arXiv:2305.00561v1 [cs.AI] 30 Apr 2023

复杂任务下部分可观测环境中自主智能体的无模型运动规划

李俊超[1*†]，蔡明宇[2†]，阚震[3]，肖少平[4†]1*爱荷华大学爱荷华技术研究所机械工程系，美国爱荷华州爱荷华市西曼斯中心 3131 号，邮编 52242。[2]理海大学机械工程系，美国宾夕法尼亚州伯利恒市西纪念大道 19 号，邮编 18015。[3] 中国科学技术大学自动化系，中国安徽省合肥市蜀山区黄山路 443 号，邮编 230022。[4]爱荷华大学爱荷华技术研究所机械工程系，美国爱荷华州爱荷华市西曼斯中心 3131 号，邮编 52242。

通讯作者。电子邮箱：junchao-li@uiowa.edu；贡献作者：mic221@lehigh.edu；zkan@ustc.edu.cn；shaoping-xiao@uiowa.edu；

†这些作者对本研究的贡献相同。

摘要

在部分已知且信息不完整的环境中，自主智能体的运动规划是一项具有挑战性的问题，尤其是涉及复杂任务时。本文提出一种无模型强化学习方法来解决这一问题。我们将运动规划表述为概率标记部分可观测马尔可夫决策过程（PL-POMDP）问题，并使用线性时序逻辑（LTL）表达复杂任务。随后，将线性时序逻辑公式转换为极限确定性广义B¨uchi自动机（LDGBA）。基于模型检测技术，我们将问题重新定义为：在PL-POMDP与LDGBA的乘积上寻找最优策略，以满足复杂任务。我们采用结合长短期记忆网络（LSTM）的深度Q学习来处理观测历史和任务识别。我们的贡献包括所提出的

方法、LTL和LDGBA的应用，以及LSTM增强型深度Q学习。我们通过在多种环境中进行仿真，展示了所提方法的适用性，这些环境包括网格世界、虚拟办公室和多智能体仓库。仿真结果表明，所提方法能够有效应对环境、动作和观测的不确定性。这表明该方法在现实应用中具有潜力，包括无人机的控制。

关键词： 运动规划、部分可观测环境、复杂任务、线性时序逻辑、强化学习、循环神经网络

1 引言

部分可观测马尔可夫决策过程 [1] 提供了一个用于建模决策问题的数学框架，其中包括自主智能体的运动规划，例如无人机。它不同于 [2] 广泛应用于机器人学和自主系统的马尔可夫决策过程，后者假设环境完全可观测。对于现实世界的应用而言，部分可观测马尔可夫决策过程更加贴近实际，因为智能体（例如机器人）可能无法通过感知获取足够的信息，因而无法完全确定环境状态。另一方面，传统寻路技术已被广泛用于简单的运动规划任务（如前往目标点） [3] ，强化学习方法也同样如此 [4]。然而，监视任务等复杂任务与现实世界的应用更为相关。因此，智能体如何学会在部分可观测环境中规划运动以完成复杂任务，是一项具有挑战性的工作，尤其是在考虑环境和动作的不确定性时。

过去十年来，研究人员尝试使用各种基于模型的强化学习（RL）算法来解决部分可观测马尔可夫决策过程（POMDP）问题（任务为简单的到达目标任务）。许多现代求解器能够处理包含数千个状态的大型空间域 [5]。一种常用方法是基于点的值迭代（PBVI） [6,7] 方法，它由基于模型的算法构成，通过在信念空间的有限子集上计算价值函数，近似求解部分可观测马尔可夫决策过程问题。需要指出的是，信念状态表示智能体可能所在状态的概率分布。每次状态转移后，都需要通过贝叶斯方法，利用转移概率和观测概率更新信念状态 [8]。事实上，这类基于模型的方法将部分可观测马尔可夫决策过程问题转化为具有相应信念状态空间的等价马尔可夫决策过程问题。

解决部分可观测马尔可夫决策过程问题的另一种方法是无模型强化学习方法，在这种方法中，智能体不知道状态转移概率和观测概率。因此，策略将一系列观测（即观测历史）映射为所选动作。Mnih等人.[9, 10] 首次提出了深度Q学习（即深度Q网络），并在多款 Atari 2600 游戏上进行了测试，其中 Q 网络经过训练，能够达到人类水平的表现。具体而言，他们的 Q 网络将最近四帧（灰度图像）作为输入，并使用卷积神经网络（CNN） [11] 提取图像

特征，供全连接神经网络用于预测 Q值（即状态-动作值）。然而，由于他们的工作将环境建模为马尔可夫决策过程，这种方法绕开了部分可观测性问题，而最近四帧使智能体能够获取有限的过去经验。

基于上述研究，Hausknecht 和 Stone [12]提出在 Q 网络架构中加入循环神经网络（RNN）。他们将 Atari 2600 游戏建模为 POMDP 问题，并提出了一种所谓的深度循环 Q 网络（DRQN）：用长短期记忆网络（LSTM）替换 Q 网络的第一个卷积后全连接层 [13], ，该网络在每个时间步都将单张图像作为输入特征。因此，这种方法能够整合完整的观测历史，而不必使用固定长度的观测序列（例如， [9]中将四帧图像作为输入）。此外，Foerster 等人。[14] 通过提出深度分布式循环 Q 网络（DDRQN），将 DRQN 扩展到部分可观测环境中的多智能体强化学习问题，其中将上一个动作作为 Q 网络的输入。Zhu 等人。[15] 开发了一种名为动作特定深度循环 Q 网络（ADRQN）的新型网络架构，其中 LSTM 层处理动作历史及其对应的观测，以计算 Q 值。其他一些类似研究，包括 [16, 17], 在具有连续状态空间的 POMDP 问题中，为控制策略实现了循环神经网络（RNN）。

需要指出的是，上述关于部分可观测马尔可夫决策过程问题的研究仅考虑了简单的到达目标任务。然而，通过线性时序逻辑（线性时序逻辑）等形式语言 [18],，复杂任务已被纳入马尔可夫决策过程问题。通常，用户定义的高级规范可以表示为线性时序逻辑公式，随后将其转换为针对无限字的、带有 B¨uchi 或 Rabin 接受条件的 ω-自动机 [19]。因此，可以通过对马尔可夫决策过程与自动机的乘积进行控制综合，来解决机器人运动规划问题。最近，这种形式化方法被用于验证任务目标，以解决具有特定时序逻辑约束的部分可观测马尔可夫决策过程问题。Chatterjee 等人 [20] 研究了无限时域部分可观测马尔可夫决策过程中定性模型检测的不可判定性。他们的方法依赖于探索整个信念空间，最适用于状态空间较小的问题。他们还指出，该方法可能无法获得最优策略，从而确保部分可观测马尔可夫决策过程中给定逻辑公式的最大满足概率。

其他研究如 [21–23]提出，将线性时序逻辑规范转换为确定性 Rabin 自动机（确定性 Rabin 自动机），再构建部分可观测马尔可夫决策过程与确定性 Rabin 自动机的乘积，以解决此类问题。具体而言，Sharan 等 [21] 和 Ahmadi 等 [23] 采用有限状态控制器来限制策略搜索，并通过价值迭代方法进行求解。此外，Bouton 等 [22] 使用近似部分可观测马尔可夫决策过程求解器 SARSOP [7], ，在乘积部分可观测马尔可夫决策过程的有限信念状态空间上搜索最优策略。然而，这些方法都是基于模型的强化学习方法，要求智能体知道状态转移概率和观测概率。这一要求限制了它们在未知环境中的应用。需要指出的是，最相关的工作由 Carr 等 [24, 25]开展，他们将循环神经网络和模型检测用于满足时序逻辑约束的部分可观测马尔可夫决策过程问题。他们提出的方法

侧重于构建一种基于策略的过程，通过实现循环神经网络来迭代改进当前策略。然而，他们的方法构建了一个底层马尔可夫决策过程来映射原始的部分可观测马尔可夫决策过程，并使用基于模型的强化学习求解器 SARSOP[7], 来近似价值函数。因此，正如他们所述，这仍然是一种基于模型的方法。

另一方面， [26] 已表明，在解决 MDP 问题时，无论采用无模型还是基于模型的强化学习，极限确定性 B¨uchi 自动机（LDBA）都比确定性 Rabin 自动机更具优势。Hasanbeig 等人 [27] 指出，将线性时序逻辑规范转换为 LDBA，所得自动机的状态空间可能比确定性 Rabin 自动机更小。此外，与 LDBA 不同，极限确定性广义 B¨uchi 自动机（LDGBA）包含多个接受集，并使智能体能够无限次遍历所有接受集。此外，转换为 LDGBA 可以避免 LDBA 导致的奖励稀疏问题，而该问题可能会减缓强化学习的收敛速度 [28]。

