specifications represented by LDGBA \mathcal{A} . Any feasible path on \mathcal{P}^X shares the intersections between an accessible path over the original PL-POMDP \mathcal{P} and a word accepted by the LTL-induced automaton \mathcal{A} . For example, a path $\sigma^{\xi^X} = (s_0, q_0)(s_1, q_1) \dots$ can be generated by the derived policy ξ^X on the product POMDP \mathcal{P}^X . If there exists $\inf(\sigma^{\xi^X}) \cap \mathcal{F}_i^X \neq \emptyset, \forall i = 1, \dots, f$, where \mathcal{F}^X captures the acceptance conditions of LDGBA \mathcal{A} , this path is accepted. In other words, the run $q_0 q_1 \dots$ satisfies the LDGBA acceptance condition, or the LDGBA accepts the corresponding word.

Therefore, according to many previous studies on solving MDP problems with LTL specifications via the product MDP [3, 4, 31, 36], an optimal policy $\xi^{X*}(\mathbf{o}_t, \mathbf{q}_t)$ on the product POMDP \mathcal{P}^X is equivalent to the optimal policy $\xi^*(\mathbf{o}_t)$ on the PL-POMDP \mathcal{P} while satisfying LTL specifications. \mathbf{o}_t is the observation history while \mathbf{q}_t is the corresponding history of transitioned automaton state. It shall be noted that it is assumed that the agent receives labels, i.e., input symbols in automaton, as part of the feedback. Therefore, if the agent is fully aware of the task, i.e., LTL-induced automaton, the history of transitioned automaton states can be derived. Then, we can reformulate Problem 1 as follows.

Problem 2 A product POMDP $\mathcal{P}^X = \mathcal{P} \times \mathcal{A}$, defined in Sect. 3.3, is constructed from a PL-POMDP \mathcal{P} describing the partially observable environment and an LDGBA \mathcal{A} expressing LTL specifications ϕ for a complex task. The objective is to find a policy $\xi^{X*}(\mathbf{o}_t, \mathbf{q}_t)$, where \mathbf{o}_t and \mathbf{q}_t denote the sequences of observations and transitioned-automaton states, respectively, over \mathcal{P}^X so that the expected return (9) is maximized.

On the other hand, if the agent is unaware of the task, the agent cannot derive automaton state transitions based on the label feedback. Consequently, the policy on the product POMDP is a function of the observation history and the perceived label history as $\xi^{X*}(\mathbf{o}_t, \mathbf{l}_t)$. Then, the above problem formulation can be corresponding revised.

4 Methodology

In this study, we propose model-free RL approaches (i.e., DQNs) on product POMDPs to synthesize optimal motion planning for the agent in a partially observable environment subject to LTL specifications. As discussed and demonstrated in Sect. 2, RNNs have the advantage of being included in Q networks for solving POMDP problems because they can capture relative temporal dependencies in the observation history. In the proposed methods, we extend the RNN-enhanced Q networks to process the perceived observations and the recognition of complex tasks.

We consider two scenarios, depending on whether the agent acknowledges the assigned task. If the agent is explicitly assigned the task, it has full knowledge of the LTL-induced automaton, including the transition function. Consequently, once it reaches a product POMDP state, i.e., an augmented state, it can derive the associated automaton state, although the POMDP state is not fully observable. It shall be noted the state labels are distinguished from the observations. We assume that the agent can acquire state labels

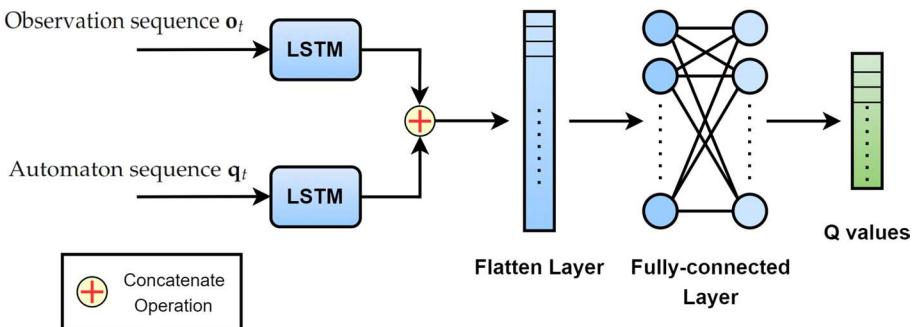


Fig. 5 The architecture of Q networks taking \mathbf{o}_t and \mathbf{q}_t as input

correctly via feedback from the environment. Therefore, in this case, in addition to the observation history, our method takes the identified automaton state as another input feature to Q networks. On the other hand, if the agent is not explicitly assigned the task and has no knowledge of the automaton's transition function, (POMDP) state labels are taken as an additional input feature to Q networks.

4.1 Automaton state sequence as additional input

Suppose the agent is fully aware of the task. In that case, the automaton is fully knowledgeable, and the automaton states (i.e., q states) can be induced from the acquired labels according to the automaton transition function. Consequently, in addition to the observation history, automaton states can be directly used as input to the Q networks. Indeed, we utilize a sequence of q states with a length of k , $\mathbf{q}_t = \bar{q}_1 \dots \bar{q}_k$, where \bar{q} denotes the induced q state, representing the automation evolution by time step t corresponding to the agent's transitions on the PL-POMDP. It shall be noted that the sequences of observations $\mathbf{o}_t = o_{t-j} \dots o_t$ have a length of $j+1$ and are generated based on the observation history from time step $t-j$ to the current time t . However, since the automaton space is usually smaller than the POMDP space, and the automaton state transition doesn't occur at each time step, it is not practical to generate the history of the automaton transition by recording the automaton state at each time step. Instead, we use First In, First Out (FIFO) to track the automaton transitions, and each transition is recorded only once. Consequently, \mathbf{q}_t may maintain the same for several consecutive steps until the next automaton state is reached. In addition, \mathbf{o}_t and \mathbf{q}_t usually don't have the same length.

After each transition $a_t^X \in A$, the observed o_{t+1} and induced automaton state q_{t+1} are used to generate the new sequences of observations \mathbf{o}_{t+1} and q states \mathbf{q}_{t+1} . Together with the previous sequences, a new experience can be written as $(\mathbf{o}_t, \mathbf{q}_t, a_t^X, r_t^X, \mathbf{o}_{t+1}, \mathbf{q}_{t+1})$, which is recorded as one data sample in the replay memory. If there is no new automation transition, $\mathbf{q}_{t+1} = \mathbf{q}_t$. It shall be noted that the ϵ -transition is one of the available actions in the product POMDP as defined in Sect. 3.3 if it exists in the LTL-induced LDGBA. In our approaches, no observation is perceived after an ϵ -transition because the agent remains at the same

POMDP state. However, the automaton state is changed and recorded in the corresponding automaton state sequence for the next time step.

Figure 5 illustrates the architecture of the Q networks in our model-free RL approaches. The sequences of observations \mathbf{o}_t and automaton states \mathbf{q}_t at time step t are pre-processed by one-hot-encoding before entering LSTMs, respectively. Since these two sequences don't have the same length, two separate LSTMs are adopted to extract the hidden states, which are then concatenated and flattened for the fully-connected layers to estimate Q values.

Similar to a general DQN, our DQN on the product POMDP also has two identical Q networks: the evaluation network $Q_E^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_E^\times)$ and the target network $Q_T^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_T^\times)$ where $a^\times \in A$. The target network is utilized for the next action selection and Q value prediction. On the other hand, the evaluation network is trained every M steps by a batch of data samples randomly selected from the replay memory. After every K steps, the target network is updated by copying the weight coefficients of the evaluation networks. Using two neural networks can prevent the bootstrapping of the DQN with a single neural network.

