





Rule-Based Reinforcement Learning for Efficient Robot Navigation With Space Reduction

Yuanyang Zhu , *Student Member, IEEE*, Zhi Wang , *Member, IEEE*, Chunlin Chen , *Senior Member, IEEE*, and Daoyi Dong , *Senior Member, IEEE*

Abstract—For real-world deployments, it is critical to allow robots to navigate in complex environments autonomously. Traditional methods usually maintain an internal map of the environment, and then design several simple rules, in conjunction with a localization and planning approach, to navigate through the internal map. These approaches often involve a variety of assumptions and prior knowledge. In contrast, recent reinforcement learning (RL) methods can provide a model-free, self-learning mechanism as the robot interacts with an initially unknown environment, but are expensive to deploy in real-world scenarios due to inefficient exploration. In this article, we focus on efficient navigation with the RL technique and combine the advantages of these two kinds of methods into a rule-based RL (RuRL) algorithm for reducing the sample complexity and cost of time. First, we use the rule of wall-following to generate a closed-loop trajectory. Second, we employ a reduction rule to shrink the trajectory, which in turn effectively reduces the redundant exploration space. Besides, we give the detailed theoretical guarantee that the optimal navigation path is still in the reduced space. Third, in the reduced space, we utilize the Pledge rule to guide the exploration strategy for accelerating the RL process at the early stage. Experiments conducted on real robot navigation problems in hex-grid environments demonstrate that RuRL can achieve improved navigation performance.

Index Terms—Hex-grid, robot navigation, rule-based reinforcement learning, space reduction.

I. INTRODUCTION

Autonomous mobile robots are becoming ubiquitous in academia, industrial applications, and our daily life [1], [2]. As one of the fundamental topics in the research of mobile robots, robot navigation can be seen as a sequence of translations and rotations for finding the destination, while avoiding obstacles in the environment [3]. Enabling mobile robots to perceive and navigate through the surroundings is essential for their successful deployment in real-world scenarios [4].

Many algorithms have been proposed for path planning and optimization in robot navigation [5]. Traditional methods usually maintain an internal map of the environment and design simple rules to navigate through the internal map. Fuzzy logic methods use fuzzy rules like IF-THEN to make robot navigation decisions [6]. Neuro-fuzzy techniques combine neural networks with fuzzy rules to improve the tracking performance under uncertain physical interaction and external dynamics [7], [8]. However, it is challenging for human experts to choose the most appropriate rules and membership functions [9]. Another line is to use robotic navigation technologies inspired by biological behavior rules, such as genetic algorithms [10], particle swarm optimization [11], and ant colony optimization (ACO) [12]. Owing to the fact that these rules need to know the prior environment model and consider extensive possible situations in advance for mimicking the cognitive process of human experts to solve decision-making problems, rule-based methods tend to converge early to suboptimal policies [13]. Hence, the real-time performance in these methods may not be sufficient to meet the requirements of planning speed and accuracy in path planning tasks [14].

Recent reinforcement learning (RL) methods offer considerable potentials for mobile robot systems [15]. RL techniques are obliged to the idea of Markov decision processes (MDPs) that aim to directly solve the optimal sequential decision-making problem of learning from interaction to achieve the goal [16]. By observing the results of navigation decisions in the real world, mobile robots can directly learn from trial-and-error experience, continuously improving their proficiency and adapting to unknown environments [2]. In recent years, RL has been widely investigated in robot navigation domains due to its self-learning and online learning capabilities [17]. However, interacting with

Manuscript received November 19, 2020; revised January 16, 2021 and February 26, 2021; accepted April 2, 2021. Date of publication April 13, 2021; date of current version April 18, 2022. Recommended by Technical Editor G. Carbone and Senior Editor K. Kyriakopoulos. This work was supported in part by the National Natural Science Foundation of China under Grant 71732003, Grant 62073160, and Grant 62006111, in part by the National Key Research and Development Program of China under Grant 2018AAA0101100, in part by the Synergistic Innovation Center of Jiangsu Modern Agricultural Equipment and Technology under Grant 4091600002, and in part by the Australian Research Councils's Discovery Projects funding scheme under Project DP190101566. (*Corresponding author: Zhi Wang.*)

Yuanyang Zhu and Zhi Wang are with the Department of Control and Systems Engineering, School of Management and Engineering, Nanjing University, Nanjing 210093, China (e-mail: yuanyang@smail.nju.edu.cn; zhiwang@nju.edu.cn).

Chunlin Chen is with the Department of Control and Systems Engineering, School of Management and Engineering, Nanjing University, Nanjing 210093, China, and also with the Synergistic Innovation Center of Jiangsu Modern Agricultural Equipment and Technology, Zhenjiang 212013, China (e-mail: clchen@nju.edu.cn).

Daoyi Dong is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2600, Australia (e-mail: daoyidong@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at <https://doi.org/10.1109/TMECH.2021.3072675>.

Digital Object Identifier 10.1109/TMECH.2021.3072675

the real world can be expensive due to practical constraints such as power usage and human supervision [18]. Model-free RL systems are capable of solving complex MDPs in a variety of complex domains, but usually at the cost of a large amount of agent-environment experience due to their limited sample efficiency [2].

Rule-based machine learning (RBML) that combines rules with learning-based methods is a promising direction for utilizing the experts' knowledge to improve the learning performance. RBML usually covers any machine learning method that identifies, learns, or evolves "rules" to store, manipulate or apply by the learning system [19]. RBML has been widely studied in a variety of fields, such as learning classifier systems, association rule mining, and artificial immune systems, which successfully combines the efficiency of rules and the autonomy of machine learning to complete complex tasks in the real world [20]. In the RL community, some researchers employ rules to improve the learning performance in dynamic simulation systems [21] and robot manipulators' navigation tasks [22]. Nevertheless, few practical implementations of rule-based RL (RuRL) methods have been systematically investigated for robot navigation.

On one hand, traditional rule-based methods generally rely on the environment model and expert knowledge to solve robot navigation tasks, and tend to converge early to suboptimal policies. On the other hand, recent RL methods can learn the global optimal policies in a model-free way as the robot interacts with an initially unknown environment, but are expensive to deploy in the real world due to inefficient exploration [15], [23]. In this article, considering the abilities of the rule-based techniques for logic reasoning and RL methods for solving complex MDPs, we combine the advantages of these two methods into RuRL methods for efficient robot navigation tasks in hex-grid environments.¹

In summary, our main contributions are threefold as follows:

- 1) We design the rule of wall-following to obtain a closed-loop trajectory from the starting point to the goal. We maintain the main angle of view tracking and the priority of action selection strategy to ensure that the mobile robot walks along the left and the right walls, respectively.
- 2) We use the reduction rule to shrink the trajectory, which effectively reduces the exploration space. We traverse the obtained trajectory to determine whether there is a shorter path between two given states than the path on the trajectory. Besides, we provide the theoretical guarantee that the optimal path is still in the reduced space.
- 3) We employ the Pledge rule to guide the mobile robot to explore more efficiently at the early learning stage. Experimental results demonstrate the effectiveness and improved performance of RuRL for robot navigation.

These rules reduce the redundant space and accelerate the early exploration to provide coarse-grained learning, which is

¹Compared to the triangular and square grids, the hexagonal grid has six equidistant action directions with higher degree of freedom, and may better conform to uneven ground under the same unit area. The formed trajectory may be smoother in hex-grid maps [24]. Moreover, biological investigations [25] also suggest that neural cognition of spatial navigation is hexagonal. Hence, we rasterize environments into hexagonal grids here.

followed by fine-grained learning using the RL methods with improved efficiency. The efficiency is verified by experiments on real-world mobile robot systems in hex-grid environments.

The rest of this article is organized as follows. Section II introduces basic concepts of RL and related work about the efficient exploration methods for RL. Section III presents the integrated RuRL algorithm, including the rule of wall-following, the reduction rule, and the Pledge rule. The experimental results are discussed in Section IV. Finally, Section V concludes this article.