本文将智能体与其部分可观测周围环境之间的交互建模为概率标记部分可观测马尔可夫决策过程（PL-POMDP）。在 POMDP 中引入概率标签，使我们能够同时考虑环境中的静态和动态事件。本研究的贡献之一是将线性时序逻辑公式转换为 LDGBA，以表示所研究 POMDP 问题中复杂任务的线性时序逻辑规范。据作者所知，此前的研究尚未报告过这一做法。生成 PL-POMDP 与 LDGBA 的乘积后，原问题——即在 PL-POMDP 中寻找满足线性时序逻辑规范的策略——可以改写为在相应的乘积部分可观测马尔可夫决策过程中寻找使累计奖励最大化的最优策略。另一项贡献是提出一种无模型强化学习方法，用于在乘积部分可观测马尔可夫决策过程中学习最优策略。我们在 Q 网络架构中引入循环神经网络，以处理智能体获取的信息：观测历史和任务识别。后者取决于智能体是否能够完全识别线性时序逻辑诱导的自动机。具体而言，分别使用自动机状态历史或状态标签历史，

本文结构安排如下：第2节回顾 PL-POMDP 的定义，并介绍如何使用带有循环神经网络的深度Q学习来解决一个简单的目标导航 POMDP 问题。第3节介绍线性时序逻辑、LDGBA 和乘积PL-POMDP。随后，本文重新表述带有线性时序逻辑规范的 PL-PO MDP 问题。第4节提出用于解决乘积PL-POMDP 问题的无模型方法，并给出详细算法。最后，第5节介绍实验及结果，并在此基础上展开讨论和总结，包括未来的研究工作。

2 背景

本节首先定义概率标记部分可观测马尔可夫决策过程（PL-POMDP），然后说明如何使用深度Q网络这一无模型强化学习方法，来解决具有简单目标到达任务的PO MDP问题。最后，我们通过一个仿真示例进行演示。本文重点是扩展深度Q网络，以解决具有复杂任务的POMDP问题；所提出的方法将在第 4 节中详细介绍。

2.1 PL-POMDP

POMDP通常用于对智能体无法充分观测并完全识别其周围环境的问题进行数学描述。我们采用PL-POMDP，并考虑静态事件和动态事件。

定义 1 (PL-POMDP)。 考虑到转移和观测的不确定性以及状态的概率标签，我们可以用元组 $\mathcal{P} = (S, A, T, s_0, R, O, \Omega, \Pi, L, P_L)$表示 *PL-POMDP*，其由以下部分组成：

- 有限状态集， $S = \{s_1, ..., s_n\}$。
- 有限动作集， $A = \{a_1, ..., a_m\}$。特别地， $A(s)$ 是智能体在当前状态 s下可用的动作集。

- 转移概率函数， T： $S \times A \times S \to [0, 1]$ 当智能体执行动作 $a \in A(s)$后，从当前状态 $s \in S$ 转移到下一个状态 $s' \in S$ 时。存在 $\sum_{s' \in S} T(s, a, s') = 1$。

- 一个初始状态， $s_0 \in S$。
- 一个奖励函数， R： $S \times A \times S \to R$。需要指出的是，奖励函数有时也可以定义为 $R(s')$、 $R(s, a)$ 或 $R(a, s')$。
- 一个有限的观测集合， $O = \{o_1, ..., o_k\}$。在当前状态 s， $O(s)$ 由智能体可以感知的观测组成。

- 一个观测概率函数，Ω： $S \times A \times O \to [0, 1]$，表示智能体在状态 $s' \in S$ 执行动作 $a \in A(s)$后能够感知到观测 o 的概率。该函数遵循 $\sum_{o \in O(s')} \Omega(s', a, o) = 1$。- 一个原子命题集合， Π。
- 一个标记函数， L： $S \to 2^{\Pi}$，输出状态 s下所有可能标签组成的集合。 2^{Π} 是 Π的幂集。

- 标记概率函数 $P_L(s, l)$表示与状态 $s \in S$ 相关联的标签 $l \in L(s)$的概率。它满足 $\sum_{l \in L(s)} P_L(s, l) = 1, \forall s \in S$。

智能体与其环境交互时，在选择并执行动作 $a \in A(s)$ 后，它以转移概率$T(s, a, s')$从状态 $s \in S$转移到状态 $s' \in S$。到达下一个状态 s'后，智能体以观测概率 $\Omega(s', a, o)$感知观测 $o \in O(s')$。此外，智能体根据奖励函数 $R(s, a, s')$，从环境中获得奖励作为反馈。

状态的标签是原子命题，表示特定状态下的事件发生。此外，标记概率函数可以描述状态 l 下的静态事件 s ，如果 $L(s) = \{l\}$ 且 $P_L(s, l) = 1$ ；否则描述动态事件。标签表示目标状态，用于处理简单的到达目标任务。在考虑复杂任务时，我们可以使用形式语言，根据标签对任务进行形式化，然后将该公式转换为有限状态自动机。标签也是导出自动机的输入符号。因此，在运动规划过程中，可以通过模型检测验证复杂任务是否完成。我们将在第 3 节提供此类方法的详细信息。

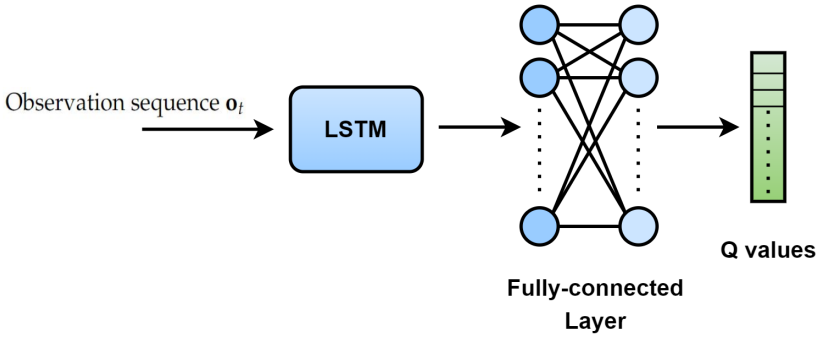


图 1 Q网络架构。

2.2 用于部分可观测马尔可夫决策过程问题的深度Q网络

为了解决转移概率未知的部分可观测马尔可夫决策过程问题，采用深度Q网络将观测序列映射为用于动作选择的Q值。我们考虑截至 j 个先前时间步的观测历史，因此观测序列表示为$\mathbf{o}_t = (o_{t-j}, o_{t-j+1}, o_{t-j+2}, ..., o_t)$，其长度为 $j+1$。因此，策略是观测序列的函数。智能体在决定动作方案时的目标是最大化如下所示的期望回报，该回报表示在策略 ξ下，智能体从当前时刻起能够获得的折扣奖励总和。

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

其中， s_t 是智能体在时间 t的状态， $\gamma \in [0,1]$ 是用于平衡即时奖励与未来奖励重要性的折扣因子。随后，最优策略可表示为 $\boldsymbol{\xi}^{*}(\mathbf{o}_t) = \text{argmax}_{\boldsymbol{\xi}} U^{\boldsymbol{\xi}}(s_0)$。

我们在 Q网络架构中采用长短期记忆网络处理观测序列，如图所示 1。长短期记忆网络是一种循环神经网络，可通过在网络架构中引入反馈回路，对观测之间的时间依赖关系进行建模。在每个时间步处理完观测后，输出会作为下一时间步的输入（与后续观测一起）反馈至网络。这样，网络便能捕捉序列中观测的顺序和时间信息。一维卷积神经网络和深度神经网络也可用于处理时间序列数据。然而，它们不会显式地对观测之间的时间依赖关系进行建模，而且深度神经网络无法处理长度可变的输入序列。在下一小节中，我们将用一个简单的目标导向仿真实例，展示循环神经网络在 Q 网络中相较于卷积神经网络和深度神经网络的优势。

深度Q网络通常有两个 Q 网络：评估Q网络 $Q_E(\mathbf{o}_t, a_t; \theta_E)$ 和目标Q网络 $Q_T(\mathbf{o}_t, a_t; \theta_T)$，其中 θ_E和 θ_T分别是网络权重。评估Q网络 Q_E 通常在每次迭代中通过随机

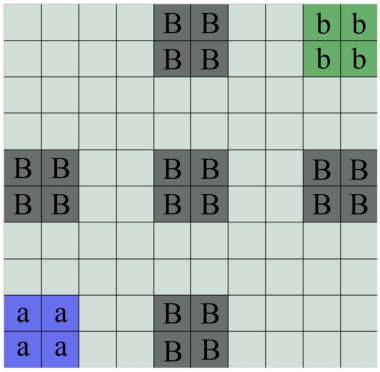


图 2 一个 10 × 10 的网格世界

从所谓的经验回放记忆库中选择一个数据样本批次 [29] ，在学习过程中进行。与此同时，目标Q网络 Q_T 保持权重不变，直到从评估Q网络复制权重 Q_E ，即偶尔进行一次， $\theta_T = \theta_E$。