The output of each data sample in the selected batch is a Q value, which can be updated as the equation below.

$$Q_{\text{new}}^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times) = Q_E^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_E^\times) + \alpha \left[r_t^\times + \gamma \max_{a_{t+1}^\times} Q_T^\times(\mathbf{o}_{t+1}, \mathbf{q}_{t+1}, a_{t+1}^\times; \theta_T^\times) - Q_E^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_E^\times) \right] \quad (10)$$

where $r_t^\times = R^\times(s_t^\times, a_t^\times, s_{t+1}^\times)$, α is the learning rate, and γ is the discount factor. Therefore, each data sample has the input features \mathbf{o}_t and \mathbf{q}_t and the output target Q_{new}^\times to train and update the evaluation Q network, Q_E . In addition, the following loss function is used to update the weight coefficients of the evaluation Q-network.

$$\mathcal{L}(\theta_E^\times) = \mathbb{E}_{(\mathbf{o}_t, \mathbf{q}_t, a_t^\times, r_t^\times, \mathbf{o}_{t+1}, \mathbf{q}_{t+1}) \sim U(D)} \left[\left(r_t^\times + \gamma \max_{a_{t+1}^\times} Q_T^\times(\mathbf{o}_{t+1}, \mathbf{q}_{t+1}, a_{t+1}^\times; \theta_T^\times) - Q_E^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_E^\times) \right)^2 \right] \quad (11)$$

Algorithm 1 demonstrates training the evaluation Q network during the agent interactions with the environment. Once converged, the Q network can predict the state-action value (i.e., Q value) function to derive the optimal policy for the studied POMDP problem with temporal logic specifications.

Algorithm 1 Deep Recurrent Q-Network for Product POMDP Problems.

```

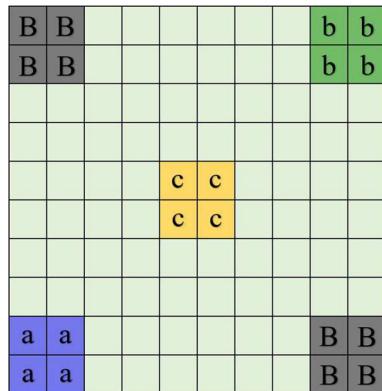
1: Initialize LTL formula  $\phi$ , POMDP  $\mathcal{P}$ .
2: Convert  $\phi$  to an LDGBA  $\mathcal{A}$ .
3: Construct the product POMDP  $\mathcal{P}^\times = \mathcal{P} \times \mathcal{A}$ .
4: Initialize the evaluation network  $Q_E^\times$ , the target network  $Q_T^\times$ , the replay memory
    $D$ , the length of observation sequence  $j + 1$ , the empty  $q$  state sequence  $\mathbf{q}_0$  with
   length of  $k$ , the learning rate  $\alpha$ , the discount factor  $\gamma$ , the total number of episodes
    $E$ , the total number of steps  $N$ , the batch size  $M$ , and the  $Q_T^\times$  update steps  $K$ .
5: while The current episode  $e$  in  $E$  do
6:   Randomly select a start state  $s_0^\times$ .
7:   while The current step  $i$  in  $N$  do
8:     Select a random action  $a_i^\times$  if  $i < j + 1$ ; otherwise, select an action via the
     $\epsilon$ -greedy technique.
9:     Obtain the observation  $o_{t+1}$  and induce the automaton state  $q_{t+1}$ .
10:    Generate  $\mathbf{o}_{t+1}$  and  $\mathbf{q}_{t+1}$ .
11:    Collect the rewards  $r_i^\times$ .
12:    Store the experience  $\langle \mathbf{o}_i, \mathbf{q}_i, a_i^\times, r_i^\times, \mathbf{o}_{i+1}, \mathbf{q}_{i+1} \rangle$  in  $D$ .
13:    if  $i > 0$  and  $i \% M = 0$  then
14:      Randomly select  $M$  data samples as  $U(D)$  from the replay memory.
15:      Compute  $Q_{new}^\times$  for each data sample.
16:      Train  $Q_E^\times$  by the batch of samples.
17:    end if
18:    if  $i > 0$  and  $i \% K = 0$  then
19:      Pass the weights of  $Q_E^\times$  to  $Q_T^\times$ .
20:    end if
21:  end while
22: end while
23: Training end and save the evaluation network  $Q_E^\times$ 

```

4.2 Label sequence as the additional input

If the agent is unaware of the task, i.e., the LTL-induced automaton is not knowledgeable, the state labels need to be the additional input for Q networks. We modified the Q network architecture in Fig. 5 by replacing the automaton state sequence with a label sequence (i.e., a sequence of input symbols) in addition to the observation sequence. As mentioned above, state labels differ from observations and can be precisely received by the agent as feedback. Consequently, the sequence of the collected experience becomes $(\mathbf{o}_t, \mathbf{l}_t, a_t^\times, r_t^\times, \mathbf{o}_{t+1}, \mathbf{l}_{t+1})$, where \mathbf{l}_t is the sequence of labels with a length of k received by the time step t . Similar to the scenario of utilizing q states, the labels corresponding to the POMDP states can be sparse. Hence, we utilize the same FIFO method (as described in Sect. 4.1) to generate the label sequence that only stores a label once it is received. On the other hand, Algorithm 1 can be correspondingly revised.

Fig. 6 A 10×10 grid world with trapping states



5 Simulations and results

We evaluate our approaches on three simulations with discrete POMDP domains. We first perform simulations over a partially observable grid world with two different tasks, considering ϵ -transition in the automaton and static/dynamic events in the PL-POMDP. Then, we test the approaches in an office scenario where we utilize two different observation settings. Finally, we also conduct a preliminary application of the proposed approach to a multiagent RL case. The simulations are programmed via Python 3.9 and Rabinizer 4. They are completed on a desktop with a 3.20 GHz eight-core CPU and 32 GB RAM. Part of the source codes and supplementary materials are provided.¹

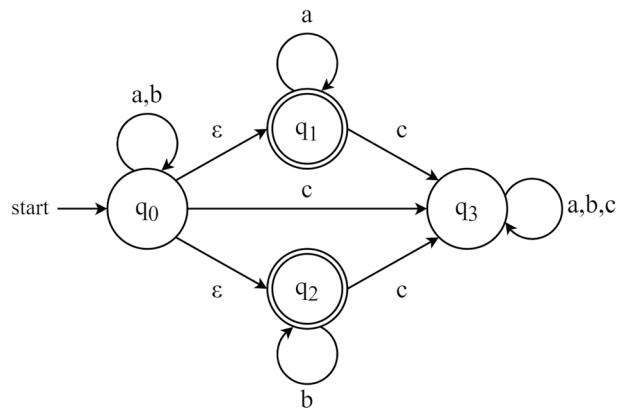
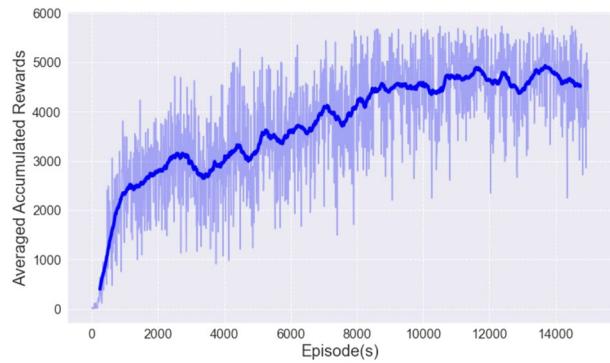
5.1 Grid world simulations

We use a 10×10 grid world workspace, shown in Fig. 6. Several states are labeled with ‘a’ in blue and ‘b’ in green, indicating two different events. The trapping states, labeled with ‘c’, indicate that the agent can never leave once entering them. Other parameters, like the transition probability and observation probability, are the same as defined in the simple go-to-goal case in Sect. 2.3. Two tasks are studied. Task 1 demonstrates that our approaches can handle LDGBA with ϵ -transitions. Task 2 considers two scenarios: static events and dynamic events, respectively. When assuming dynamic events, each event has a 90% probability of occurring at its labeled states and a 10% probability at the other labeled states. Both tasks are simulated for 15,000 episodes with 600 steps per episode, using the observation sequence with a length of $j + 1 = 5$, the automaton state or label sequence with a length of $k = 3$, batch size $M = 32$, and the number of steps to copy the evaluation Q network $K = 50$.

5.1.1 Task 1

The first task tested in the grid world requires the agent to visit states labeled ‘a’ or ‘b’ infinitely many times. The LTL formula is expressed below, and the induced LDGBA is shown in Fig. 7.

¹ https://github.com/JunchaoLi001/Model-free_DRL_LSTM_on_POMDP_with_LDGBA.