II. PRELIMINARIES AND RELATED WORK

A. Reinforcement Learning

RL is originated from the idea of MDPs in the field of optimal sequential decision-making problems. A finite MDP is a tuple of $\langle S, A, T, R, \gamma \rangle$, where S is the set of states, A is the set of actions, $T : S \times A \times S \rightarrow [0, 1]$ is the state transition probability upon taking action a in state s , $R : S \times A \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. A policy, $\pi : S \times A \rightarrow [0, 1]$, defines how a learner interacts with the environment by mapping perceived environmental states to actions, and $\sum_{a \in A} \pi(a|s) = 1 \forall s \in S$. The success of an agent depends on how to maximize the total rewards in the long run when acting under some policy π . The goal of RL is to find an optimal policy $\pi^* = \arg \max_{\pi} J(\pi)$ that maximizes the expected long-term return from the distribution

$$J(\pi) = \mathbb{E}_{\tau \sim \pi(\tau)}[r(\tau)] = \mathbb{E}_{\tau \sim \pi(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \quad (1)$$

where $\tau = (s_0, a_0, s_1, a_1, \dots)$ is the learning episode, $\pi(\tau) = p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$, r_t is the immediate reward received on the transition from s_t to s_{t+1} under action a_t .

The expected total reward when executing action a in state s is related to the optimal action-value function $Q^*(s, a)$ as

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_t = s, a_t = a \right] \quad (2)$$

and satisfies the Bellman optimality equation [26]

$$Q^*(s, a) = \mathbb{E}_{s'} \left[r + \gamma \max_{a'} Q^*(s', a') | s, a \right]. \quad (3)$$

For a discrete state-action space, the popular Q-learning [16] updates the action-value function with a learning rate α as

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]. \quad (4)$$

In learning, the transition tuples (s, a, r, s') are generated by a behavior policy that can be any exploration policy in principle. Using these transitions, the Q-function is iteratively updated until converging to the optimal value function $Q^*(s, a)$, and the optimal policy is naturally derived as $\pi^*(a|s) = \arg \max_a Q^*(s, a)$. To ensure that π converges to the optimal policy, we can select the behavior policy to be ε -soft (e.g., the ε -greedy policy) so that each state-action pair will be visited for an infinite number of times theoretically. More details about Q-learning can be found in [16].

B. Related Work

Learning control methods have been widely investigated for solving complex robot control problems. It signifies that the control system develops representations of the system's mathematical model and derives optimal control laws. Many learning control techniques have been applied to control systems, mainly consisting of iterative feedback tuning, control loop learning, and machine learning methods. For example, to cope with a class of second-order servo systems, the iterative feedback tuning approach employed numerical iterative optimization techniques based on Hessians of output errors and control signal data from the closed-loop system [27]. A control loop learning method applied two PID control loops to a parallel manipulator aiming to identify models of the robot with a manual approach [28]. To reduce traffic fatalities, supervised learning was employed to classify different movement events of pedestrians [29].

While RL has confirmed its ability to learn control strategies for various tasks, e.g., robot navigation, its performance in terms of sample efficiency is still a major challenge in complex applications. Many exploration techniques have been investigated to improve the learning efficiency of RL, which can be categorized into undirected and directed exploration strategies according to whether they utilize exploration-specific knowledge of the learning process itself. Undirected exploration strategies explore the environment based on randomness, such as ϵ -greedy [16], Boltzmann-distributed [16] and Gaussian noise methods [30]. Gaussian noise methods apply Gaussian noise to the action space or parameter space to generate noisy actions for provably efficient exploration [30]. Without utilizing any internal information of the learning process, these exploration strategies bring exponential regret in discrete MDPs and are limited to linear function approximations [31].

Directed exploration strategies utilize the previous history of the learning process and influence the portion of the environment explored in the future, including count-based [32], curiosity-driven [31], [33], and upper confidence bounds (UCB) exploration [34]. For tabular-based RL, count-based exploration strategies give an extra exploration bonus to frequently visited states [32]. In large or continuous state spaces, the pseudocounts methods employ the density model to obtain pseudocounts from the raw pixels and convert them into an exploration bonus [32]. Neural density model methods utilize the PixelCNN to provide an exploration bonus derived from an online density model [35]. In large state-action space where the states are rarely visited multiple times, count-based methods are easy to obtain suboptimal policies owing to paying more attention to visited states only [36]. By comparison, we use rules to efficiently reduce the redundant exploration space in complex environments, and theoretically prove that the optimal policy is still in reduced space.

In contrast to count-based methods, curiosity-driven exploration uses a mechanism for generating intrinsic reward signals towards seeking out state-action regions that the agent rarely explores [31]. Intrinsically motivated goal exploration processes explore more states, which are fewer experienced in disentangled goal space, and lead to more efficient exploration

than the entangled one [37]. The curious object-based search agent method [38] learns representations of the environment without extrinsic reward during the task-free exploration phase, and can be subsequently applied well in other tasks. Since the exploration bonus is not dependent on the reward, the main disadvantage is that the exploration may concentrate on irrelevant aspects of the environment [34]. In contrast, our method explores efficiently in smaller space, paying less attention to the irrelevant aspects of the environment.

Compared to these methods, the UCB methods design a mechanism for computing the upper confidence bounds of Q-values, and add decaying exploration bonuses to frequently visited states for optimistic exploration [39]. The discounted UCB1-tuned method considers the variance of reward, and uses the weighted variance of the Q-values to reduce exploration regrets [40]. UCB exploration via the Q-ensemble method computes the empirical mean and standard deviation of an ensemble of Q-value estimates to reduce the uncertainty for exploration [34]. UCB Bernstein approach achieves lower regrets by deriving a coarse bound on the empirical variance of value functions [39]. While these methods are more efficient with UCB exploration strategies, the agent may not efficiently learn in large state space since the regret scales linearly in the dimension of state space [41]. Here, the improved learning efficiency obtained by our method is more pronounced in a multiroom environment, which is supposed to benefit from the rule for efficiently reducing the redundant exploration space.

III. RuRL FOR NAVIGATION

In this section, we present the framework of RuRL with specific implementations of the rules for generating the closed-loop trajectory, reducing the exploration space, and guiding the early exploration strategy. Then, we give the integrated RuRL algorithm using these implementations.

A. Framework

We focus on the mobile robot navigation problem using RL in the hexagonal map environments. The simultaneous localization and mapping (SLAM) system running on the robot operating system (ROS) platform [42] is employed to construct the map of the unknown environment. The environment perception tasks can be tackled by utilizing the scan matching technique from a fusion between the lidar and ultrasonic sensors. An inertial measurement unit (IMU) sensor is used to estimate the rotational angle for improving the accuracy of the scan matching method. The map is created by Cartographer algorithms [42], and rasterized into hexagonal grids using the double-width coordinate system [43].²

One fundamental problem faced by RL for robot navigation is that the state space can be vast, and consequently, there may be a long delay before the reward is received. By applying a machine learning method to automatically discover useful rules, RBML allocates the learning mode in the cooperation rules, making the

²Based on two orthogonal axes, the double-width coordinate system steps to the right by 1 unit, and steps to the below by 2 units [44]

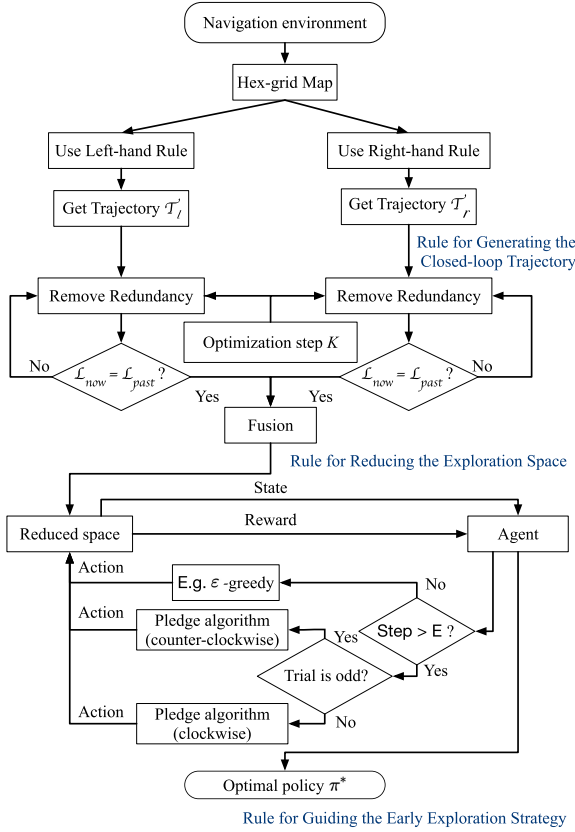


Fig. 1. Flow diagram of rule-based RL for robot navigation.

algorithm effective and flexible. The individually interpretable rules are clearly defined and applied to challenging tasks that are time-consuming or difficult for data-driven methods. These rules can model domain-specific knowledge and help speed up the RL process. Hence, we design several rules to facilitate the navigation performance using RL under a hex-grid map environment. First, a closed-loop trajectory is generated using two specific trajectories, i.e., T_l and T_r , obtained by the left- and right-hand rules, respectively. Second, based on the closed-loop trajectory, the space reduction rule is employed to form a reduced closed-loop state-action space for reducing redundant exploration space using a proper optimization step K . Finally, in the reduced space, when the number of learning steps exceeds a threshold value E without reaching the goal, the Pledge rule is utilized to guide the exploration strategy for finding the goal with fewer steps at the early learning stage. In the article, we adopt the widely used Q-learning as the basic implementation algorithm. The framework of RuRL is illustrated by the flow diagram, as shown in Fig. 1, and the rules and the integrated algorithm are presented in detail in the following sections.