在学习过程中，每一步（例如，时间 t），目标Q网络（Q_T）都会根据观测序列 $\mathbf{o}_t$ 预测 Q 值，以便智能体通过 ϵ-贪心技术 [4]选择动作 a_t 。执行所选动作并感知到观测 o_{t+1}后，会生成新的观测序列 $\mathbf{o}_{t+1}$ =(o_{t-j+1}， o_{t-j+2}， o_{t-j+3}，..... o_{t+1})。因此，形成一条经验： e_t =($\mathbf{o}_t$， a_t， r_t， $\mathbf{o}^{t+1}$)，其中 $r_t = R(s_t, a_t, s_{t+1})$。这条经验会作为一个数据样本记录在经验回放记忆库 D中；从中随机选取数据样本，以更新新的 Q 值并进行 Q 网络更新。用于更新与 a_t 相关联的 Q 值的方程如下所示。

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

其中， α 是学习率。

2.3 到达目标示例

我们以网格世界中的一个简单到达目标示例来演示 DQN，即一种无模型强化学习方法，用于解决部分可观测马尔可夫决策过程问题。图 2 展示了一个 10×10 网格世界，其中标记为“a”的蓝色区域和标记为“b”的绿色区域分别是初始状态和目标状态。障碍状态，即障碍物，标记为“B”。智能体（例如，网格世界中的移动机器人）必须从初始状态移动到目标状态。在每个状态下，它可以采取四种动作： up、 $left$、 $down$ 和 $right$。由于存在动作不确定性，智能体以 0.9 的概率采取所选动作；否则，它会以相同的概率采取两个侧向动作。此外，如果下一状态位于网格世界之外或是障碍物，智能体将停留在当前状态。到达下一状态后，智能体以 0.9 的概率观测到该状态，并以总计 0.1 的均匀分布概率观测到相邻状态。折扣因子 γ 设为 0.98。

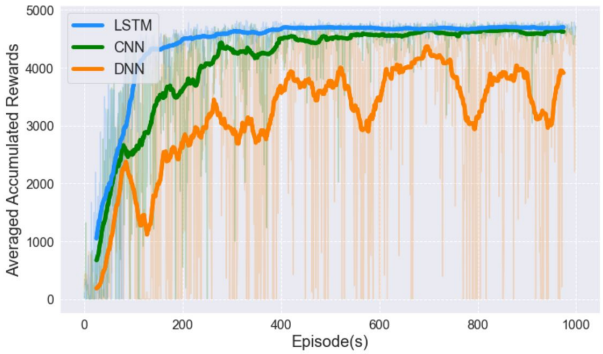


图 3 使用带有 LSTM、CNN 和 DNN 的 Q 网络所获得的累积奖励比较。

需要注意的是，在这个网格世界示例中，观测就是状态。因此，由于状态的行索引和列索引属于序数数据，观测序列需先通过独热编码进行预处理。随后，将生成的一维向量作为输入馈入 LSTM 层。提取出的特征接着传入两个全连接层，每层分别包含 16 个神经元，用于预测相应的 Q值。Q 网络采用修正线性单元（ReLU）作为激活函数。该问题的学习过程包含 1,000 个回合，每回合 500 步。两个 Q 网络——评估网络 Q_E 和目标网络Q_T——均采用随机初始化。评估网络的训练批次大小为 32；每隔 50 步，通过复制 Q_E 的权重系数来更新目标网络。

我们还通过实现 CNN 和深度神经网络（DNN）测试了另外两种 Q 网络架构。基于 CNN 的 Q 网络在全连接层之前，先使用一个二维卷积层（滤波器=6，内核大小=（3,3），步幅=1），随后连接另一个二维卷积层（滤波器=12，内核大小=（2,2），步幅=1）。DNN 仅使用全连接层。图 3 比较了智能体每十个回合收集的平均累积奖励（趋势线表示每 50 个回合计算一次奖励的简单移动平均（SMA））。结果表明，带有 LSTM 的 Q 网络能够实现更快的收敛速度和更高的累积奖励。

收敛后，可以从 Q 网络中派生出最优策略。图 4展示了智能体在网格世界中移动以完成这一简单的趋向目标任务时，根据派生策略生成的一条路径。初始状态标记为较大的紫色实心圆，到达的状态标记为浅红色圆点。颜色更亮的红点以及黑色路线的弯折表明该状态曾被访问多次。需要注意的是，即使策略已经收敛并趋近于最优，生成路径也可能并非最优。这是由于观测与动作不确定性造成的。

3 问题定义

在本研究中，我们提出了一种框架，通过将线性时序逻辑公式转换为 LDGBA，以表示 POMDP 的任务变量和安全约束，并生成乘积

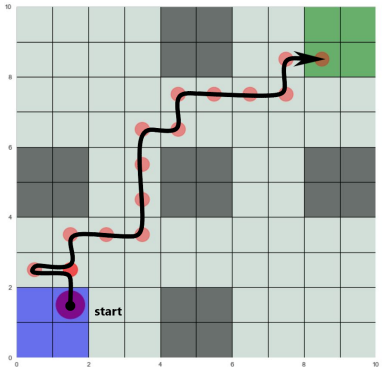


图 4 由派生策略生成的一条路径。

POMDP 与 LDGBA，从而求解带有线性时序逻辑规范的 PL-POMDP 问题。在 POMDP 中满足给定 LTL 约束的问题，等价于到达乘积 POMDP 中的（B¨uchi）接受状态。

3.1 线性时序逻辑（LTL）

线性时序逻辑 [18]是一种用于描述线性时间性质的逻辑形式体系，表示顺序执行中状态标签之间的关系。除了布尔连接词之外，LTL 还通过引入一些时序运算符扩展了命题逻辑，其中包括两个基本运算符 $\bigcirc$ （读作“下一时刻”）和 $\mathcal{U}$ （读作“直到”）。本研究假设 $a \in \Pi$ 是一个原子命题，且 ϕ，ϕ_1 和 ϕ_2 都是单个线性时序逻辑公式。于是，线性时序逻辑公式可以按照以下文法构成[30]:

$$\phi := \text{True} \mid a \mid \phi_1 \wedge \phi_2 \mid \neg\phi \mid \bigcirc\phi \mid \phi_1 \mathcal{U} \phi_2 \tag{3}$$

其中，否定（$\neg$）和合取（$\wedge$）是布尔运算符。公式 $\bigcirc\phi$ 在当前时刻为真，当且仅当 ϕ 在下一时刻为真。此外，公式 $\phi_1 \mathcal{U} \phi_2$ 在当前时刻为真，当且仅当 ϕ_2 在某个未来时刻为真，并且 ϕ_1 在该未来时刻之前的所有时刻都为真。

其他常用的时序算子有 $\Diamond$ （读作“最终”）和$\Box$ （读作“始终”）。公式 $\Diamond\phi$ 确保 ϕ 在未来某个时刻为真，而 $\Box\phi$ 从现在起将永远为真。它们可以推导如下:

$$\begin{aligned} &\text{eventually}: \quad \Diamond\phi \equiv \text{True}\, \mathcal{U} \phi \\ &\text{always}: \quad\;\; \Box\phi \equiv \neg(\Diamond\neg\phi) \end{aligned} \tag{4}$$

使用 $\models$ 表示满足关系时，我们可以将词上的线性时序逻辑公式语义解释如下。词是一个无限序列 $\boldsymbol{w} = w_0 w_1 \ldots$

其中 Π 是对所有 $i \geq 0$而言的原子命题集合。

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

3.2 限定确定性广义 B¨uchi 自动机（LDGBA）

上一小节介绍了，用户指定的复杂任务可以通过线性时序逻辑进行表述。然后，我们可以将线性时序逻辑公式转换为自动机，包括 LDGBA[31],以通过模型检测评估任务满足情况 [18]。

定义 2（LDGBA）：*LDGBA* $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$ 由有限状态集合 Q、有限字母表（即有限输入符号集合） $\Sigma = 2^{\Pi}$ （其中 Π是原子命题集合）、转移函数 δ: $Q \times (\Sigma \cup \{\epsilon\}) \rightarrow 2^Q$、初始状态 $q_0 \in Q$以及接受集集合 $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_f\}$ 组成，其中 $\mathcal{F}_i \subseteq Q$、$\forall i \in \{1, \ldots, f\}$。此外，状态集合 Q 可分解为确定性集合和非确定性集合，即分别为 Q_D 和 Q_N。它们满足以下要求。

- $Q_D \cup Q_N = Q$ 和 $Q_D \cap Q_N = \emptyset$。
- Q_D 中的状态转移是完全的，即 $|\delta(q, \alpha)| = 1$。
- Q_D 中的转移仅限于其中，即对于每个状态 $q \in Q_D$ 和 $\alpha \in \Sigma$，都有 $\delta(q, \alpha) \subseteq Q_D$。
- ϵ-转移不读取输入符号，且仅在从 $q \in Q_N$ 到$q' \in Q_D$时有效。
- 接受集由接受状态组成，且仅在确定性集合中定义。换言之，对于每个 $\mathcal{F}_i \in \mathcal{F}$，都有 $\mathcal{F}_i \subseteq Q_D$。