Fig. 7 The LDGBA of ϕ_1 **Fig. 8** The averaged accumulated rewards of task ϕ_1 

$$\phi_1 = (\square \diamond a \vee \square \diamond b) \wedge \square \neg c \quad (12)$$

It can be seen that the LDGBA contains ϵ -transitions that are included in the set of actions, a^\times , in the generated product POMDP. Figure 7 also shows that the ϵ -transitions can transition the automaton states from the initial state q_0 to either q_1 or q_2 , resulting in the agent to keep visiting only ‘a’ or ‘b’, respectively. Visiting ‘c’ will lead to a state (q_3) without an outlet as a “trapping” state.

According to the definition of LDGBA in Sect. 3.2, ϵ -transitions don’t take any input symbols. They are only valid to enter the deterministic set of automaton states where the transitions are restricted. Therefore, after an ϵ -transition, the agent will be at the augmented states associated with either q_1 or q_2 to complete the task. Without the loss of generality, we give each episode a random probability for selecting the ϵ -transition, which only occurs once. Otherwise, the other actions, i.e., $a^\times \in A$, will be chosen based on the ϵ -greedy method. As mentioned in Sect. 3.3, the agent doesn’t perceive observations right after taking ϵ -transitions, and Q networks only predict Q values of the actions other than ϵ -transitions.

In Task 1, we only consider static events, i.e., $P_L(s_a, \{a'\}) = P_L(s_b, \{b'\}) = 1$. Also, it is assumed that the agent is aware of the task. Therefore, the observation and q state histories are input to predict Q values via Q networks for action selection. Figure 8 shows

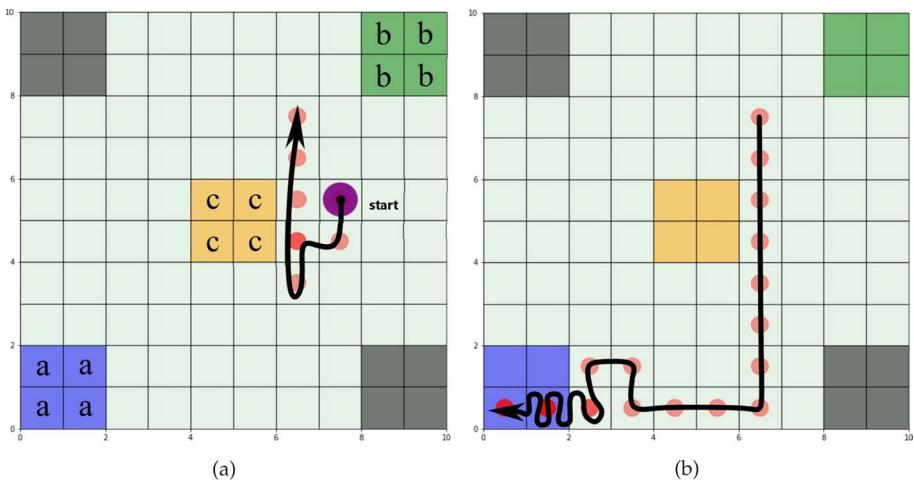


Fig. 9 A single round path of task ϕ_1 : **a** The path on q_0 . **b** The path on q_1

the evolution of accumulated reward, averaged per 10 episodes with SMA 50 episodes. Since the agent may accidentally move into the ‘trapping’ state, averaged rewards better presents the trend of the rewards’ convergence. The reward is set as 10 whenever the agent visits accepting states. After obtaining the optimal policy, we generate one path for the agent to accomplish the task as shown in Fig. 9.

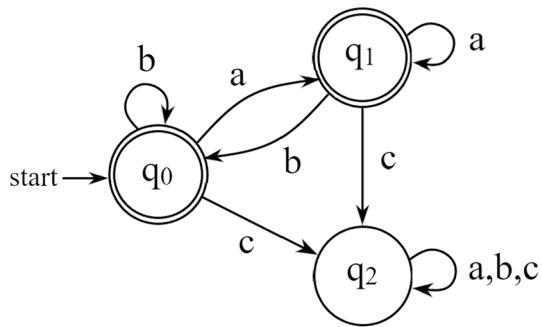
At first, the agent randomly selects actions, as shown in Fig. 9a, before generating the first observation and q state sequences. Then, Q values can be predicted, and the greedy action selection is applied unless an ϵ -transition is taken. After the ϵ -transition, the automaton state is transitioned from q_0 to q_1 , and the agent desires to visit the state labeled ‘a’ infinitely many times. This can be seen in Fig. 9b, where the agent moves down and then left to keep visiting state ‘a’. Due to the transition uncertainty, the agent occasionally visits some states multiple times, indicated as bright red dots in Fig. 9.

5.1.2 Task 2

The second task requires the agent to visit states labeled ‘a’ then ‘b’ in order infinitely many times, subject to dynamic events due to labeling uncertainty. The LTL formula is expressed as

$$\phi_2 = \square \diamond (a \wedge \diamond b) \wedge \square \neg c \quad (13)$$

In the LTL-induced LDGBA, directly utilizing the accepting sets may fail to find the deterministic policy as discussed in [32, 34]. Inspired by their works, we modify the automaton structure and reward function for easing the training process. Specifically, we augment the accepting states to separate transitions with the input symbols ‘a’ or ‘b’, respectively. Only single-labeled transitions are kept for demonstration, as shown in Fig. 10 for Task 2. We then redesign the reward function (7), shown below, by adding a constraint to the reward function so that the agent can visit the accepting sets repeatedly.

Fig. 10 The LDGBA of ϕ_2 

$$R^{\times}(s^{\times}, a^{\times}, s'^{\times}) = \begin{cases} R(s, a^{\times}, s') & a^{\times} \in A, l \in L(s), q' = \delta(q, l) \in \mathcal{F}_i, \mathcal{F}_i \in \mathcal{F}, \text{ and } q' \neq q \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

where $q \neq q'$ prevents the repeated transitions at the same automaton accepting state by removing the rewards on the associated labeled POMDP states. After applying this constraint to the reward function, the derived optimal policy satisfies the desired surveillance task specification. An alternative approach can be implementing the frontier-tracking function introduced by Cai et al. [3].

For Task 2, we investigate the agent's learning in the environment with dynamic or static events combined with two scenarios that depend on the agent's awareness of the task. At first, to provide a detailed demonstration, we consider dynamic events and assume that the agent fully understands the task. The labeling uncertainty is introduced so that an event has a 90% probability of occurrence at its labeled states and the other labeled states otherwise. For example, at the states labeled 'a', event 'a' has a 90% probability of occurrence while event 'b' has a 10% probability. In addition, a q state sequence is utilized as the input of Q networks in addition to the observation sequence.

Figure 11(1) demonstrates a single round path generated from the learned policy for the agent to accomplish Task 2 when the agent is fully aware of the task. In Fig. 11a, the agent starts from the initial state (purple dot) and performs five random movements on q_0 until it generates the first observation sequence (along with the q state sequence) for decision-making. It then navigates around the trapping states and heads to the bottom left for the states labeled 'a' shown as the blue square. However, event 'a' doesn't occur in the state when the agent first visits the blue area due to the labeling uncertainty. Therefore, the agent moves downward to the next state, where event 'a' occurs at the next time step. The bend of the black route at the top of the path is caused by motion uncertainty.

After visiting the state labeled 'a', the automaton transition happens from q_0 to q_1 . In Fig. 11b, the agent moves around the area of 'a' states then bypasses the trapping states in yellow to reach the states labeled 'b' in the top right corner. Again, event 'b' doesn't occur on the agent's first visit. Then, the agent moves one more step to the right, and event 'b' occurs. Consequently, the agent completes a single round to visit 'a' and then 'b'. At the same time, the agent is back to the automaton state q_0 . Figure 11c shows the agent tries to move back to visit states labeled 'a' for the second round, but it keeps visiting 'b' states a few times before heading to 'a' states.