B. Rule for Generating the Closed-Loop Trajectory

After the SLAM system obtains the environment map, we rasterize it into hexagonal grids, as shown in Fig. 2. In hexagonal grids, each state has more action options than in the square grids, making the planned path smoother. We adopt the widely

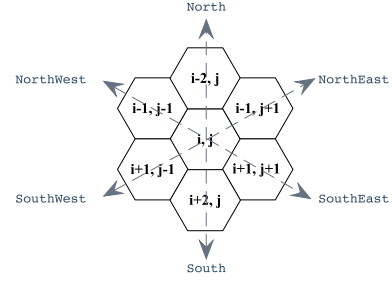


Fig. 2. Coordinate system of the hexagonal map.

used double-width coordinate system to calibrate the hexagonal environment. The origin of the coordinate system is at the top left corner, and the adjacent hexagonal grids along the horizontal and the vertical coordinate axis differ by one and two units, respectively. Let l and w denote the length and width of the map. Then, the numbers of columns and rows of the hex-grid map m and n are calculated as

$$\begin{cases} (n-1) * \frac{\sqrt{3}}{2} * a = w \\ (m+1) * \frac{3}{2} * a - a = l \end{cases} \quad (5)$$

where a is the hexagonal edge length. Given a state that corresponds to a hexagonal grid, there are six available actions. Correspondingly, from the first perspective of the mobile robot, the six available actions are: front (F), right front (RF), right rear (RR), rear (R), left rear (LR), and left front (LF).

Since the direction of the main angle of view changes when the robot moves, we need to record the previous action a_{t-1} to determine the perspective of the robot at the current time step t . For example, when $a_{t-1} = \text{RF}$, the main angle of view is the direction of the right front from the perspective of the mobile robot. Now, we aim to use the left- and the right-hand rules to generate a closed-loop trajectory along the wall. The right-hand rule is explained as follows, and the left-hand rule can be understood in a similar way. To design the right-hand rule of always walking right, we define the priority of action selection as $\text{RF} > F > \text{LF} > \text{LR} > R > \text{RR}$. That is, in any state s_t , the mobile robot will first try to choose the RF action if it can pass through the right front direction. If not, the mobile robot will try to select the action in the order of $F, \text{LF}, \text{LR}, R, \text{RR}$ until it can find a direction to take a valid step. The action selection strategy is executed following the right-hand rule until the mobile robot navigates to the goal point. We record the sequence of the states and actions as the right-hand trajectory T_r and the left-hand trajectory T_l as

$$\begin{aligned} T_r &= \{s_{r_1}, a_{r_1}, s_{r_2}, a_{r_2}, \dots, s_{r_p}\} \\ T_l &= \{s_{l_1}, a_{l_1}, s_{l_2}, a_{l_2}, \dots, s_{l_q}\} \end{aligned} \quad (6)$$

where p and q are the numbers of the traversed states in the right and the left trajectories, respectively. s_{r_p} and s_{l_q} are the same goal state. To better illustrate the left- and the right-hand rules, Fig. 3 presents a simple example of navigating in a hexagonal grid map, and Algorithm 1 summarizes the rule for generating the closed-loop trajectory.

Algorithm 1: Right-hand (Left-hand) Rule.**Input:** Starting point B ; Goal point G **Output:** Right-hand (left-hand) trajectory \mathcal{T}_r (\mathcal{T}_l)

```

1 Initialize  $\mathcal{T}_r = \emptyset$  ( $\mathcal{T}_l = \emptyset$ )
2 Initialize  $t = 0, s_t$ 
3 while  $s_t$  is not terminal do
4   Select  $a_t$  according the right-hand (left-hand) rule
5   Execute  $a_t$ , observe  $s_{t+1}$ 
6   Add  $(s_t, a_t)$  to  $\mathcal{T}_r$  ( $\mathcal{T}_l$ ) using Eq. (6)
7    $t \leftarrow t + 1$ 
8 end

```

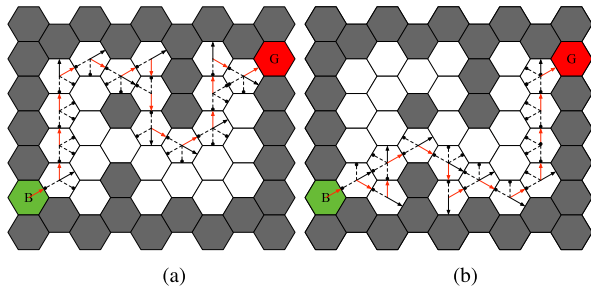


Fig. 3. Simple example of using the left- and right-hand rules to navigate in a hexagonal grid map. B is the starting point and G is the goal point. The selected actions are indicated by the red arrows. The other available actions are indicated by the dotted solid circular arrows. The main angle of view is indicated by the black arrows. (a) Left-hand rule. (b) Right-hand rule.

C. Rule for Reducing the Exploration Space

After obtaining the left- and the right-hand trajectories, the available states that the mobile robot can access are on or inside the closed loop. Obviously, the trajectory itself, \mathcal{T}_l or \mathcal{T}_r , is a feasible while not necessarily optimal path that navigates from the starting point to the goal. If we can properly reduce the length of the left- and right-hand trajectories, we will obtain a smaller closed-loop trajectory that can effectively reduce the redundant exploration space. Hence, we employ the reduction rule to optimize the two trajectories \mathcal{T}_l and \mathcal{T}_r , respectively. The reduction process on the right-hand trajectory is explained as follows, and the operation on the left-hand trajectory can be understood in a similar way.

Our main idea is that, given two states on the right-hand trajectory, we aim to find out whether there exists a shorter path between the two states than the path on the trajectory. If it does exist, we can replace the original path on the trajectory with the shorter one, thus obtaining a new trajectory with a reduced length. To implement this idea, we need to first formally define the step distance between two given states and the trajectory distance between two states on the trajectory.

Definition 1: (Step distance): The step distance is defined as the number of the least steps of actions needed to transit from s to s' (analogous to the definition in [45]). Specifically, the step distance between the same states is 0 and the step distance from one state to its adjacent states is set as 1.

Definition 2: (Trajectory distance): Given two states s and s' on the left- or the right-hand trajectory, their trajectory distance is defined as the number of the least steps of actions needed to transit from s to s' , while the intermediate states should also be on the trajectory.

Based on the definition of step distance and the fact that there are $6K$ K -step hexagonal grids around the center grid, we further define the K -step reachable states.

Definition 3: K -: (step reachable states): Given the state $s = (i, j)$, a state s' is called the K -step reachable state of s if the step distance between s' and s is K . There exist $6K$ states whose step distance from s is K , and these $6K$ states are called the K -step reachable states of s .

Based on the definition of K -step reachable states, we are able to easily figure out the potentially shorter path between two states on the trajectory. Given a state s on the right-hand trajectory, we first obtain its K -step reachable states. For example, when $K=2$, those 12 states are $(i-4, j), (i-3, j+1), (i-2, j+2), (i, j+2), (i+2, j+2), (i+3, j+1), (i+4, j+1), (i+3, j-1), (i+2, j-2), (i, j-2), (i-2, j-2), (i-3, j-1)$. If any reachable state is on the \mathcal{T}_{temp} , which is the sequence after the current state on the right-hand trajectory, we compute the trajectory distance from the given state to the reachable state. If the trajectory distance is greater than K , it indicates that we have discovered a shorter path between them instead of the original path on the trajectory. Hence, we can replace the original path with the new K -step path \mathcal{T}_j , resulting in an improved right-hand trajectory.³ We apply this reduction rule for every state in the right-hand trajectory, and obtain an optimized trajectory \mathcal{T}_r^K . In similar way, we can obtain the optimized left-hand trajectory \mathcal{T}_l^K . Together, a smaller closed-loop trajectory is formed. Algorithm 2 summarizes the rule for reducing the exploration space.