给定输入词 $\boldsymbol{w} = w_0 w_1 \ldots$，LDGBA 的一次运行可根据转移函数 $\delta(q_i, w_i) = q_{i+1}$ 表示为$\boldsymbol{q} = q_0 q_1 \ldots$。令 $\inf(\boldsymbol{q})$ 表示 $\boldsymbol{q}$ 的无限部分。从理论上讲，如果存在 $\inf(\boldsymbol{q}) \cap \mathcal{F}_i \neq \emptyset, \forall i \in \{1, \ldots f\}$，则 $\boldsymbol{q}$ 满足 LDGBA 接受条件，即 LDGBA 接受词 $\boldsymbol{w}$。有关自动机生成的更多详情，请读者参阅 Owl [19] 。本研究旨在解决下文定义的带有 LTL 约束的 POMDP 问题。

问题 1。 给定一个 *PL-POMDP* $\mathcal{P}$ 以及一个通过线性时序逻辑公式表达的复杂任务。目标是寻找一个策略 $\xi^*(\mathbf{o}_t)$，其中$\mathbf{o}_t$ 表示在 $\mathcal{P}$上的观测序列，使其能够通过满足线性时序逻辑诱导的 *LDGBA* 的接受条件来完成该任务。

3.3 乘积部分可观测马尔可夫决策过程

定义 3（乘积部分可观测马尔可夫决策过程）。给定一个 *PL-POMDP* $\mathcal{P} = (S, A, T, s_0, R, O、\Omega\ \Pi, L)$ 和一个 *LDGBA* $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$，乘积 *PL-POMDP*（简称乘积 *POMDP*）定义为$\mathcal{P}^\times = \mathcal{P} \times \mathcal{A} = (S^\times, A^\times, T^\times, s_0^\times, R^\times, O、\Omega^\times, \mathcal{F}^\times)$，由以下组件构成。

- 一个有限的带标签状态集合，$S^\times = S \times Q$ 或 $s^\times = \langle s, q\rangle \in S^\times$，其中 $s \in S$ 且$q \in Q$。
- 一个有限的动作集合，$A^\times = A \cup \{\epsilon\}$。
- 一个转移函数，$T^\times = S^\times \times A^\times \times S^\times \to [0,1]$，以及

$$T^\times\left(s^\times, a^\times, s^{\times\prime}\right) = \begin{cases} T(s, a^\times, s') & q' = \delta(q, l), l \in L(s')\ and\ a^\times \in A \\ 1 & a^\times \in \{\epsilon\}\ and\ q' \in \delta(q, \epsilon)\ and\ s' = s, \\ 0 & otherwise. \end{cases} \tag{6}$$

  其中 $s^{\times\prime} = \langle s', q'\rangle$。
- 一个初始状态 $s_0^\times = \langle s_0, q_0\rangle \in S^\times$，其中 $s_0 \in S$ 且 $q_0 \in Q$。
- 奖励函数 $R^\times = S^\times \times A^\times \times S^\times \to \mathcal{R}$，以及

$$R^\times(s^\times, a^\times, s^{\times\prime}) = \begin{cases} R(s, a^\times, s')\ a^\times \in A, l \in L(s'), q' = \delta(q, l) \in \mathcal{F}_i,\ and\ \mathcal{F}_i \in \mathcal{F} \\ 0 \qquad\qquad otherwise. \end{cases} \tag{7}$$

- 观测函数 $\Omega^\times = S^\times \times A^\times \times O \to [0,1]$ 是

$$\Omega^\times(s^{\times\prime}, a^\times, o) = \Omega(s', a^\times, o) \tag{8}$$

  如果 $a^\times \in A$。否则，如果 $a^\times \in \{\epsilon\}$，智能体将保持在同一状态 s，即 $s' = s$，但 $q' = \delta(q, \epsilon)$，且无法感知到任何观测。
- 一组接受集 $\mathcal{F}^\times = \left\{\mathcal{F}_1^\times, \mathcal{F}_2^\times, ..., \mathcal{F}_f^\times\right\}$，其中 $\mathcal{F}_i^\times = \{\langle s, q\rangle | s \in S;\ q \in \mathcal{F}_i\}$且 $i = 1, ...f$。

乘积部分可观测马尔可夫决策过程上的随机路径，表示为 $(s_0, q_0)(s_1, q_1)\ldots$，是部分可观测马尔可夫决策过程上的路径 $s_0 s_1 \ldots$与 LDGBA 上的路径 $q_0 q_1 \ldots$的组合。与 (1) 类似，在策略 $\xi^\times$下，从初始状态出发，在乘积部分可观测马尔可夫决策过程上的期望回报可写为

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

需要指出的是，乘积部分可观测马尔可夫决策过程 $\mathcal{P}^\times$ 可视为具有增广状态空间的 PL-POMDP $\mathcal{P}$，该增广状态空间包含自动机状态空间。因此，乘积部分可观测马尔可夫决策过程考虑了由 LDGBA$\mathcal{A}$表示的时序逻辑规范。在 $\mathcal{P}^\times$ 上的任何可行路径，都兼具原始 PL-POMDP $\mathcal{P}$ 上可访问路径与线性时序逻辑诱导的自动机 $\mathcal{A}$所接受词之间的交集。

例如，派生策略$\xi^{\times}$可在乘积部分可观测马尔可夫决策过程 $\mathcal{P}^{\times}$上生成路径 $\sigma^{\xi^{\times}} = (s_0, q_0)(s_1, q_1)$ .....。如果存在 $\inf(\sigma^{\xi^{\times}}) \cap \mathcal{F}_i^{\times} \neq \emptyset, \ \forall i = 1, \ ...f$，其中$\mathcal{F}^{\times}$ 刻画了 LDGBA $\mathcal{A}$的接受条件，则该路径被接受。换言之，运行 $q_0 q_1 \ldots$满足 LDGBA 的接受条件，或者说，LDGBA 接受相应的词。

因此，根据许多以往通过乘积马尔可夫决策过程求解带有线性时序逻辑规范的马尔可夫决策过程问题的研究， [30, 32–34], 乘积部分可观测马尔可夫决策过程上的最优策略 $\xi^{\times *}(\mathbf{o}_t, \mathbf{q}_t)$ $\mathcal{P}^{\times}$ 等价于 PL-POMDP 上的最优策略 $\xi^{*}(\mathbf{o}_t)$ $\mathcal{P}$ ，同时满足线性时序逻辑规范。$\mathbf{o}_t$ 表示观测历史，而$\mathbf{q}_t$表示相应的自动机状态转移历史。需要注意的是，这里假设智能体会将标签（即自动机中的输入符号）作为反馈的一部分接收。因此，如果智能体完全了解任务，即了解线性时序逻辑诱导的自动机，则可以推导出自动机状态转移历史。由此，我们可以将问题1 重新表述如下。

问题 2。乘积部分可观测马尔可夫决策过程 $\mathcal{P}^{\times} = \mathcal{P} \times \mathcal{A}$定义见第*3.3*节，由描述部分可观测环境的 *PL-POMDP* $\mathcal{P}$ 和表达复杂任务的线性时序逻辑规范的 *LDGBA*$\mathcal{A}$ 构造。目标是找到一个策略$\xi^{\times *}(\mathbf{o}_t, \mathbf{q}_t)$，其中$\mathbf{o}_t$ 和$\mathbf{q}_t$ 分别表示关于 $\mathcal{P}^{\times}$ *s*的观测序列和自动机状态转移序列，使期望回报（*9*）最大化。

另一方面，如果智能体不了解任务，就无法根据标签反馈推导自动机状态转移。因此，乘积部分可观测马尔可夫决策过程上的策略是观测历史和感知到的标签历史的函数，表示为 $\xi^{\times *}(\mathbf{o}_t, \mathbf{l}_t)$。随后，可以对上述问题表述进行相应修订。

4 方法论

在本研究中，我们提出在乘积部分可观测马尔可夫决策过程上采用无模型强化学习方法（即深度 Q 网络），以综合生成满足线性时序逻辑规范的部分可观测环境中智能体的最优运动规划。如第 2 节所讨论和展示的，循环神经网络具有嵌入 Q 网络以解决部分可观测马尔可夫决策过程问题的优势，因为它们能够捕获观测历史中的相对时间依赖关系。在所提出的方法中，我们扩展了循环神经网络增强的 Q 网络，以处理感知到的观测并识别复杂任务。

我们考虑两种情形，具体取决于智能体是否知晓指定任务。如果明确为智能体指定了任务，它就完全了解线性时序逻辑诱导的自动机，包括其转移函数。因此，一旦到达乘积部分可观测马尔可夫决策过程的状态，即增广状态，即使部分可观测马尔可夫决策过程的状态并非完全可观测，它也能推导出相应的自动机状态。需要注意的是，状态标签与观测有所区别。我们假设智能体能够通过环境反馈正确获取状态标签。因此，在这种情况下，除观测历史外，我们的方法还将已识别的自动机状态作为另一项输入特征提供给 Q 网络。另一方面，如果没有明确为智能体指定任务，且它不了解自动机的