We also conduct the simulation in an environment with dynamic events when the agent is unaware of the task. The observation and label sequences are input to the Q networks in this case. Figure 11(2) illustrates the path generated from the learned policy. We observe a

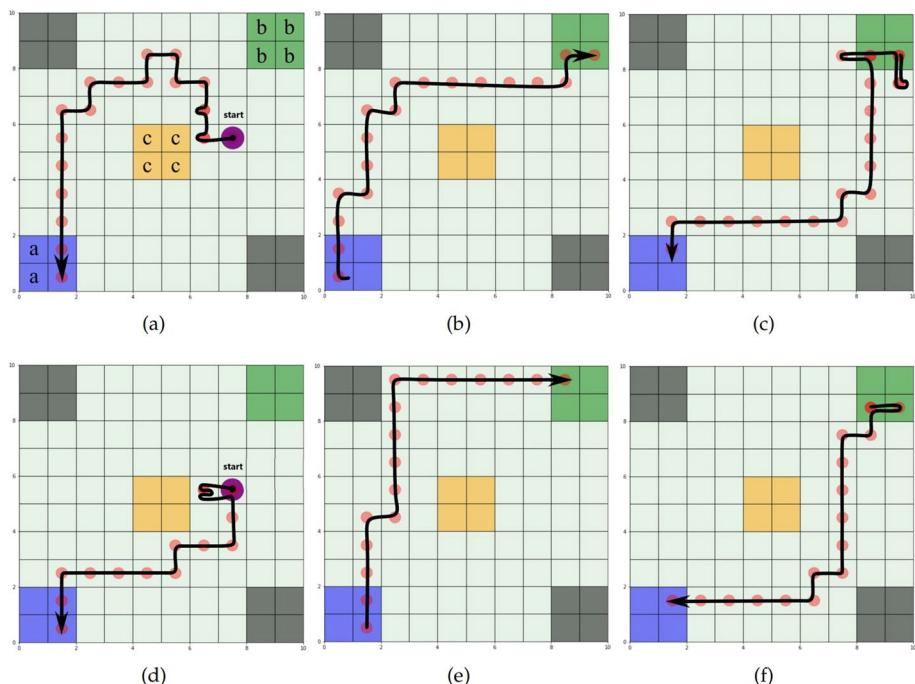


Fig. 11 A generated path for Task in ϕ_2 of dynamic event with labeling uncertainty: (1) If it is fully aware of the task (q state sequence): **a** The path on q_0 . **b** The path on q_1 . **c** The path back on q_0 . (2) If it is not aware of the task (label sequence): **d** The path on q_0 . **e** The path on q_1 . **f** The path back on q_0

similar phenomenon in which the agent makes a few attempts to visit ‘a’ states before leaving for ‘b’ states. However, after we generate more paths, we find that the agent occasionally visits the states labeled ‘b’ first because event ‘a’ has 10% probability of occurring on those states. This phenomenon is uniquely observed when the agent is unaware of the task.

Next, the events are assumed to be static, i.e., without labeling uncertainty. We consider both scenarios depending on whether the agent is fully aware of the task. After obtaining optimal policies for each scenario, we generate paths to demonstrate the agent accomplishing the task. Figure 12 (1) shows the agent planning the motion based on the observation history and the q state history. On the other hand, Fig. 12 (2) demonstrates that the agent can accomplish the task following the policy regarding the observation history and the label-receiving history. By following both paths, the agent can complete the task, and the performances between the two approaches are similar. Comparing the dynamic event cases, we don’t observe that the agent tries to visit ‘a’ or ‘b’ states multiple times before leaving for the other.

Figure 13 compares the evolution of the accumulated rewards of Task 2 for four cases discussed above. It can be observed that static event cases reach higher accumulated rewards than dynamic event cases. It may be due to the labeling uncertainty in dynamic events. In this 10×10 grid world challenge, given the simple and straightforward label assignment, both automaton state sequence and label sequence provide the agent with adequate information for the decision-making process. Consequently, the learning performance is comparable whether the agent is aware of the task or not. However, as we

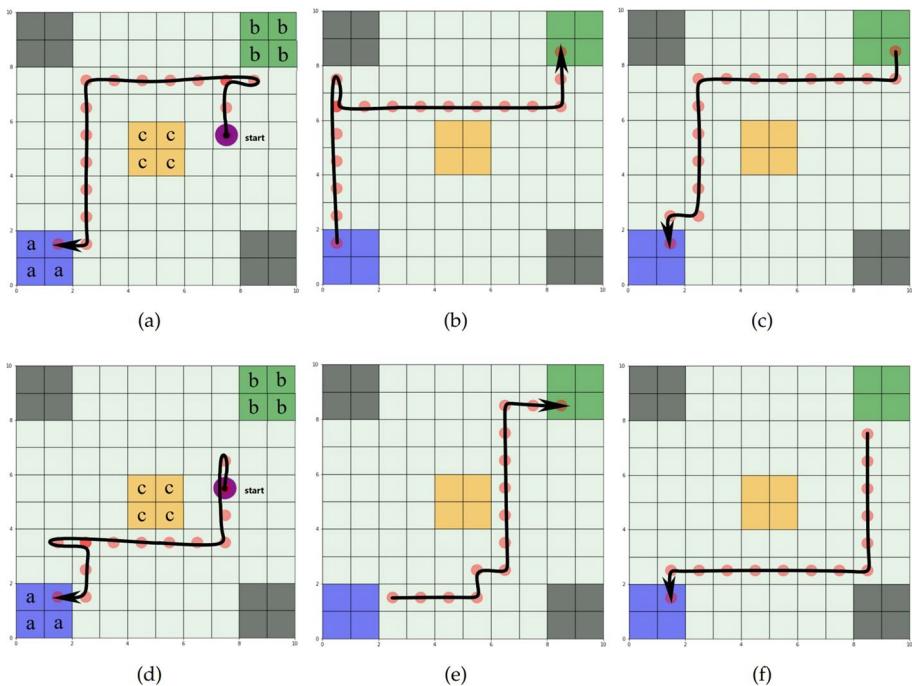
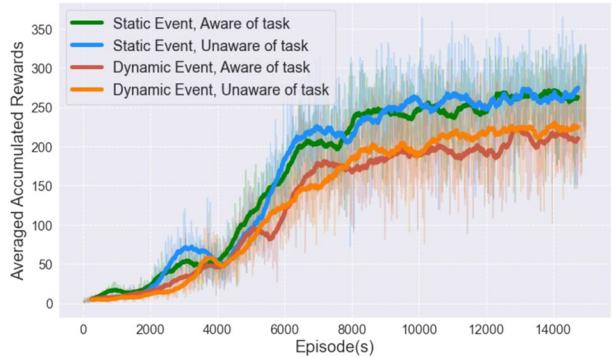


Fig. 12 A generated path for Task in ϕ_2 of static event: (1) If it is fully aware of the task (q state sequence): **a** The path on q_0 . **b** The path on q_1 . **c** The path back on q_0 . (2) If it is not aware of the task (label sequence): **d** The path on q_0 . **e** The path on q_1 . **f** The path back on q_0

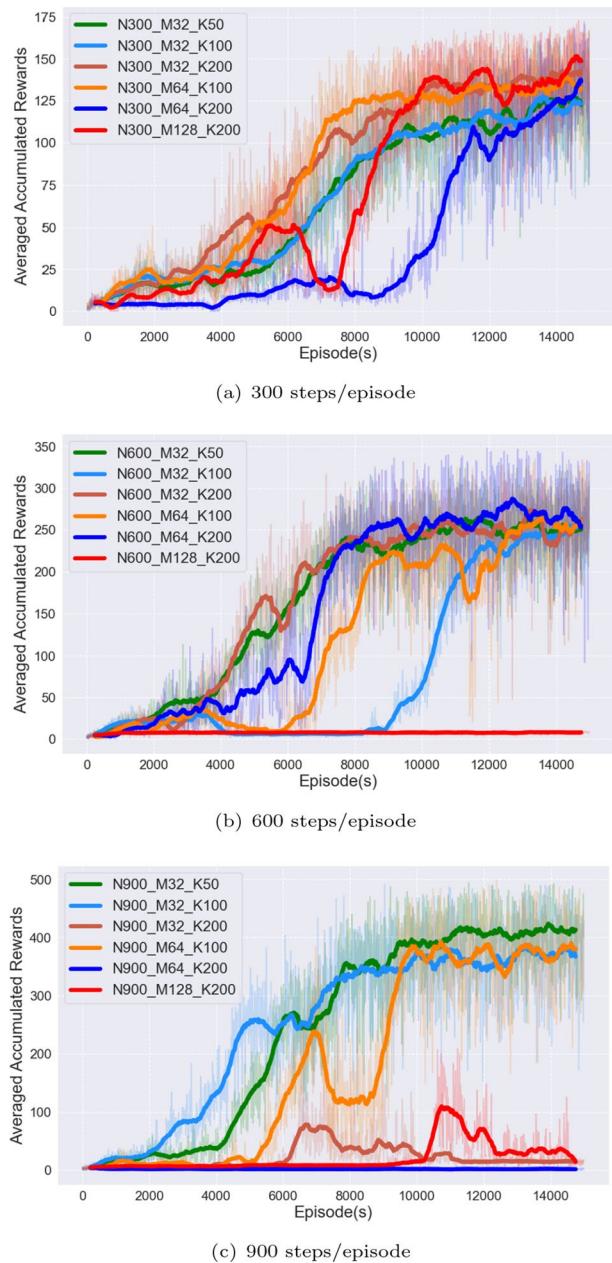
Fig. 13 Comparison of the reward evolution for task ϕ_2