Fig. 4 presents a simple example of optimizing the closed loop trajectory. In **Fig. 4(a)**, the path trajectory has been marked as a light gray area, in which actions are selected using the left-hand rule. In **Fig. 4(b)**, from the start state to the end state of the trajectory obtained in **Fig. 4(a)**, we sequentially generate a 1-step reachable state \mathcal{T}_{reach}^1 for each state and match it with the subsequent trajectory. If the match is successful, we replace it with a new 1-step path, and otherwise we do the same for the next state. For example, the original path is $C \rightarrow C2 \rightarrow C3$ and $D \rightarrow D1 \rightarrow D2$, which can be optimized as $C \rightarrow C3$ and $D \rightarrow D2$. After optimization, we obtain the optimized closed-loop space, which is inside the light gray trajectories, as shown in **Fig. 4(c)**. In the closed-loop space formed after the optimization of $K = 1$, we will optimize with $K = 2$. The process is similar to that in **Fig. 4(a)–(c)**, as a detailed process shown in **Fig. 4(d)–(f)**.

Intuitively, the optimized trajectories can form an improved closed loop for more efficient exploration. Furthermore, Theorem 1 presents the theoretical analysis that the optimized trajectories correctly reduce the redundant exploration space.

³When the optimized step size is K , there are $2^K - 1$ K -step reachable paths, and \mathcal{T}_j is one of them.

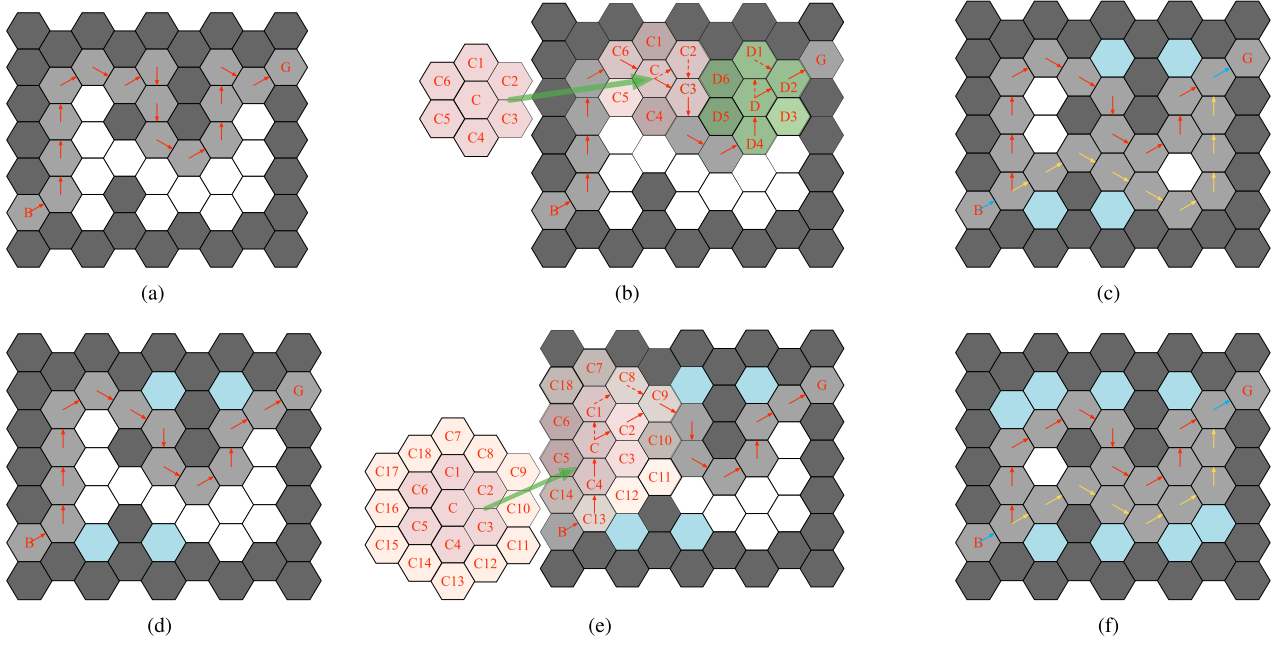


Fig. 4. Simple example of optimizing the closed-loop trajectory using different optimization steps of $K = 1$ (a)–(c) and $K = 2$ (d)–(f). The selected actions are indicated as red or yellow solid arrows. The actions before optimizing the trajectory are indicated as dotted arrows. The closed-up view on the left side of (b) and (e) shows the K -Step reachable states. The reduced exploration space is indicated as light blue hexagonal grids.

Algorithm 2: Rule for Reducing Exploration Space.

Input: Optimization step K ;
 left- and right-hand trajectories $\mathcal{T}_l, \mathcal{T}_r$
Output: Reduced trajectories $\mathcal{T}_l^K, \mathcal{T}_r^K$

```

1 for  $\mathcal{T}_{opt} = \mathcal{T}_l, \mathcal{T}_r$  do
2   Get  $l$  from the length of  $\mathcal{T}_{opt}$ 
3   while  $l$  is changing do
4     Update the length of  $l$ 
5     for  $i = 1, \dots, l - K$  do
6       Get  $\mathcal{T}_{temp}$  after the  $i$ -th state in  $\mathcal{T}_{opt}$ 
7       Obtain the  $K$ -step reachable states  $\mathcal{T}_{reach}^K$ 
8       if  $\mathcal{T}_{reach}^K$  matches state  $s_j$  in  $\mathcal{T}_{temp}$  then
9         Replace the sequence from  $s_i$  to  $s_j$  in
           $\mathcal{T}_{opt}$  with the new  $K$ -step path  $\mathcal{T}_j$ 
10        Update  $\mathcal{T}_{opt}$ 
11      end
12    end
13  end
14   $\mathcal{T}_l^K, \mathcal{T}_r^K \leftarrow \mathcal{T}_{opt}$ 
15 end
    
```

Theorem 1: Let s, s' be two states on the right-hand trajectory, with their step distance being K and their trajectory distance being J . Let $v(s')$ be the value function of state s' . Let π_1 and π_2 denote the policies of navigating from state s to state s' following the original path on the right-hand trajectory \mathcal{T}_r and the optimized path on the optimized right-hand trajectory \mathcal{T}_r^K , respectively. $v_{\pi_1}(s)$ and $v_{\pi_2}(s)$ are the value functions of state s when executing policy π_1 and π_2 , respectively. Then, for any $J \geq K$, we have $v_{\pi_1}(s) \leq v_{\pi_2}(s)$.

Proof: When executing policy π_1 , assume that $s_{\pi_1}^1, \dots, s_{\pi_1}^{J-1}$ are the sequential states between states s and s' . According to the Bellman equation [16], [26], the value function of state s is

$$\begin{aligned}
 v_{\pi_1}(s) &= \mathbb{E}_{\pi_1} \left[\sum_{i=0}^{\infty} \gamma^i r_i \mid s_0 = s \right] \\
 &= \sum_a \pi_1(a|s) \sum_{s''} p(s''|s, a) (r + \gamma v_{\pi_1}(s'')) \\
 &= r + \gamma v_{\pi_1}(s_{\pi_1}^1) = r + \gamma (r + \gamma v_{\pi_1}(s_{\pi_1}^2)) = \dots \\
 &= (r + \gamma r + \dots + \gamma^{J-1} r) + \gamma^{J-1} v_{\pi_1}(s') \\
 &= \frac{1 - \gamma^J}{1 - \gamma} r + \gamma^{J-1} v_{\pi_1}(s').
 \end{aligned}$$

In a similar way, the value function of state s when executing policy π_2 is

$$v_{\pi_2}(s) = \frac{1 - \gamma^K}{1 - \gamma} r + \gamma^{K-1} v_{\pi_2}(s').$$

We can set state s' as the terminal state. Then, we have $v_{\pi_1}(s') = v_{\pi_2}(s') = 0$, and

$$v_{\pi_1}(s) - v_{\pi_2}(s) = \frac{\gamma^K - \gamma^J}{1 - \gamma} r.$$

Since $J \geq K$, $0 \leq \gamma \leq 1$, and $r \leq 0$ in the navigation domains, we have $v_{\pi_1}(s) \leq v_{\pi_2}(s)$. ■

Theorem 1 proves the effectiveness of trajectory optimization. Suppose that \mathcal{T}^* is the optimal path in the original closed loop formed by trajectories \mathcal{T}_l and \mathcal{T}_r . Next, we present and prove

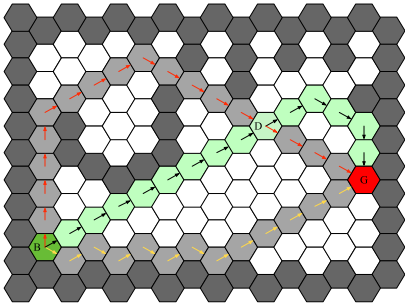


Fig. 5. Illustrative example for proving that the optimal path is still within the closed-loop \mathcal{C} . The trajectories $\mathcal{T}_l^K, \mathcal{T}_r^K$ are indicated as red and yellow solid arrows. The optimal path assumed \mathcal{T}^* is indicated as black solid arrows. The intersecting positions of $\mathcal{T}_l^K, \mathcal{T}^*$ are indicated as D and G .