转移函数，则将（部分可观测马尔可夫决策过程的）状态标签作为 Q 网络的附加输入特征。

4.1 将自动机状态序列作为额外输入

假设智能体完全了解任务。在这种情况下，自动机是完全已知的，可以根据自动机转移函数从已获取的标签推导出自动机状态（即 q 状态）。因此，除了观测历史外，还可以将自动机状态直接用作 Q 网络的输入。具体而言，我们使用长度为 k 的 q 状态序列$\mathbf{q}_t = \bar{q}_1 \dots \bar{q}_k$，其中 $\bar{q}$ 表示推导出的 q 状态，代表智能体在 PL-POMDP 上执行转移时自动机随时间步 t 的演化。需要注意的是，观测序列$\mathbf{o}_t = o_{t-j} \dots o_t$ 的长度为 $j+1$，并根据从时间步 $t-j$ 到当前时间 t的观测历史生成。然而，由于自动机空间通常小于部分可观测决策过程空间，且自动机状态并非每个时间步都会转移，因此通过记录每个时间步的自动机状态来生成自动机转移历史并不实际。相反，我们使用先进先出（FIFO）来跟踪自动机转移，每次转移只记录一次。因此，在到达下一个自动机状态之前，$\mathbf{q}_t$ 可能会连续多个步保持不变。此外，$\mathbf{o}_t$ 和$\mathbf{q}_t$ 的长度通常不同。

在每次转移 $a_t^\times \in A$ 后，观测到的 o_{t+1} 和导出的自动机状态q_{t+1} 用于生成新的观测序列 $\mathbf{o}_{t+1}$ 和 q 状态序列 $\mathbf{q}_{t+1}$。结合之前的序列，可以将新的经验表示为 $(\mathbf{o}_t, \mathbf{q}_t, a_t^\times, r_t^\times, \mathbf{o}_{t+1}, \mathbf{q}_{t+1})$，并将其作为一个数据样本记录在经验回放存储器中。如果没有发生新的自动机转移，则 $\mathbf{q}_{t+1} = \mathbf{q}_t$。需要注意的是，如果线性时序逻辑诱导的 LDGBA 中存在 ϵ-转移，那么根据第 3.3 节中的定义，该转移就是乘积部分可观测马尔可夫决策过程中的一个可用动作。在我们的方法中，执行 ϵ-转移后不会感知到新的观测，因为智能体仍处于相同的部分可观测马尔可夫决策过程状态。然而，自动机状态会发生变化，并被记录在相应的自动机状态序列中，以供下一个时间步使用。

图5 展示了我们无模型强化学习方法中 Q 网络的架构。观测序列 $\mathbf{o}_t$ 和自动机状态序列 $\mathbf{q}_t$ 在时间步t 分别经过独热编码预处理后，再输入长短期记忆网络。由于这两个序列长度不同，因此采用两个独立的长短期记忆网络提取隐藏状态，随后将其拼接并展平，供全连接层估计 Q 值。

与常规深度Q网络类似，我们在乘积部分可观测马尔可夫决策过程上的深度Q网络也包含两个相同的 Q 网络：评估网络 $Q_E^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_E^\times)$ 和目标网络$Q_T^\times(\mathbf{o}_t, \mathbf{q}_t, a_t^\times; \theta_T^\times)$，其中 $a^\times \in A$。目标网络用于下一步的动作选择和 Q值预测。另一方面，评估网络每经过 M 步，便从经验回放记忆库中随机选取一批数据样本进行训练。每经过 K 步，便通过复制评估网络的权重系数来更新目标网络。使用两个神经网络可以避免单个神经网络中深度Q网络的自举。

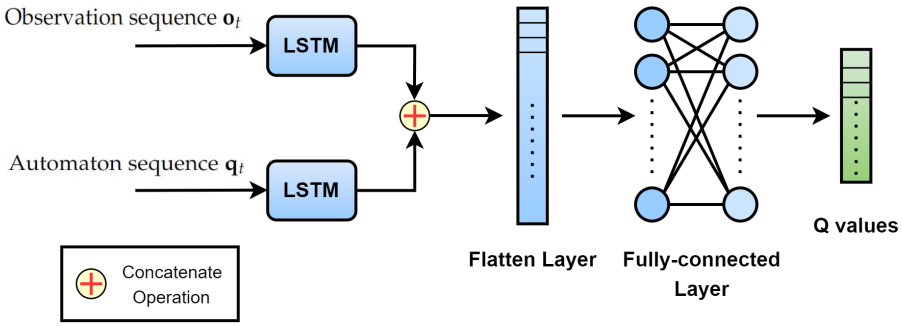


图 5 以$\mathbf{o}_t$ 和$\mathbf{q}_t$ 为输入的 Q 网络架构。

所选批次中每个数据样本的输出都是一个 Q值，可以按照下面的公式进行更新。

$$Q^{\times}_{\text{new}}(\mathbf{o}_t, \mathbf{q}_t, a^{\times}_t) = Q^{\times}_E(\mathbf{o}_t, \mathbf{q}_t, a^{\times}_t; \theta^{\times}_E) + \alpha \left[r^{\times}_t + \gamma \max_{a^{\times}_{t+1}} Q^{\times}_T(\mathbf{o}_{t+1}, \mathbf{q}_{t+1}, a^{\times}_{t+1}; \theta^{\times}_T) - Q^{\times}_E(\mathbf{o}_t, \mathbf{q}_t, a^{\times}_t; \theta^{\times}_E) \right] \tag{10}$$

其中，$r^{\times}_t = R^{\times}(s^{\times}_t, a^{\times}_t, s^{\times'}_t)$，$\alpha$ 为学习率，γ 为折扣因子。因此，每个数据样本都具有输入特征$\mathbf{o}_t$ 和$\mathbf{q}_t$，以及输出目标$Q^{\times}_{\text{new}}$，用于训练和更新评估 Q 网络 Q_E。此外，使用以下损失函数更新评估Q网络的权重系数。

$$\mathcal{L}(\theta^{\times}_E) = \mathbb{E}_{(\mathbf{o}_t, \mathbf{q}_t, a^{\times}_t, r^{\times}_t, \mathbf{o}_{t+1}, \mathbf{q}_{t+1}) \sim U(D)} \left[\left(r^{\times}_t + \gamma \max_{a^{\times}_{t+1}} Q^{\times}_T(\mathbf{o}_{t+1}, \mathbf{q}_{t+1}, a^{\times}_{t+1}; \theta^{\times}_T) - Q^{\times}_E(\mathbf{o}_t, \mathbf{q}_t, a^{\times}_t; \theta^{\times}_E) \right)^2 \right] \tag{11}$$

算法 1 展示了在智能体与环境交互过程中训练评估 Q 网络的过程。收敛后，Q 网络可以预测状态-动作值（即 Q值）函数，从而为具有时序逻辑规范的所研究部分可观测马尔可夫决策过程问题推导出最优策略。

4.2 将标签序列作为额外输入

如果智能体不了解任务，即线性时序逻辑诱导的自动机不可知，则需要将状态标签作为额外输入提供给 Q 网络。我们通过将图 5 中的 Q 网络架构里的自动机状态序列替换为标签序列（即输入符号序列），并保留观测序列，对其进行了修改。如上所述，状态标签不同于观测，智能体可以将其作为反馈精确地接收。因此，收集到的经验序列变为（$\mathbf{o}_t$，$\mathbf{l}_t$，$a^{\times}_t$，$r^{\times}_t$，$\mathbf{o}t{+}1$，$\mathbf{l}t{+}1$），其中 $\mathbf{l}_t$ 是截至时间步 t所接收到的标签序列，长度为 k。与使用 q 状态的情形类似，与部分可观测马尔可夫决策过程状态对应的标签可能是稀疏的。因此，我们采用相同的先进先出方法（如第 4.1节所述）生成标签序列，该序列仅

Alg算法 1深度 p 循环 Q-网络处理乘积部分可观测马尔可夫决策过程问题。

1：初始化线性时序逻辑公式 ϕ、部分可观测马尔可夫决策过程 $\mathcal{P}$。2：将 ϕ 转换为 LDGBA $\mathcal{A}$。3：构造乘积部分可观测马尔可夫决策过程 $\mathcal{P}^{\times} = \mathcal{P} \times \mathcal{A}$。4：初始化评估网络 $Q_E^{\times}$ 、目标网络 $Q_T^{\times}$ 、经验回放记忆库D、观测序列长度 $j+1$、长度为 k的空 q 状态序列 q_0 、学习率 α、折扣因子 γ、回合总数E、步数总数 N、批次大小 M以及$Q_T^{\times}$更新步数 K。5：当当前回合 e 属于 E 时循环。6：随机选择一个初始状态 $s_0^{\times}$。7：当当前步 i 属于 N 时循环。8：若满足 $i < j+1$，则随机选择一个动作 $a_i^{\times}$ ；否则，通过ϵ-贪心策略选择一个动作。9：获取观测 o_{t+1} 并导出自动机状态 q_{t+1}。10：生成 o_{t+1} 和 q_{t+1}。11：收集奖励 $r_i^{\times}$ 。12：将经验 $\langle \mathsf{o}_i$、q_i、 $a_i^{\times}$ 、 $r_i^{\times}$ 、o_{i+1}、$\mathsf{q}_{i+1}\rangle$ 存入 D。13：如果 $i > 0$ 且 $i\%M{=}0$ ，则 14：从经验回放记忆库中随机抽取 M 个数据样本，作为 U(D)。15：计算每个数据样本的 $Q_{new}^{\times}$ 。16：使用该批次样本训练$Q_E^{\times}$ 。17：结束 if。18：如果 $i > 0$ 且 $i\%K{=}0$ ，则 19：将 $Q_E^{\times}$ 的权重传递给 $Q_T^{\times}$ 。20：结束 if。21：结束循环。22：结束循环。23：训练结束并保存评估网络 $Q_E^{\times}$