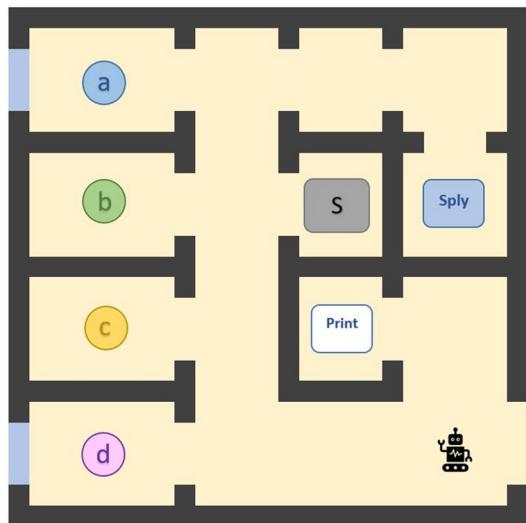
will mention in the discussion section, there can be notable differences in the training performance between these two scenarios.

Figure 14 shows the comparison between the reward evolution for task ϕ_2 with various batch sizes M , Q-network update steps K , and steps N per episode. It shows that a batch size M of 32 exhibits a more stable trend compared to M of 64 and 128, although their overall performances are similar. Q-network update steps K at 50 and 100 result in generally stable performances compared to K at 200. On the other hand, both 600 and

Fig. 14 Comparison of the reward evolution for task ϕ_2 with various batch sizes M , Q-network update steps K , and steps per episode N



(c) 900 steps/episode

Fig. 15 The office environment

900 steps per episode yield better convergence. However, considering the computational efficiency, we selected 600 steps per episode in this example.

5.2 Pybullet TurtleBot simulations

Figure 15 illustrates a virtual office environment, representing the scenario of TurtleBot simulation conducted with PyBullet 3.0 [38]. In the office space, there are four office rooms ‘a’, ‘b’, ‘c’, and ‘d’, a storage room ‘S’, a printer room ‘Print’, and a supply station ‘Sply’ to recharge the TurtleBot, i.e., the agent. In addition, there are two big windows in Offices ‘a’ and ‘d’ and multiple doors in the office space. We discretize this office space into a four-by-four grid world to generate the POMDP model. Considering the motion uncertainties, we assume that the TurtleBot has a probability of 0.9 to successfully execute its navigation controller by moving along the desired direction. However, it can move to other possible directions, uniformly sharing a probability of 0.1. Moving toward the wall will keep the TurtleBot stay at the same location. Assuming the agent is fully aware of the assigned tasks, We test the proposed model-free RL approach with two different observation settings in this office scenario. All simulations are conducted via 10,000 episodes with 300 steps per episode. The other computation settings are the same as in Sect. 5.1. We map the office scenario for each simulation to a grid world in which the optimal policy is learned. Then, we apply the derived optimal policy to the virtual TurtleBot in the PyBullet platform to validate the task accomplishment.

5.2.1 Observation of the surroundings

This setting assumes that at the current state, the TurtleBot can collect the surrounding observations in all four directions, following a specific order from ‘North’, ‘West’, ‘South’ to ‘East’. The observation elements are ‘wall’, ‘hallway’, ‘door’, and ‘window’. The agent can only observe one element in each direction. It shall be noted that there is only one

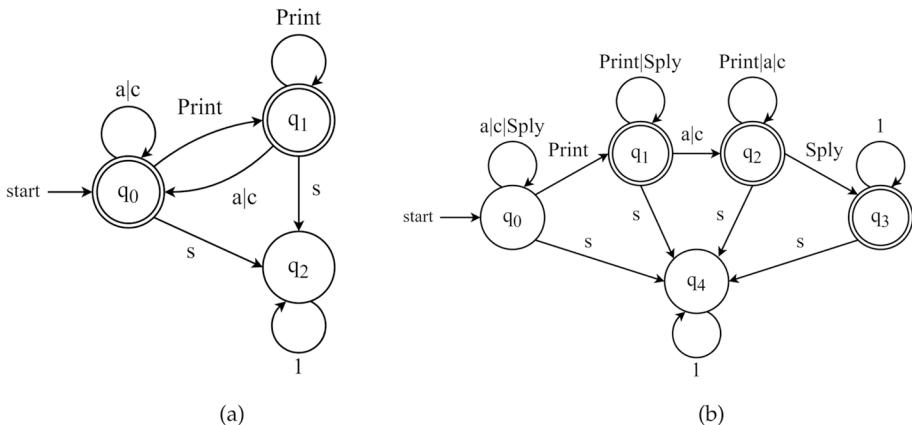


Fig. 16 The induced LDGBAs: **a** φ_{task1} . **b** φ_{task2}

observation at each state, i.e., $\Omega(s', a, o) = 1$. However, the agent may perceive the same observation in two or more states. For example, offices ‘b’ and ‘c’ have the same observation: $o(‘b’)=o(‘c’)=\{‘wall’ ‘wall’ ‘wall’ ‘door’\}$. Consequently, the set of observation O in this POMDP problem consists of 13 distinct observations.

Task 1: Task 1 requires that the TurtleBot visits the printer room to collect the documents and then carries the documents to Office ‘a’ or ‘c’, repeatedly. At the same time, the TurtleBot shall always avoid entering the storage room ‘S’. Similar to Eq. (13), this task can be expressed as an LTL formula in Eq. (15). Figure 16a shows the induced LDGBA, and we also employ the redesigned reward in Eq. (14).

$$\varphi_{task1} = \square \diamond (Print \wedge \diamond(a \vee c)) \wedge \square \neg S \quad (15)$$

After the training process is converged, the optimal policy can be derived from Q networks. The generated paths are then plotted in figures as straight lines to clearly demonstrate the successful task completion by the TurtleBot. In addition, a simulation in PyBullet was also conducted to provide a practical verification of the TurtleBot’s trajectory. Yet, this result is not pivotal to the primary aim of our paper, hence not included in the simulation results. Figure 17a illustrates a generated path with which the TurtleBot can complete Task 1. It can be seen that after leaving the initial state, office ‘b’, the TurtleBot moves towards the printer room, indicated via a yellow path. It visits offices ‘b’ and ‘c’ more than once on the way to the printer room due to motion uncertainty. After the TurtleBot arrives at the printer room and collects the documents, a blue path demonstrates that it leaves the printer room and moves to office ‘c’ for the delivery.