Theorem 2, which demonstrates that the optimal path \mathcal{T}^* is still within the closed loop formed by \mathcal{T}_l^K and \mathcal{T}_r^K .

Theorem 2: Let \mathcal{C} denote the closed-loop region formed by \mathcal{T}_l^K and \mathcal{T}_r^K . For any optimal path \mathcal{T}^* , we have $\mathcal{T}^* \subseteq \mathcal{C}$.

Proof: We assume that there is an optimal path \mathcal{T}^* , and $\mathcal{T}^* \not\subseteq \mathcal{C}$. Since both the starting and goal points are on \mathcal{C} , \mathcal{T}^* , and \mathcal{T}_l^K (or \mathcal{T}^* and \mathcal{T}_r^K) intersect at least twice. We assume that the two intersecting positions are D and G , as shown in Fig. 5. The trajectory distances from B to G on \mathcal{T}_l^K and \mathcal{T}^* are J and I , respectively. The trajectory distances from D to G on \mathcal{T}_l^K and \mathcal{T}^* are j and i , respectively. Let π^*, π_1 , and π_2 denote policies of navigating from state B to state G following the path on \mathcal{T}^* , the path on \mathcal{T}_l^K , and the path composed of the trajectory $B \rightarrow D$ on \mathcal{T}^* and $D \rightarrow G$ on \mathcal{T}_l^K , respectively. Since \mathcal{T}^* is the optimal path, we have $I < J$, $v_{\pi^*}(B) > v_{\pi_1}(B)$, and $v_{\pi^*}(B) > v_{\pi_2}(B)$. According to the reduction rule, we have $j < i$. Furthermore, we have $v_{\pi_2}(D) > v_{\pi^*}(D)$ according to Theorem 1. Since π^* and π_2 share the same subpath from B to D , we have $v_{\pi_2}(B) > v_{\pi^*}(B)$. It is contradictory to the assumption that \mathcal{T}^* is the optimal path. Hence, the assumption does not hold. Then, for any \mathcal{T}^* , we have $\mathcal{T}^* \subseteq \mathcal{C}$. ■

When the optimization step is K , there are $6K$ hexagon grids around the center point to be optimized. Let n denote the length of the trajectory to be optimized. Then, the complexity of the optimization algorithm is $6nK$, i.e., $O(nK)$. It can be observed that selecting K with an enormous value will linearly increase the computational cost of trajectory reduction. On the other hand, a larger K generally leads to a smaller closed-loop trajectory, which can reduce the redundant exploration space to a more considerable extent. In practice, a moderate value of K (e.g., 2 – 4) is sufficient to obtain efficient performance.

D. Rule for Guiding the Early Exploration Strategy

In the reduced exploration space, we use the closed loop formed by the optimized trajectories as the new navigation environment. While the agent can learn more efficiently in the reduced space, the agent still needs to explore many steps to find the goal point at the early stage. We employ the Pledge rule [46] to accelerate the early learning performance when the number

Algorithm 3: Pledge Algorithm (Counter-clockwise).

Input: Sum of turns θ ; current state s ; turns θ' at state s ; previous action a_{past} ; goal point G

Output: Action a ; updated θ

```

1 Initialize  $\theta' = 0, a_{past}$ 
2 if  $\theta = 0$  at state  $s$  then
3   Select  $a$  according to  $\Theta_0$  rule and  $a_{past}$ 
4   Calculate the turns  $\theta'$ 
5    $\theta += \theta'$ 
6    $a_{past} \leftarrow a$ 
7 else
8   Select  $a$  according the right-hand rule and  $a_{past}$ 
9   Calculate the turns  $\theta'$ 
10   $\theta += \theta'$ 
11   $a_{past} \leftarrow a$ 
12 end

```

of steps exceeds a threshold value without finding the goal. The counter-clockwise method of the Pledge rule is explained as follows. Similarly, the clockwise way operates.

In the Pledge rule (counter-clockwise), the agent first needs to choose an initial action direction, and moves towards this action direction with priority. Next, to meet the priority of the initial action direction selected, we employ the sum of turns θ to record the changes of the action direction and update the main angle view of the mobile robot by the previous action a_{past} . When an obstacle is met, the sum of turns θ is added by 1 per 60° if the clockwise turn is positive, and is subtracted by 1 per 60° otherwise. Finally, for the purpose of avoiding traps, if the overall turning angle θ is 0, the agent will take action in the priority order of $F > LF > LR > R > RR > RF$ (defined as Θ_0 rule). Otherwise, the agent will choose an action in the priority order of $RF > F > LF > LR > R > RR$, which operates in the same way as the right-hand rule in Section III-B. By recording the sum of the turns θ and keeping the initial direction, the Pledge algorithm can find the goal point, regardless of the initial position of the agent. The clockwise method that counts the overall turning angle θ is opposite to the counter-clockwise method, and the Pledge algorithm is summarized in Algorithm 3.

To utilize the Pledge algorithm to improve the exploration efficiency, we use the counter-clockwise method when the number of learning episodes η is odd and employ the clockwise method otherwise. We design a decay function for the threshold value of learning steps as

$$E = \frac{M_{\max}}{\omega * \eta + b} \quad (7)$$

where M_{\max} is the maximum learning steps per episode and η is the number of the current episode. b tunes the expected number of threshold steps in previous episodes, and ω controls the decay rate. Generally, a smooth decay function (e.g., $\omega = 0.15$ and $b = 10$) can obtain effective improvement with the Pledge rule for guiding the agent to explore.

Algorithm 4: RuRL Algorithm.

Input: Learning rate α ; small ε ; Pledge rule usage episodes N ; threshold steps of each episode E ; discounting factor γ ; optimization step K

Output: Optimal policy π^*

```

1  $\mathcal{T}_l, \mathcal{T}_r \leftarrow$  Get trajectories using Algorithm 1
2  $\mathcal{T}_l^K, \mathcal{T}_r^K \leftarrow$  Optimize the trajectories using Algorithm 2
3 Env  $\leftarrow$  Connect two optimized trajectories  $\mathcal{T}_l^K, \mathcal{T}_r^K$ 
4 Initialize  $Q(s, a)$  arbitrarily,  $\forall s \in S, a \in A(s)$ 
5 for  $\eta$  up to  $T_{max}$  do each episode
6   Initialize  $s$ 
7   for  $s$  is terminal do each step of episode
8     if  $\eta \leq N$  and steps  $\geq E$  using Eq. (7) then
9       if  $\eta$  is odd then
10         Choose  $a$  using policy derived from
           Pledge Rule (counter-clockwise)
11       else
12         Choose  $a$  using policy derived from
           Pledge Rule (clockwise)
13       end
14     else
15       Choose  $a$  using policy derived from  $Q$ 
16     end
17     Take action  $a$ , observe  $r, s'$ 
18      $Q(s, a) \leftarrow$ 
        $Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$ 
19      $s \leftarrow s'$ 
20   end
21 end

```

E. Integrated RuRL Algorithm

With the abovementioned implementations, the integrated RuRL for navigation algorithm is summarized in Algorithm 4. First, in the original environment, the RL agent obtains trajectories $\mathcal{T}_l, \mathcal{T}_r$ through left- and right-hand rules in Line 1. Second, we employ the rule for optimizing the initial trajectories and reducing exploration space in Lines 2 and 3, and generate a smaller navigation environment. Third, in the new environment with the reduced space, we employ RL to learn the optimal policies in Lines 4–21. Finally, we employ the Pledge rule to accelerate the early learning performance for a small number of episodes when the number of steps exceeds a threshold value without finding the goal in Lines 8–13.