收到标签后将其存储。另一方面，算法1 可以进行相应的修改。

5 仿真与结果

我们在三个离散部分可观测马尔可夫决策过程领域的仿真中评估所提出的方法。首先，我们在一个部分可观测网格世界中进行仿真，研究两种不同的任务，并考虑自动机中的 ϵ-转移以及 PL-POMDP 中的静态/动态事件。随后，我们在办公室场景中测试这些方法，并采用两种不同的观测设置。最后，我们还将所提出的方法初步应用于多智能体强化学习案例。仿真使用 Python 3.9 和 Rabinizer 4 编写，并在配备 3.20 GHz 八核 CPU 和 32 GB RAM 的台式机上完成。部分源代码和补充材料见 [1]。

[1]https://github.com/JunchaoLi001/Model-free 深度强化学习 长短期记忆网络 在部分可观测马尔可夫决策过程中的 LDGBA

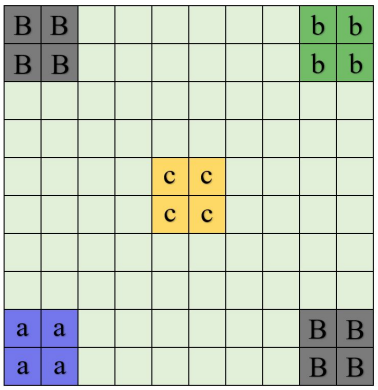


图 6 一个包含陷阱状态的 10 × 10 网格世界

5.1 网格世界仿真

我们使用一个 10×10 网格世界工作区，如图所示。6。若干状态标记为蓝色的'a'和绿色的'b'，分别表示两种不同的事件。标记为'c'的陷阱状态表示智能体一旦进入便永远无法离开。转移概率和观测概率等其他参数与第 2.3 节简单前往目标情形中的定义相同。本文研究两项任务。任务 1 展示了我们的方法能够处理带有 LDGBA 的 ϵ-转移。任务 2 考虑两种情形：静态事件和动态事件。假设事件为动态事件时，每个事件在其带标签状态发生的概率为 90%，在其他带标签状态发生的概率为 10%。两项任务均使用长度为 $j+1=5$ 的观测序列、长度为 $k=3$ 的自动机状态或标签序列、批次大小 $M=32$，并以每回合 600 步进行 15,000 回合仿真；评估 Q 网络的复制步数为 $K=50$。

5.1.1 任务 1

在网格世界中测试的第一个任务要求智能体无限次访问标记为"a"或"b"的状态。线性时序逻辑公式如下所示，其诱导的 LDGBA 如图 7 所示。

$$\phi_1 = (\square\Diamond a \mid \square\Diamond b) \wedge \square\neg c \tag{12}$$

可以看出，LDGBA 包含 ϵ-转移，这些转移被纳入生成的乘积部分可观测马尔可夫决策过程中的动作集合 $a^{\times}$。图 7 还显示，ϵ-转移可以使自动机状态从初始状态 q_0 转移到 q_1或 q_2，从而使智能体分别持续访问仅标记为"a"或"b"的状态。访问"c"将导致进入一个无出口的状态（q_3），即"陷阱"状态。

根据第 3.2 节中对 LDGBA 的定义 ϵ 转移不接受任何输入符号。它们仅用于进入确定性自动机状态集合，在该集合中转移受到限制。因此，在执行 ϵ 转移后，智能体将处于与 q_1 或 q_2 相关联的增广状态，以完成任务。在没有

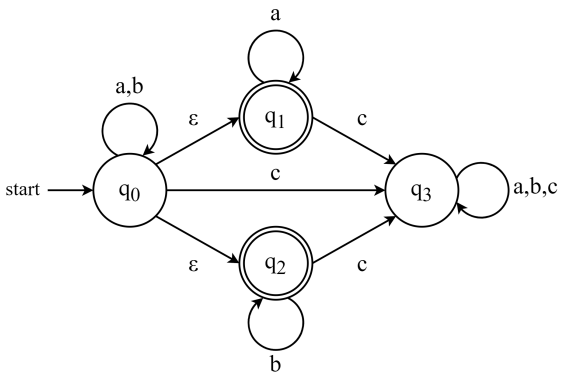


图 7 ϕ_1 的 LDGBA。

在不失一般性的情况下，我们为每个回合随机设置选择 ϵ 转移的概率，该转移仅发生一次。否则，其他动作，即 $a^{\times} \in A$，将基于 ϵ-贪心方法进行选择。如第3.3节所述，智能体在执行 ϵ 转移后不会立即感知观测，Q 网络仅预测除 ϵ 转移之外的动作的 Q值。

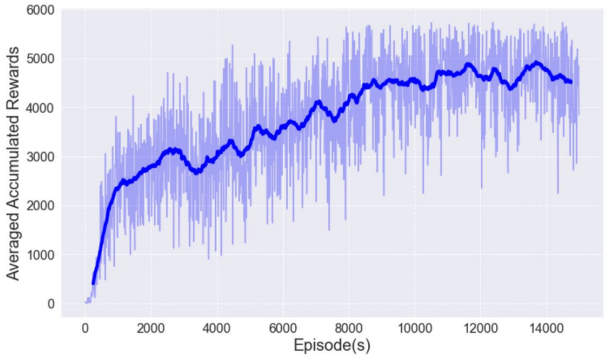
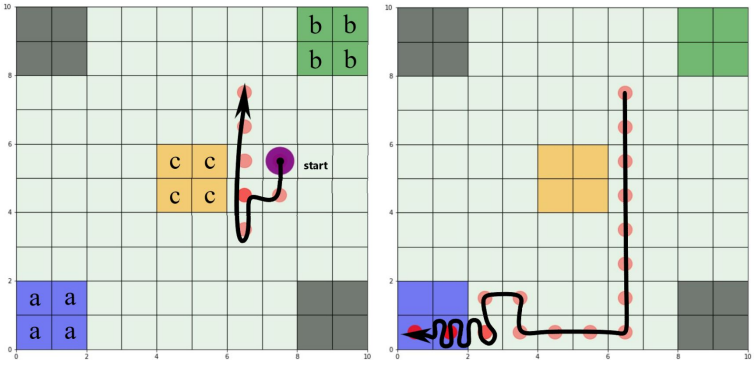


图 8 任务 ϕ_1 的平均累积奖励。

在任务 1 中，我们只考虑静态事件，即 $P_L(s_a, \text{`}a') = P_L(s_b, \text{`}b') = 1$。此外，假设智能体了解该任务。因此，将观测和 q 状态历史输入 Q 网络，以预测 Q值并进行动作选择。图 8展示了累积奖励的变化趋势，该值按每 10 个回合取平均，并采用 50 个回合的简单移动平均。由于智能体可能意外进入“陷阱”状态，平均奖励更能呈现奖励收敛的趋势。每当智能体访问接受状态时，奖励设为 10。获得最优策略后，我们为智能体生成一条完成任务的路径，如图 9所示。

起初，智能体随机选择动作，如图 9(a) 所示，然后生成第一个观测和 q 状态序列。随后可以预测 Q值，



(a) (b)

图 9 任务 ϕ_1的单轮路径：**(a)** q_0上的路径。**(b)** q_1上的路径。

除非发生 ϵ-转换，否则将采用贪心动作选择。发生 ϵ- 转换后，自动机状态从 q_0 转换为 q_1，智能体希望无限次访问标记为 'a' 的状态。如图 9(b) 所示，智能体先向下移动，再向左移动，以持续访问状态 'a'。由于动作存在不确定性，智能体偶尔会多次访问某些状态，这些状态在图 9 中以亮红色的点表示。

5.1.2 任务 2

第二项任务要求智能体按照先后顺序访问标记为'a'和'b'的状态，并无限次重复这一过程，同时还需应对标记不确定性导致的动态事件。线性时序逻辑公式表示为

$$\phi_2 = \square\diamondsuit(a \wedge \diamondsuit b) \wedge \square\neg c \tag{13}$$