Task 2: Here, we extend Task 1 to a more complex task. Task 2 indicates that the TurtleBot must go to the supply station for recharge after delivering the documents and before repeating Task 1. The LTL formula of Task 2 can be expressed as Eq. (16), and Fig. 16b depicts the corresponding LDGBA, keeping the transitions with the single label only to simplify the illustration.

$$\varphi_{task2} = (\neg(a \vee c) \wedge Print) \wedge (\neg Sply \wedge (a \vee c)) \wedge (\diamond Sply) \wedge (\square \neg S) \quad (16)$$

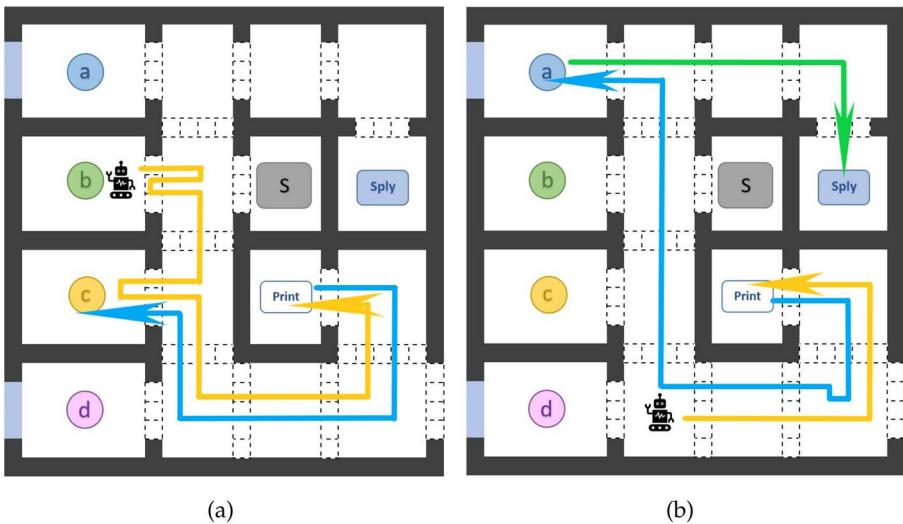


Fig. 17 Generated paths for the TurtleBot to accomplish tasks if it can observe surroundings in all four directions: **a** Task 1. **b** Task 2

Figure 17b shows a path generated from the learned policy for the agent to accomplish Task 2. The TurtleBot starts outside office ‘d’. It then moves to the printer room, collects the documents, and leaves for office ‘a’, indicated as the yellow and blue routes, respectively. Finally, the green route shows that the agent arrives at the ‘Sply’ station for recharge after delivering the documents.

5.2.2 Observation of a single direction

We also consider another observation setting, in which the TurtleBot is supposed to observe only one direction randomly at each state. Consequently, the observation uncertainty increases significantly. We add a few items in the PyBullet office environment to enhance policy convergence, as shown in Fig. 18. Therefore, the observable space includes ‘hallway’, ‘wall’, ‘door’, ‘window’, ‘table’, ‘paint on the wall’, and ‘flower by the wall.’ For example, the agent can observe ‘wall’, ‘table’, and ‘door’ with the probabilities of 50%, 25%, and 25%, respectively, in offices ‘b’ and ‘c’.

Figure 19 demonstrates the generated paths for the agent to accomplish the same tasks as in Sect. 5.2.1, respectively. Compared with the paths in Fig. 17 with four directional observations, the single observation element provides an agent with less sense of the current state. Figure 19 also indicates that the agent encounters difficulty deciding the right moves. For example, the yellow path in Fig. 19a illustrates that the agent moves back and forth a few times in the hallway with multiple paints on the wall before finally heading to the printer room. It is mainly because of observation uncertainty in addition to motion uncertainty.

Fig. 18 The modified office environment where the agent can observe only one direction

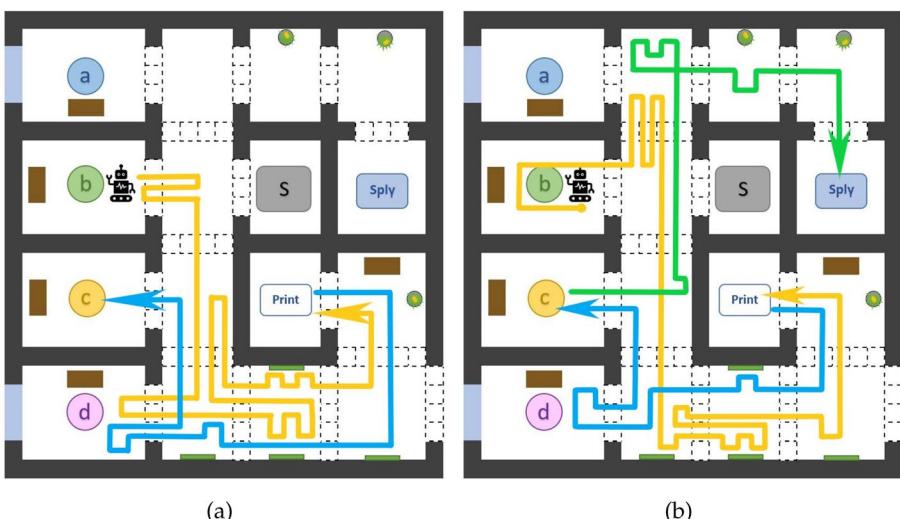


Fig. 19 The paths for the agent to accomplish the assigned tasks for a single round when the agent can observe only one direction: **a** Task 1. **b** Task 2

5.3 Multi-agent warehouse simulation

We also preliminarily apply the developed model-free RL approaches to a multi-agent problem. A mini-factory warehouse is modeled as an 8×8 grid world as shown in Fig. 20. There are two agents, and each must repeatedly move a box from one of the locations labeled as ‘a’ in blue to any spot labeled as ‘b’ in green to drop off the box on the convey belts. The darker and lighter gray states indicate the wall and other immovable packages. In this case, each agent can move ‘up’, ‘down’, ‘right’, ‘left’ and ‘stay’, but they cannot move

Fig. 20 The warehouse environment



to the same location simultaneously. Once both agents complete the task within one round, more boxes await at locations ‘a’. Each agent is rewarded for successfully moving a box to the goal location. In addition, the agents’ initial locations are shown in Fig. 20. The observation and transition probabilities are set as the same as in Sect. 2.3.

We can formulate the team task via LTL as $\phi = \square \diamond (a \wedge \diamond b)$, similar to the formula in Eq. (13). Also, we implement the Reward Redesign mentioned in Sect. 5.1.2 for task feasibility. Multi-agent reinforcement learning (MARL) studies how multiple agents interact in a common environment. According to agents’ actual task requirements, MARL can be categorized as the following three broad classes [39]:

- Cooperative: Agents cooperate to achieve the team goal, in which no agent can perform the whole task alone.
- Competitive: Agents compete against each other.
- Mixed: Agents maximize the utility that may require cooperating and/or competing.

In addition, the multi-agent system can also be modeled in a centralized or decentralized framework, where a central policy or multiple independent policies can be learned by the agents [39], respectively.

In this case, the agents work in the same environment towards a common goal. This MARL problem can be categorized as a cooperative case. We model this problem as a simplified decentralized POMDP (Dec-POMDP) framework. Each agent has a set of actions A^i and a set of observations O^i , where i denotes the index of agents: $i \in \{1, 2\}$. However, since they are identical agents in the same state space, we define $A^1 = A^2$ and $O^1 = O^2$. Only static events are considered in this case.

There have been numerous previous works to solve MARL problems from various perspectives. Inspired by the work of Zhou et al. [40], we set up cooperative communication between the agents and implement it into our model-free RL algorithms. We assume the agents have full knowledge of the task. Two independent Q networks are initialized for each agent. For each Q network, the input consists of the observation sequence, the q state sequence, and the q state sequence from the other agent. The agents share the task recognition information through the messages between the agents in the same communication network. The corresponding Q networks are shown as Fig. 21. Since two Q networks are