Remark 1: When the starting point is not adjacent to the wall, the Pledge rule can solve this problem [46]. When the goal point is not adjacent to the wall, we can solve it by setting the subtarget point near the goal point. To highlight the usage of rules, we only consider the situation where both the starting and goal points are adjacent to the wall in this article.

IV. EXPERIMENTS

We conduct two sets of experiments to evaluate the feasibility and effectiveness of RuRL. One is the single-room navigation tasks consisting of the obstacle-free map and the map with

obstacles. The other is the multiroom navigation with a large and complex map, where conventional methods tend to explore inefficiently or converge to suboptimal policies.

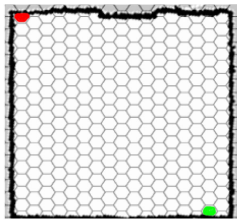
A. Experimental Settings

Since we aim to utilize rules to improve the exploration efficiency of RL in navigation domains, we focus on comparing RuRL to three baselines: RL without rules, RL with count-based exploration, and RL with UCB-based exploration [35], [40]. We employ the learning curve and total learning steps as the performance metrics. For single-room experiments, the Q-learning [16] algorithm with the ε -greedy strategy is investigated to evaluate the effectiveness of RuRL. Furthermore, the Q-learning [16] and SARSA [16] algorithms with the ε -greedy and the Softmax exploration strategies are investigated in the complex task. Besides, we compare our method with classic robot navigation methods, including A* and ACO heuristic algorithms, and utilize the length of the final planned path and the direction switching times of the path as the performance metrics. More details can be found in [47] and [48].

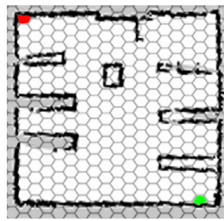
In all environments, we use the double-width coordinate system, where the state is the two-dimensional coordinate (i, j) , and available actions are: North towards $(i - 2, j)$, Northeast towards $(i - 1, j + 1)$, Southeast towards $(i + 1, j + 1)$, South towards $(i + 2, j)$, Southwest towards $(i + 1, j - 1)$, and Northwest towards $(i - 1, j - 1)$. When robots collide with obstacles, they bounce to the previous position. For all experiments, the reward is 100 if reaching the target, -100 if heading towards obstacles, and -1 otherwise. The hyperparameters are the same for all tested algorithms in each group of experiments: learning rate $\alpha = 0.01$ and discount factor $\gamma = 0.99$. Additionally, in the count-based exploration strategy, we use $N(s, a)$ to explicitly refer to the number of visits of a state-action pair in the learning process, and use an exploration bonus of the form $R^+(s, a) = \sqrt{\frac{\beta}{\log[N(s, a) + 1]}}$, where β is set as a constant 0.4. In RL with UCB-based method, we use the same exploration ratio settings as the ε -greedy strategy, and set the same parameters to all the tested algorithms: the damping factor $d = 0.9$ and the tendency of exploration constant $C' = 0.01$. More details about parameter settings of RL with the UCB-based method can be found in [40]. Additionally, the Euclidean distance is used for the heuristic function in the A* algorithm. In the ACO method, the number of ants is set as 100, and the other parameter settings can be found in [48]. The navigation maps, constructed from real environments by a SLAM mobile robot with high-precision lidar, ultrasonic, and IMU sensors running on the ROS Kinetic platform, are hexagonally rasterized on the MATLAB simulation platform. The environment perception control systems operate on an industrial PC (CPU: ARMv8 1.2 GHz, GPU: 400 MHz VideoCore IV). All the experiments are carried out on an Intel Core i7-7700 3.60 GHz PC with 16 GB RAM under Windows 10. The experimental results given are averaged over 50 runs.



(a)



(b)



(c)

Fig. 6. Physical size of the single-room navigation experimental environments is 465×458 cm. Using (5), the environment is rasterized into a hexagonal map with 35 rows and 19 columns after setting the grid length as $a = 15.8$ cm. (a) Real environment. (b) Env 1: obstacle-free. (c) Env 2: with obstacles.

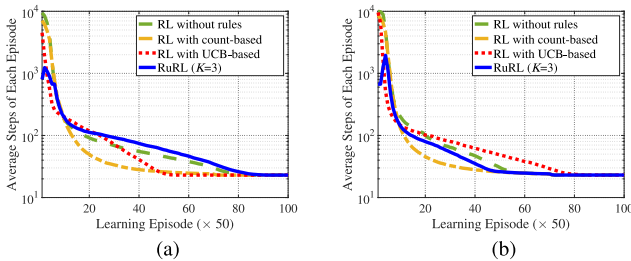


Fig. 7. Average steps per episode of RL without rules, RL with count-based, RL with UCB-based, and RuRL in single-room navigation tasks. (a) Env 1. (b) Env 2.

B. Tasks in Single-Room Environments

The real single-room environment is shown as in Fig. 6(a). Using a SLAM mobile robot, we constructed maps of the obstacle-free environment and the environment with obstacles, as shown in Fig. 6(b) and (c), respectively. The maps are rasterized with hexagonal grids, where the gray indicates the unknown area, the black indicates obstacles, and the white indicates the feasible area. For both environments, the starting point is set as (34, 17) with the green mark, and the goal point is set as (3, 2) with the red mark. The maximum number of learning episodes is set as 7000, and the maximum number of learning steps per episode is set as $M_{\max} = 10000$. The optimization step K is set as three for reducing the exploration space. The exploration ratio is $e^{-0.001*\eta}$ ($\eta < 3500$) and 0 otherwise. Parameters of the Pledge rule are set as $N = 100$ and $E = \frac{10000}{0.2*\eta+8}$. Fig. 7 shows the learning curve and Table I presents corresponding numerical results.

TABLE I

NUMERICAL RESULTS IN TERMS OF TOTAL LEARNING STEPS OF ALL TESTED ALGORITHMS IN SINGLE-ROOM TASKS

Implementation algorithm	Q-learning (ϵ -greedy)	
	Env 1	Env 2
RL without Rules ($\times 10^5$)	18.18	19.06
RL with count-based ($\times 10^5$)	14.90	15.60
RL with UCB-based ($\times 10^5$)	11.27	11.08
RuRL ($K = 3$) ($\times 10^5$)	8.82	7.14

First, all tested methods are implemented by the Q-learning algorithm. The ϵ -greedy exploration strategy is used for RuRL, RL without rules, and RL with count based. As shown in Fig. 7(a), all methods can find the optimal path with 23 steps in Env 1. We observe that RL with count-based, RL with UCB-based, and RuRL methods perform well compared with RL without rules. The RL with count-based method demonstrates that internal reward signals positively affect the middle learning stage. Compared to the RL without rules, RL with UCB-based approach obtains slightly superior performance with reduced regrets for efficient exploration throughout the learning process. In contrast, RuRL obtains the best performance by improving the exploration efficiency, which may benefit from the Pledge rule, where the efficiency of the rule for reducing the redundant space is minor. Table I shows that the steps are already reduced by 51.46% when $K = 3$ compared to RL without rules. Under the situation that they all find the optimal path, A* and ACO algorithms need to switch directions for 11 and 13 times, respectively, while RuRL only requires three times.⁴ Frequently switching directions will slow down the speed of mobile robots with more energy consumed and may be critical for the performance of the robot motion kinematics [49].

Next, we test RuRL in Env 2 with obstacles, as shown in Fig. 6(c). From Fig. 7(b), we can find that all tested algorithms can obtain the optimal path with 23 steps. We also find that RuRL has greater performance improvement than that in Env 1, but the performance of RL with count-based and RL with UCB-based methods is slightly improved. When there are more obstacles, the rules of reducing the redundant space play a more active role in enhancing exploration efficiency. It can be obtained from Table I that the learning steps of RuRL are nearly 62.52% less than that of RL without rules. In general, RuRL can accelerate the learning process to a large extent, especially in the environments with obstacles. Furthermore, the planned paths of A*, ACO, and RuRL methods need to switch directions for 11, 9, and 7 times, respectively.⁴

C. Tasks in Multiroom Environments

To further test the performance of RuRL, we use a larger and more complex navigation environment composed of multiple rooms, as shown in Fig. 8. The starting point and target point of Env 3 are set as (85, 56) and (3, 22). The maximal numbers of learning episodes and learning steps per episode are set as

⁴See Appendix A in Supplementary Materials for details.