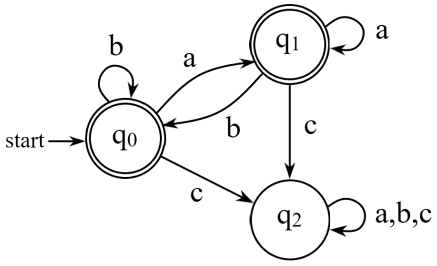


图 10 ϕ_2的 LDGBA。

在线性时序逻辑诱导的 LDGBA 中，如 [27, 28] 中所讨论的，直接利用接受集可能无法找到确定性策略。受他们工作的启发，我们修改了自动机结构和奖励函数，以简化训练过程。具体而言，我们扩展接受状态，以分离输入符号 'a' 或

分别为“b”。仅保留单标签转移作为示例，如图所示。10 针对任务 2。然后，我们重新设计奖励函数（7），如下所示，通过在奖励函数中添加约束，使智能体能够重复访问接受集。

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

其中，$q \neq q'$ 通过移除相关带标签的部分可观测马尔可夫决策过程状态上的奖励，避免在同一自动机接受状态处重复转移。将此约束应用于奖励函数后，得到的最优策略便满足所需的监视任务规范。另一种方法是实现 M. Cai 等人 [33]提出的前沿跟踪函数。

对于任务 2，我们研究智能体在包含动态事件或静态事件的环境中的学习，并结合两种取决于智能体对任务的认知程度的情形。首先，为了进行详细演示，我们考虑动态事件，并假设智能体完全了解任务。引入标记不确定性，使得某个事件在其带标签状态处发生的概率为 90%，而在其他带标签状态处发生的概率为 10%。例如，在标记为‘a’的状态处，事件‘a’发生的概率为 90%，而事件‘b’发生的概率为 10%。此外，除了观测序列之外，还将状态序列用作 Q 网络的输入。

图 11（1）展示了智能体在完全了解任务时，根据学习策略生成的一条单轮路径，以完成任务 2。在图11（a）中，智能体从初始状态（紫色圆点）出发，在 q_0 上进行五次随机移动，随后生成第一条用于决策的观测序列（以及相应的 q 状态序列）。之后，它绕过陷阱状态，前往左下方标记为‘a’的状态（蓝色方块）。然而，由于标记不确定性，智能体首次到达蓝色区域时，该状态并未发生事件‘a’。因此，智能体继续向下移动到下一个状态，并在下一个时间步发生事件‘a’。路径顶部黑色路线的弯折是由运动不确定性造成的。

访问标记为‘a’的状态后，自动机从 q_0 转移到q_1。在图 11（b）中，智能体绕过标记为‘a’的状态区域，然后避开黄色的陷阱状态，到达右上角标记为‘b’的状态。同样，智能体首次到达时，事件‘b’并未发生。随后，智能体再向右移动一步，事件‘b’随之发生。因此，智能体完成了单轮访问，先访问‘a’，再访问‘b’。与此同时，智能体回到了自动机状态q_0。图11（c）显示，智能体尝试返回标记为‘a’的状态以进行第二轮访问，但在前往‘a’状态之前，它连续几次访问了标记为‘b’的状态。

此外，我们还在智能体不了解任务的情况下，模拟了一个存在动态事件的环境。在这种情况下，观测序列和标签序列被输入 Q 网络。图 11 (2) 展示了根据学习策略生成的路径。我们观察到类似的现象：智能体会尝试几次访问标记为‘a’的状态，然后再前往标记为‘b’的状态。然而，在生成更多路径后，我们发现智能体偶尔会先访问标记为‘b’的状态，因为事件‘a’在这些状态上发生的概率为

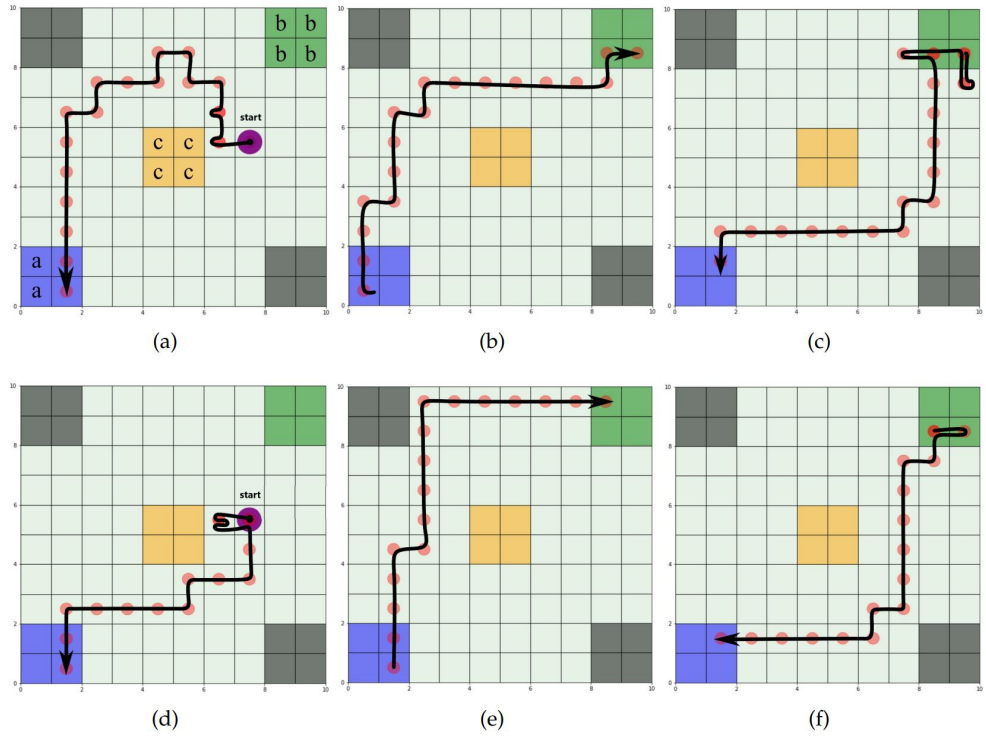


图 11 存在标记不确定性的动态事件中任务在 ϕ_2 下的一条生成路径：**(1)** 若智能体完全了解任务（q 状态序列）：**(a)** q_0上的路径。**(b)** q_1上的路径。**(c)** 返回 q_0的路径。**(2)** 若智能体不了解任务（标签序列）：**(d)** q_0上的路径。**(e)** q_1上的路径。**(f)** 返回 q_0的路径。

10%。这种现象仅在智能体不了解任务时出现。

接下来，假设事件是静态的，即不存在标记不确定性。我们根据智能体是否完全了解任务，考虑两种情形。在获得每种情形下的最优策略后，我们生成路径，以展示智能体完成任务的过程。图 12 **(1)** 展示了智能体如何基于观测历史和 q 状态历史规划运动。另一方面，图 12**(2)** 展示了智能体如何遵循基于观测历史和标签接收历史的策略完成任务。沿着两条路径，智能体都能完成任务，两种方法的表现也相近。与动态事件情形相比，我们没有观察到智能体在离开前会多次尝试访问'a' 或 'b' 状态。

图 13 比较了上述四种情形下任务 2 累积奖励的变化情况。可以观察到，静态事件情形获得的累积奖励高于动态事件情形。这可能是由于

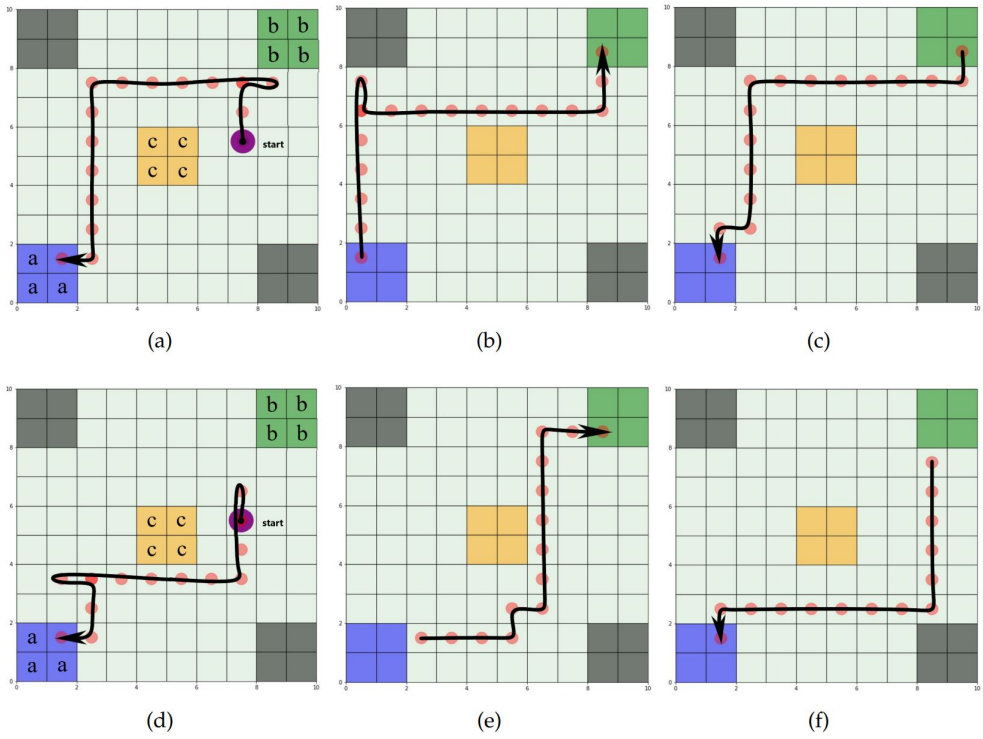


图 12 静态事件中任务在 ϕ_2 的一条生成路径：**(1)** 如果完全了解任务（q 状态序列）：**(a)** q_0 上的路径。**(b)** q_1 上的路径。**(c)** 返回 q_0 上的路径。**(2)** 如果不了解任务（标签序列）：**(d)** q_0 上的路径。**(e)** q_1 上的路径。**(f)** 返回 q_0上的路径。