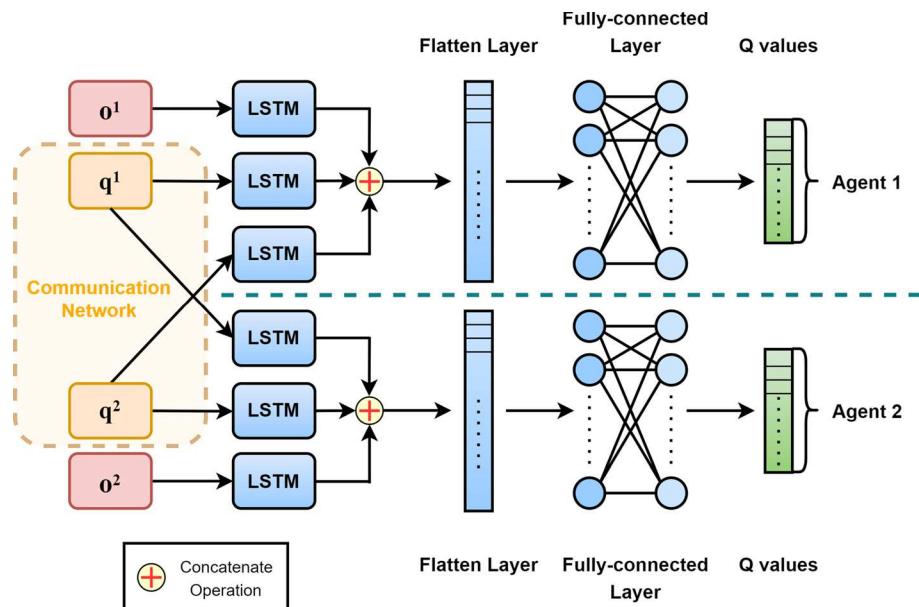
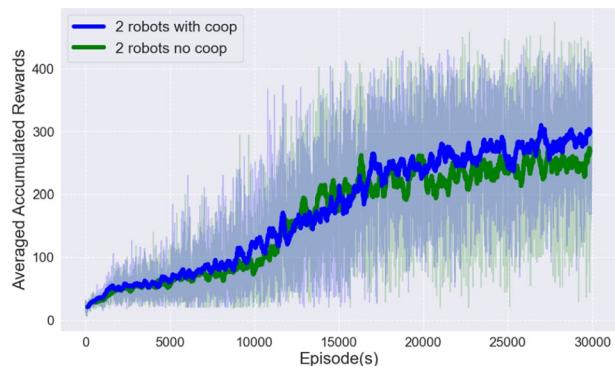


Fig. 21 Q network architectures for a MARL problem

Fig. 22 The accumulated rewards of the MARL case (two agents with cooperation or without cooperation)



trained independently to estimate Q values for each agent, we consider this approach a decentralized training and execution. The simulation took 30,000 episodes for 300 steps each. Figure 22 shows the trend of the accumulated reward vs. episode, where the case without cooperation is included for comparison.

Figure 23 shows the derived paths for both agents, represented by red and yellow routes, respectively. At the beginning of the task in Fig. 23a, both agents stay at their original positions to gather sufficient observations of the surrounding environment for decision-making. Shortly after, they leave for the loading zone in blue, and the yellow agent waits at a state on its way to prevent a collision with the other agent (red). Once the packages are loaded, the agents head towards the green zone to drop them off on the convey belts, as shown in Fig. 23b. The second run starts right after the completion, shown in Fig. 23c. It shall be noted that the yellow agent selects the action ‘stay’ a few times to wait for the other

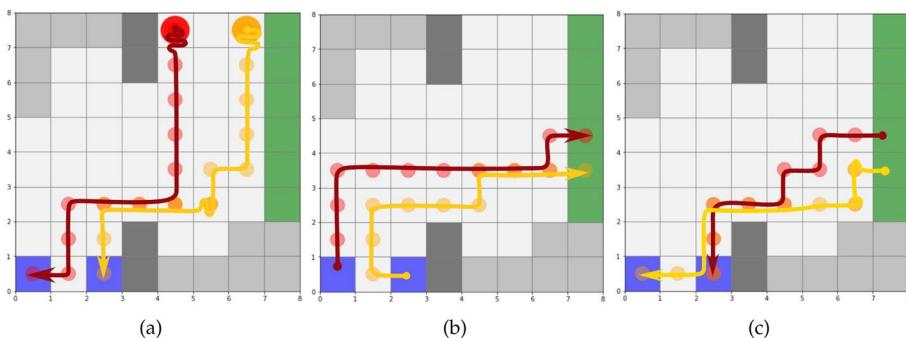


Fig. 23 The paths generated for the red and yellow agents to accomplish the assigned task

agent to pass because we prioritize the red agent. The generated paths show that the desired cooperative task is achieved even though separate policies are trained for the agents with limited communication. It is evident from Fig. 22 that the trend of rewards in the cooperative scenario shows a slight improvement compared to the non-cooperative setup. This variance could be attributed to the behavior observed in non-cooperative agents, as they tend to visit both pickup locations ('a' states) to ensure the retrieval of stored boxes.

6 Discussion

This study formulates the motion planning of autonomous agents in partially observable environments as a PL-POMDP problem. We also address high-level complex tasks by expressing them through LTLs and converting them to LDGBAs for model checking. Consequently, such a motion planning problem became equivalent to finding an optimal policy on the product of PL-POMDP and the induced LDGBA, and a model-free RL approach is proposed. We employ LDGBAs for model checking because they result in a smaller automaton state space than the corresponding DRAs. In addition, they have multiple accepting sets, enabling the agent to visit all accepting sets infinitely often. With the implementation of the reward redesign, the modified LDGBA can address the sparsity of rewards caused by LDBA in RL for motion planning [34]. In the future, we plan to leverage the tracking-frontier function proposed in [3] to keep track of non-visited accepting sets in LDGBA. This could be particularly useful when more complicated surveillance tasks for UAVs are required in partially observable environments. By further developing this function, we aim to improve the performance and applicability of our proposed method.

Deep Q learning is employed in the proposed model-free RL approach to learning optimal policies on the product PL-POMDP. Specifically, LSTM is implemented into Q network architectures to process the agent's observation history and task recognition. Using the induced automation state or perceived label history to represent task recognition depends on whether the agent is fully aware of the task. As the simulation examples demonstrate, either representation can enhance the agent's learning in partially observable environments subject to complex tasks. We also investigate the performances of Q networks by using different lengths of the input sequences. Choosing too long or too short sequences can lead to an unstable training process. The sequence lengths used in this study provide the agent with sufficient information for decision-making and avoid

lengthy delays before initiating action selection and following the trained policy. It shall be noted that the length selection may depend on the solved problem. In this paper, the automaton state and label sequences had the same length. Utilizing the automaton state sequence may outperform the label sequence if the environment has highly uncertain observations and complicated label assignments. For instance, in the second example with an office scenario, the label sequence was unable to provide the same level of information as the automaton state sequence. Additionally, this office example had a small state space, and using a long label sequence was inappropriate. As a result, the optimal policy could not be achieved when the agent was unaware of the task. In the future, while considering the scalability, long label sequences are expected to capture sufficient information for agent making optimal decisions. In addition, the proposed approach can be easily updated with other value-based or policy-based deep reinforcement learning (DRL) methods in future work.

We apply the proposed model-free RL approach to the motion planning of autonomous agents in discrete environments only in this study. Future works will extend this approach to tackle POMDP problems with continuous state and action spaces. Additionally, we will also consider the concept of associating labels with observations, which can be potentially more aligned with certain scenarios in real-world applications. Furthermore, a general LDGBA can consist of non-deterministic and deterministic sets. Although most LDGBA corresponding to the tasks considered in our simulation examples have deterministic sets only, one example demonstrates that the proposed approach can handle ϵ -transitions, which are non-deterministic. Even if the initial state is in the non-deterministic set, the automaton state transitions are restricted in the deterministic set after an ϵ -transition. Because the accepting states are in the deterministic set only, the proposed approaches can still achieve optimal policies.

We also provide a preliminary simulation of a cooperative multi-agent system to demonstrate the proposed approach's versatility and adaptability. By enabling the communication of task recognition between the agents, the acquired policies show a high level of collaboration that results in successful task completion, which has the potential to address a system of multiple UAVs for cooperative missions. In our scalability tests involving three and four agents within identical state spaces, we observe a substantial increase in computation times - double and triple respectively - compared to the scenario with two agents to achieve optimal policies. The escalating complexity and computation requirements are notable as the agent count increases, indicating potential limitations of our current approach. The limitation stems from our algorithm's reliance on each agent's Q network, which necessitates the utilization of task recognition sequences from all agents to aid individual decision-making. Consequently, this decentralized approach demands extensive communications between agents, posing computational challenges as the number of agents scales up, and often leads to non-stationary behavior [39]. Additionally, human intervention is required to modify the input features and structures of the DQN, enabling the neural network to adapt and accommodate more agents. In the future, it may be more appropriate to use a centralized training and decentralized execution (CTDE) framework capable of handling a continuous state space. Moreover, the agents' task recognition communication relies on their full knowledge of the tasks. An alternative approach can use local observations from other adjacent agents, so an additional neural network is needed to approximate the communication mechanism.