TABLE II
HYPERPARAMETERS IN ENV 3

Implementation algorithm	Q-learning (ϵ -greedy)	Q-learning (Softmax)	SARSA (ϵ -greedy)	SARSA (Softmax)
Exploration ratio	$\begin{cases} e^{-0.001*\eta} (\eta < 5000) \\ 0 \end{cases}$	$\begin{cases} \frac{35}{\eta*0.011+1} (\eta < 3000) \\ 1 \end{cases}$	$\begin{cases} \frac{1}{\eta*0.4} (\eta < 500) \\ 0 \end{cases}$	$\begin{cases} \frac{35}{\eta*0.011+1} (\eta < 3000) \\ 1 \end{cases}$
# of episodes with Pledge rule N	700	1000	500	1000
Threshold steps of each episode E	$\frac{20000}{0.1*\eta+8}$	$\frac{20000}{0.1*\eta+10}$	$\frac{20000}{0.1*\eta+10}$	$\frac{20000}{0.1*\eta+10}$

TABLE III
NUMERICAL RESULTS IN TERMS OF TOTAL LEARNING STEPS OF ALL TESTED ALGORITHMS IN MULTIROOM TASKS (ENV 3)

Implementation algorithm	RL without rules ($\times 10^7$)	RL with count-based ($\times 10^7$)	RL with UCB-based ($\times 10^7$)	RuRL ($K=3$) ($\times 10^6$)	Maximal reduction (%)
Q-learning (ϵ -greedy)	1.69	1.17	1.67	4.77	71.79%
Q-learning (Softmax)	2.36	2.51	1.24	7.08	69.93%
SARSA (ϵ -greedy)	1.23	0.79	1.24	4.72	61.69%
SARSA (Softmax)	2.78	1.99	1.24	7.58	72.78%

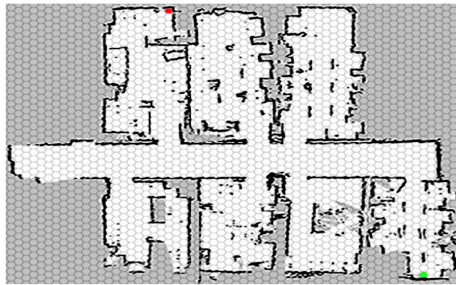


Fig. 8. Env 3: the multiroom navigation environment (1640 cm \times 1960 cm) rasterized into an 87×59 hexagonal grid map using (5).

15 000 and 20 000, respectively. The bonus coefficient of count-based exploration strategy β is set as 0.5. Other hyperparameters corresponding to the exploration strategy and the Pledge rule can be found in Table II.

In the multiroom task, the A* and ACO approaches obtain suboptimal paths with 73 and 74 steps and need to switch directions for 29 and 27 times, respectively. In contrast, RuRL requires only 14 times of switching directions while learning the optimal path with 72 steps, where RuRL is implemented by the Q-learning algorithm with ϵ -greedy exploration strategy. Compared with A* and ACO algorithms, RuRL can find smoother paths with a potentially better optimality guarantee.⁴

Furthermore, Fig. 9 presents the learning steps of all tested algorithms implemented by Q-learning and SARSA with the ϵ -greedy and Softmax strategies, and Table III shows corresponding numerical results. It is clear that the number of learning steps of RL with count-based, RL with UCB-based, and RuRL methods is lower than that of RL without rules. The RL with count-based method improves the performance with the ϵ -greedy strategy, and obtains suboptimal policy with 74 steps due to paying more attention to visited states. However, it obtains reduced performance improvement with the Softmax strategy. Compared to the single-room tasks, the improvement of RL with UCB based is reduced to 1.18% since the regret scales linearly in the dimension of the state space. In contrast, RuRL enables

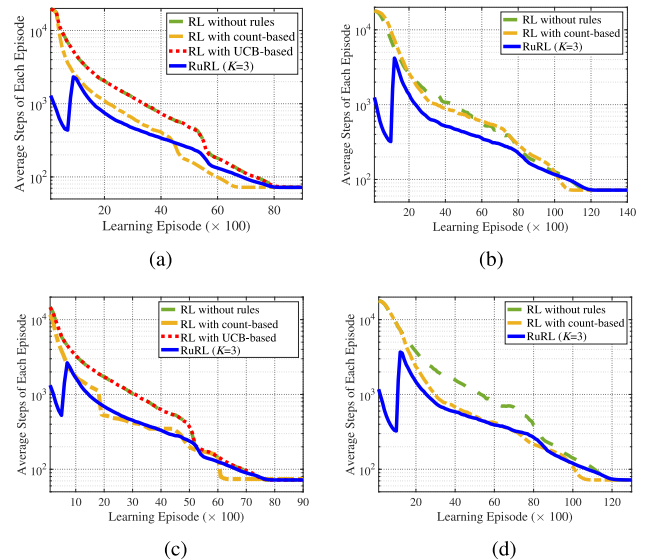


Fig. 9. Performance of all tested methods implemented by Q-learning and SARSA in the multiroom environment. The ϵ -greedy and Softmax strategies are used for RuRL, RL with count based, and RL without rules. (a) Q-learning with ϵ -greedy. (b) Q-learning with Softmax. (c) SARSA with ϵ -greedy. (d) SARSA with Softmax.

the agent to make quicker progress than the others on finding the optimal policy. Taking the implementation of Q-learning with ϵ -greedy strategy as an example, the total learning steps are reduced by 71.79% when using the proposed rules with $K=3$. Consistent with observations in Section IV-B, RuRL better improves the learning performance in multiroom tasks, which is supposed to benefit from the distinct space reduction and the Pledge rule for finding the goal with fewer steps.

In addition, we analyze the learning performance of RuRL as the optimization step K increases. We also adopt a statistical approach to analyze the relationship between the number of episodes N with the Pledge rule employed and the performance of RuRL (see Appendix B and Appendix C in Supplementary Materials for details).

V. CONCLUSION

In this article, we propose a RuRL algorithm for efficient robot navigation with space reduction, where three rules are applied to reduce the redundant exploration space and guide the exploration strategy. Then, we evaluate RuRL on the single-room environments and a multiroom environment, where the maps are built using a SLAM mobile robot. Experimental results demonstrate that RuRL can efficiently improve the navigation performance with good scalability. Our future work will focus on more practical rules for advanced RL methods in the field of complex robotic control.

REFERENCES

- [1] A. Faust *et al.*, "PRM-RL: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 5113–5120.
- [2] C. Chen, H.-X. Li, and D. Dong, "Hybrid control for robot navigation-A hierarchical Q-learning algorithm," *IEEE Trans. Robot. Autom.*, vol. 15, no. 2, pp. 37–47, Jun. 2008.
- [3] Z. Wang, C. Chen, H.-X. Li, D. Dong, and T.-J. Tarn, "Incremental reinforcement learning with prioritized sweeping for dynamic environments," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 2, pp. 621–632, Apr. 2019.
- [4] Z. Li, T. Zhao, F. Chen, Y. Hu, C. Su, and T. Fukuda, "Reinforcement learning of manipulation and grasping using dynamical movement primitives for a humanoidlike mobile manipulator," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 121–131, Feb. 2018.
- [5] R. C. Holte, M. B. Perez, R. M. Zimmer, and A. J. MacDonald, "Hierarchical A*: Searching abstraction hierarchies efficiently," in *Proc. Nat. Conf. Artif. Intell. Innovative Appl. Artif. Intell. Conf.*, 1996, vol. 1, pp. 530–535.
- [6] J. H. Lilly, "Evolution of a negative-rule fuzzy obstacle avoidance controller for an autonomous vehicle," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 4, pp. 718–728, Aug. 2007.
- [7] J. Li *et al.*, "Neural fuzzy approximation enhanced autonomous tracking control of the wheel-legged robot under uncertain physical interaction," *Neurocomputing*, vol. 410, pp. 342–353, 2020.
- [8] J. Li *et al.*, "Parallel structure of six wheel-legged robot trajectory tracking control with heavy payload under uncertain physical interaction," *Assem. Autom.*, vol. 40, no. 5, pp. 675–687, 2020.
- [9] C. Fu, A. Sarabakha, E. Kayacan, C. Wagner, R. John, and J. M. Garibaldi, "Input uncertainty sensitivity enhanced nonsingleton fuzzy logic controllers for long-term navigation of quadrotor UAVs," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 2, pp. 725–734, Apr. 2018.
- [10] Y. Hu and S. X. Yang, "A knowledge based genetic algorithm for path planning of a mobile robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2004, pp. 4350–4355.
- [11] Y.-D. Hong and B. Lee, "Real-time feasible footstep planning for bipedal robots in three-dimensional environments using particle swarm optimization," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 1, pp. 429–437, Feb. 2020.
- [12] B. Englot and F. Hover, "Multi-goal feasible path planning using ant colony optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 2255–2260.
- [13] S. S. A. Naser and M. A. Al-Nakhal, "A ruled based system for ear problem diagnosis and treatment," *World Wide J. Multidiscip. Res. Dev.*, vol. 2, no. 4, pp. 25–31, 2016.
- [14] H. Ju, J. Zhang, X. Pei, and G. Xing, "Evolutionary fuzzy navigation for security robots," in *Proc. World Congr. Intell. Control Autom.*, 2008, pp. 5739–5743.
- [15] D. Dong, C. Chen, J. Chu, and T. Tarn, "Robust quantum-inspired reinforcement learning for robot navigation," *IEEE/ASME Trans. Mechatronics*, vol. 17, no. 1, pp. 86–97, Feb. 2012.
- [16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [17] Z. Wang, C. Chen, and D. Dong, "Lifelong incremental reinforcement learning with online Bayesian inference," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021, early access, Feb. 11, 2021, doi: 10.1109/TNNLS.2021.3055499.
- [18] Y. Pan, G. I. Boutselis, and E. A. Theodorou, "Efficient reinforcement learning via probabilistic trajectory optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5459–5474, Nov. 2018.
- [19] G. W. Bassel, E. Glaab, J. Marquez, M. J. Holdsworth, and J. Bacardit, "Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets," *Plant Cell*, vol. 23, no. 9, pp. 3101–3116, 2011.
- [20] R. J. Urbanowicz and J. H. Moore, "Learning classifier systems: A complete introduction, review, and roadmap," *J. Artif. Evol. Appl.*, vol. 2009, no. 1, pp. 1–25, 2009.
- [21] P. Muñoz, R. Barco, J. M. Ruiz-Avilés, I. De la Bandera, and A. Aguilar, "Fuzzy rule-based reinforcement learning for load balancing techniques in enterprise LTE femtocells," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1962–1973, Jun. 2013.
- [22] K. Althoefer, B. Kregelberg, D. Husmeier, and L. Seneviratne, "Reinforcement learning in a rule-based navigator for robotic manipulators," *Neurocomputing*, vol. 37, no. 1–4, pp. 51–70, 2001.
- [23] Z. Wang, H. Li, and C. Chen, "Incremental reinforcement learning in continuous spaces via policy relaxation and importance weighting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 1870–1883, Jun. 2020.
- [24] D. H. Rothman and S. Zaleski, *Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics*, vol. 5. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [25] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.
- [26] R. Bellman, *Dynamic Programming*. Chelmsford, MA, USA: Courier Corporation, 2013.
- [27] S. Preitl *et al.*, "Iterative feedback and learning control. servo systems applications," *Int. Fed. Autom. Control*, vol. 40, no. 8, pp. 16–27, 2007.
- [28] G. Carbone, E. Villegas, and M. Ceccarelli, "Design and validation of force control loops for a parallel manipulator," in *Proc. Adv. Eng. Comput. Meth. Intell. Mech. Rob.*, 2013, pp. 206–224.
- [29] M. U. Ahmed *et al.*, "A machine learning approach to classify pedestrians' events based on IMU and GPS," *Int. J. Artif. Intell.*, vol. 17, no. 2, pp. 154–167, 2019.
- [30] I. Osband, B. V. Roy, and Z. Wen, "Generalization and exploration via randomized value functions," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2377–2386.
- [31] R. Houthoofd *et al.*, "VIME: Variational information maximizing exploration," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2016, pp. 1109–1117.
- [32] M. Bellemare *et al.*, "Unifying count-based exploration and intrinsic motivation," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2016, pp. 1471–1479.
- [33] J. Li, X. Shi, J. Li, X. Zhang, and J. Wang, "Random curiosity-driven exploration in deep reinforcement learning," *Neurocomputing*, vol. 418, pp. 139–147, 2020.
- [34] R. Y. Chen, S. Sidor, P. Abbeel, and J. Schulman, "UCB exploration via Q-ensembles," 2017, *arXiv:1706.01502*.
- [35] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, "Count-based exploration with neural density models," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 2721–2730.
- [36] H. Tang *et al.*, "# exploration: A study of count-based exploration for deep reinforcement learning," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2017, pp. 2753–2762.
- [37] A. Laversanne-Finot, A. Péré, and P.-Y. Oudeyer, "Curiosity driven exploration of learned disentangled goal spaces," in *Proc. Conf. Robot Learn.*, 2018, pp. 487–504.
- [38] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner, "COBRA: Data-efficient model-based RL through unsupervised object discovery and curiosity-driven exploration," 2019, *arXiv:1905.09275*.
- [39] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Proces. Syst.*, 2018, pp. 4863–4873.
- [40] K. Saito, A. Notsu, and K. Honda, "Discounted UCB1-tuned for Q-learning," in *Proc. Joint Int. Conf. Soft Comput. Intell. Syst. Int. Symp. Adv. Intell. Syst.*, 2014, pp. 966–970.
- [41] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proc. Conf. Learn. Theory*, vol. 125, pp. 2137–2143, 2020.
- [42] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2D lidar SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 1271–1278.
- [43] C. P. Birch, S. P. Oom, and J. A. Beecham, "Rectangular and hexagonal grids used for observation, experiment and simulation in ecology," *Ecological Model.*, vol. 206, no. 3/4, pp. 347–359, 2007.
- [44] E. Hoogeboom, J. W. T. Peters, T. S. Cohen, and M. Welling, "Hexaconv," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [45] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Proc. Conf. Robot. Learn.*, 2020, pp. 394–406.

- [46] H. Abelson and A. A. DiSessa, *Turtle Geometry: The Computer as a Medium for Exploring Mathematics*. Cambridge, MA, USA: MIT Press, 1986.
- [47] F. Duchoň *et al.*, “Path planning with modified a star algorithm for a mobile robot,” *Procedia Eng.*, vol. 96, pp. 59–69, 2014.
- [48] I. Châari, A. Koubaa, H. Bennaceur, S. Triguí, and K. Al-Shalfan, “Smart-path: A hybrid ACO-GA algorithm for robot path planning,” in *Proc. IEEE Congr. Evol. Comput.*, 2012, pp. 1–8.
- [49] A. Ravankar, A. A. Ravankar, Y. Kobayashi, Y. Hoshino, and C.-C. Peng, “Path smoothing techniques in robot navigation: State-of-the-art, current and future challenges,” *Sensors*, vol. 18, no. 9, 2018, Art no. 3170.



Chunlin Chen (Senior Member, IEEE) received the B.E. degree in automatic control and the Ph.D. degree in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is currently a Professor and the Chair of the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. His current research interests include machine learning, intelligent control, and quantum

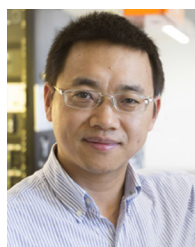
control.

Dr. Chen is the Chair of Technical Committee on Quantum Cybernetics, IEEE Systems, Man, and Cybernetics Society.



Yuanyang Zhu (Student Member, IEEE) received the B.E. degree in automation from the Department of Automation, Huaiyin Institute of Technology, Huai'an, China, in 2017, and the M.S. degree in control engineering, in 2020, from the Department of Control and Systems Engineering, School of Management and Engineering, Nanjing University, Nanjing, China, where he is currently working toward the Ph.D. degree in management science and engineering.

His current research interests include reinforcement learning, machine learning, and robotics.



Daoyi Dong (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively, both in engineering.

He is currently a Scientia Associate Professor with the University of New South Wales, Canberra, Australia. His research interests include machine learning and quantum cybernetics.

Dr. Dong was the recipient of an ACA Temasek Young Educator Award by the Asian Control Association and is a recipient of an International Collaboration Award, Discovery International Award and an Australian Postdoctoral Fellowship from the Australian Research Council, and Humboldt Research Fellowship from Alexander von Humboldt Foundation in Germany. He is an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and a Technical Editor for the IEEE/ASME TRANSACTIONS ON MECHATRONICS.



Zhi Wang (Member, IEEE) received the B.E. degree in automation from Nanjing University, Nanjing, China, in 2015, and the Ph.D. degree in machine learning from the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, in 2019.

He is currently an Assistant Professor with the Department of Control and Systems Engineering, Nanjing University. His current research interests include reinforcement learning, machine

learning, and robotics.