Acknowledgements Li and Xiao would like to thank US Department of Education (ED#P116S210005) and NSF (#2226936) for supporting this research.

Funding This research was funded by US Department of Education (ED#P116S210005) and NSF (#2226936).

Code availability Code files are available at https://github.com/JunchaoLi001/Model-free_DRL_LSTM_on_POMDP_with_LDGBA.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kurniawati, H. (2022). Partially observable Markov decision processes and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*. <https://doi.org/10.1146/annurev-contr-042920-092451>
2. Kaufman, H., & Howard, R. A. (1961). Dynamic programming and Markov processes. *The American Mathematical Monthly*. <https://doi.org/10.2307/2312519>
3. Cai, M., Xiao, S., Li, B., Li, Z., & Kan, Z. (2021). Reinforcement learning based temporal logic control with maximum probabilistic satisfaction (Vol. 2021). <https://doi.org/10.1109/ICRA48506.2021.9561903>
4. Cai, M., Xiao, S., Li, Z., & Kan, Z. (2021). Optimal probabilistic motion planning with potential infeasible ltl constraints. *IEEE Transactions on Automatic Control*. <https://doi.org/10.1109/TAC.2021.3138704>
5. Perez, A., Platt, R., Konidaris, G., Kaelbling, L., & Lozano-Perez, T. (2012). Lqr-rrt*: Optimal sampling-based motion planning with automatically derived extension heuristics. In *2012 IEEE International Conference on Robotics and Automation* (pp. 2537–2542). IEEE.
6. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
7. Shani, G., Pineau, J., & Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*. <https://doi.org/10.1007/s10458-012-9200-2>
8. Pineau, J., Gordon, G., & Thrun, S. (2003). Point-based value iteration: An anytime algorithm for pomdps.
9. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces (Vol. 4) (2009). <https://doi.org/10.15607/rss.2008.iv.009>
10. Li, J., Cai, M., Wang, Z., & Xiao, S. (2023). Model-based motion planning in pomdps with temporal logic specifications. *Advanced Robotics*, 37(14), 871–886.
11. Mnih, V., Silver, D., & Riedmiller, M. (2013). Playing atari with deep q learning. Nips.
12. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*. <https://doi.org/10.1038/nature14236>
13. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*. <https://doi.org/10.1145/3065386>
14. Hausknecht, M., & Stone, P. (2015). Deep recurrent q-learning for partially observable mdps (Vol. FS-15-06).
15. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*. <https://doi.org/10.1162/neco.1997.9.8.1735>
16. Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016). Learning to communicate to solve riddles with deep distributed recurrent q-networks.
17. Zhu, P., Li, X., Poupart, P., & Miao, G. (2017). On improving deep reinforcement learning for pomdps.
18. Heess, N., Hunt, J. J., Lillicrap, T. P., & Silver, D. (2015). Memory-based control with recurrent neural networks.
19. Meng, L., Gorbet, R., & Kulic, D. (2021). Memory-based deep reinforcement learning for pomdps. <https://doi.org/10.1109/IROS51168.2021.9636140>
20. O Baier, C., & Katoen, J.-P. (2008). Principles of model checking (Vol. 950).
21. Křetínský, J., Meggendorfer, T., & Sickert, S. (2018). Owl: A library for ω -words, automata, and ltl, vol. 11138 LNCS. https://doi.org/10.1007/978-3-030-01090-4_34

22. Chatterjee, K., Chmelík, M., Gupta, R., & Kanodia, A. (2015). Qualitative analysis of pomdps with temporal logic specifications for robotics applications, vol. 2015-June. <https://doi.org/10.1109/ICRA.2015.7139019>
23. Icarte, R. T., Waldie, E., Klassen, T. Q., Valenzano, R., Castro, M. P., & McIlraith, S. A. (2019). Learning reward machines for partially observable reinforcement learning, vol. 32.
24. Icarte, R. T., Klassen, T. Q., Valenzano, R., & McIlraith, S. A. (2022). Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73, 173–208.
25. Sharan, R., & Burdick, J. (2014). Finite state control of pomdps with ltl specifications. <https://doi.org/10.1109/ACC.2014.6858909>
26. Bouton, M., & Kochenderfer, M. J. (2020). Point-based methods for model checking in partially observable Markov decision processes. <https://doi.org/10.1609/aaai.v34i06.6563>
27. Ahmadi, M., Sharan, R., & Burdick, J. W. (2020). Stochastic finite state control of pomdps with ltl specifications.
28. Carr, S., Jansen, N., Wimmer, R., Serban, A., Becker, B., & Topcu, U. (2019). Counterexample-guided strategy improvement for pomdps using recurrent neural networks (vol. 2019). <https://doi.org/10.24963/jcai.2019/768>.
29. Carr, S., Jansen, N., & Topcu, U. (2020) Verifiable rnn-based policies for pomdps under temporal logic constraints (Vol. 2021). <https://doi.org/10.24963/ijcai.2020/570>.
30. Hahn, E. M., Perez, M., Schewe, S., Somenzi, F., Trivedi, A., & Wojtczak, D. (2019). Omega-regular objectives in model-free reinforcement learning (Vol. 11427). LNCS. https://doi.org/10.1007/978-3-030-17462-0_27
31. Cai, M., Hasanbeig, M., Xiao, S., Abate, A., & Kan, Z. (2021). Modular deep reinforcement learning for continuous motion planning with temporal logic. *IEEE Robotics and Automation Letters*. <https://doi.org/10.1109/LRA.2021.3101544>
32. Hasanbeig, M., Kantaros, Y., Abate, A., Kroening, D., Pappas, G. J., & Lee, I. (2019). Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees (Vol. 2019). <https://doi.org/10.1109/CDC40024.2019.9028919>
33. Hasanbeig, H., Kroening, D., & Abate, A. (2023). Certified reinforcement learning with logic guidance. *Artificial Intelligence*, 322, 103949.
34. Oura, R., Sakakibara, A., & Ushio, T. (2020). Reinforcement learning of control policy for linear temporal logic specifications using limit-deterministic generalized Büchi automata. *IEEE Control Systems Letters*. <https://doi.org/10.1109/LCSYS.2020.2980552>
35. Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*. <https://doi.org/10.1007/bf00992699>
36. Bozkurt, A. K., Wang, Y., Zavlanos, M. M., & Pajic, M. (2020). Control synthesis from linear temporal logic specifications using model-free reinforcement learning. <https://doi.org/10.1109/ICRA40945.2020.9196796>
37. Sickert, S., Esparza, J., Jaax, S., & Kretínský, J. (2016). Limit-deterministic Büchi automata for linear temporal logic (Vol. 9780). https://doi.org/10.1007/978-3-319-41540-6_17.
38. Coumans, E., & Bai, Y. PyBullet, a Python module for physics simulation for games, robotics and machine learning. <http://pybullet.org> (2016–2021)
39. Oroojlooy, A., & Hajinezhad, D. (2022). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*. <https://doi.org/10.1007/s10489-022-04105-y>
40. Zhou, W., Li, J., & Zhang, Q. (2022). Joint communication and action learning in multi-target tracking of uav swarms with deep reinforcement learning. *Drones*, 6(11), 339.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Junchao Li¹ · Mingyu Cai² · Zhen Kan³ · Shaoping Xiao¹

✉ Junchao Li
junchao-li@uiowa.edu

Mingyu Cai
mingyu.cai@ucr.edu

Zhen Kan
zkan@ustc.edu.cn

Shaoping Xiao
shaoping-xiao@uiowa.edu

¹ Department of Mechanical Engineering, Iowa Technology of Institute, The University of Iowa, Seamans Center, 3131, Iowa City, IA 52242, USA

² Department of Mechanical Engineering, University of California, Bourns Hall A342 900 University Ave., Riverside, CA 92521, USA

³ Department of Automation, University of Science and Technology of China, No. 443 Huangshan Road, Shushan District, Hefei 230022, Anhui, China