动态事件中的标记不确定性。此外，无论智能体是否了解任务，收敛后的累积奖励都相近。

5.2 PyBullet TurtleBot 仿真

图。 14 展示了由 PyBullet 3 生成的虚拟办公室环境。0 [35]。在办公空间中，有四间办公室‘a’、‘b’、‘c’和‘d’，一间储藏室‘S’、一间打印机室‘Print’，以及一个用于为 TurtleBot（即智能体）充电的补给站‘Sply’。此外，办公室‘a’和‘d’中有两扇大窗户，办公空间内还有多扇门。我们将该办公空间离散化为一个四乘四的网格世界，以生成部分可观测马尔可夫决策过程模型。考虑到运动不确定性，我们假设 TurtleBot 沿期望方向移动时，有 0.9 的概率能够成功执行其导航控制器；而它也可能朝其他方向移动，这些方向均分剩余的 0.1 概率。朝墙壁移动时，TurtleBot 将停留在原地。假设智能体完全了解分配的任务，我们在此办公室中，采用两种不同的观测设置测试所提出的无模型强化学习方法

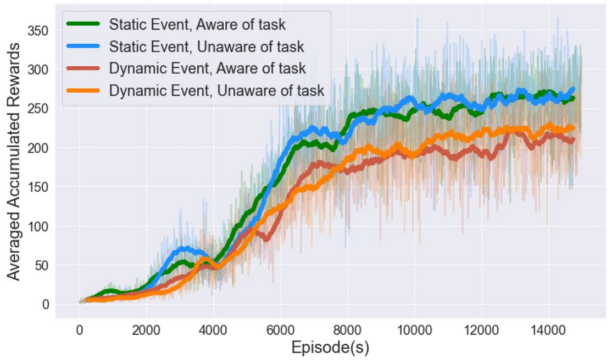


图13 任务 ϕ_2的奖励演化对比。

场景。所有仿真均进行 10,000 回合，每回合 300 步。其他计算设置与第 5.1 节相同。我们将每次仿真的办公室场景映射为网格世界，并在其中学习最优策略。随后，将所得最优策略应用于 PyBullet 平台上的虚拟 TurtleBot，以验证任务完成情况。

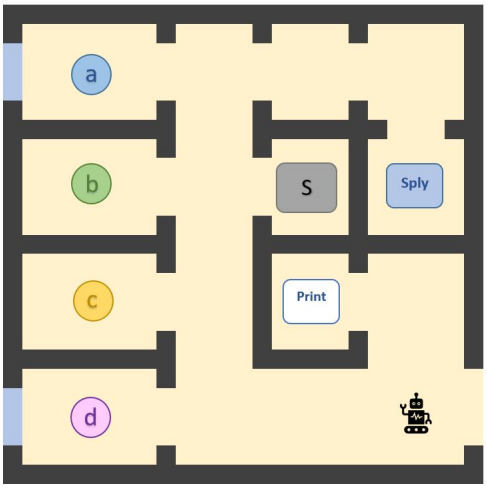


图14 办公室环境。

5.2.1 周围环境观测

该设定假设，在当前状态下，TurtleBot 可以按照“北”“西”“南”到“东”的特定顺序，收集四个方向上的周围观测。观测元素包括“墙”“走廊”“门”和“窗户”。智能体在每个方向上只能观测到一个元素。需要注意的是，

每个状态下只有一个观测，即 $O(s',\ a,\ o) = 1$。然而，智能体可能在两个或更多状态下感知到相同的观测。例如，办公室“b”和“c”具有相同的观测：$o(\text{“b”}) = o(\text{“c”}) = \{\text{“墙”“墙”“墙”“门”}\}$。因此，该部分可观测马尔可夫决策过程问题中的观测集合 O 包含 13 种不同的观测。

任务 ***1***：

任务 1 要求 TurtleBot 反复前往打印机室收集文件，然后将文件送到办公室‘a’或‘c’。同时，TurtleBot 必须始终避免进入储藏室‘S’。与公式（13）类似，这项任务可以用公式（15）中的线性时序逻辑公式表示。图 15（a）展示了诱导出的 LDGBA，我们还采用了公式（14）中重新设计的奖励。

$$\varphi_{task1} = \square\diamondsuit(Print \wedge \diamondsuit(a \mid c)) \wedge \square\neg S \tag{15}$$

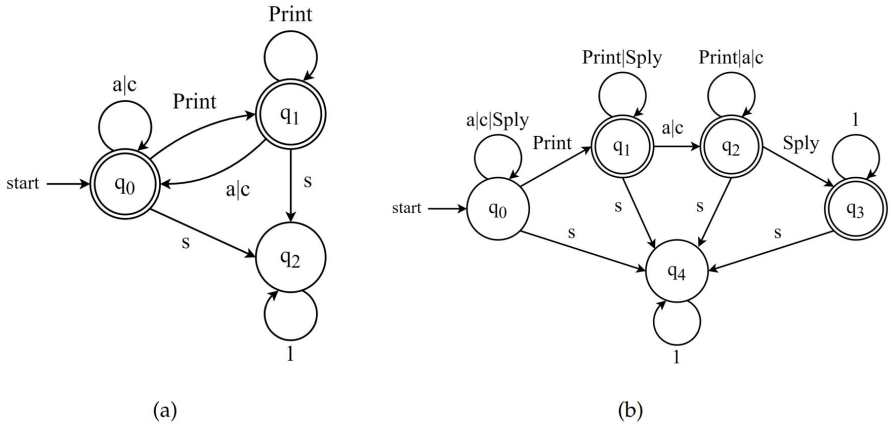


图 15 诱导出的极限确定性广义布奇自动机：（**a**）φ_{task1}。（**b**）φ_{task2}。

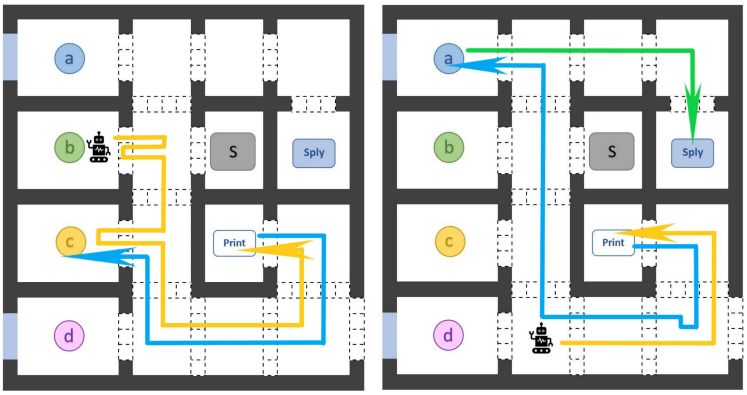
训练过程收敛后，可以从 Q 网络中导出最优策略。图 16（a）展示了一条生成路径，TurtleBot 可沿此路径完成任务 1。可以看到，TurtleBot 离开初始状态（办公室‘b’）后，朝打印机室移动，黄色路径表示其行进路线。由于运动不确定性，它在前往打印机室的途中多次经过办公室‘b’和‘c’。TurtleBot 到达打印机室并收集文件后，蓝色路径显示它离开打印机室，前往办公室‘c’送交文件。

任务 ***2***：

在此，我们将任务1扩展为一项更复杂的任务。任务 2 表示，TurtleBot 必须先递送文件，再前往补给站充电，然后重复执行任务 1。任务 2 的线性时序逻辑公式可表示为方程（16），

和图 15(b) 展示了对应的 LDGBA。为简化图示，仅保留带有单个标签的转移。

$$\varphi_{task2} = (\neg(a \mid c)\mathcal{U}Print) \wedge (\neg Sply\mathcal{U}(a \mid c)) \wedge (\Diamond Sply) \wedge (\Box\neg S) \tag{16}$$



(a) (b)

图 16 TurtleBot 在能够观测四个方向的周围环境时完成任务所生成的路径：**(a)** 任务 1。**(b)** 任务 2。

图16(b) 展示了智能体根据学习策略生成的路径，以完成任务 2。TurtleBot 起始于办公室‘d’外。随后，它前往打印机室，收集文件，再前往办公室‘a’，分别以黄色和蓝色路径表示。最后，绿色路径显示智能体在递送文件后到达“Sply”站充电。

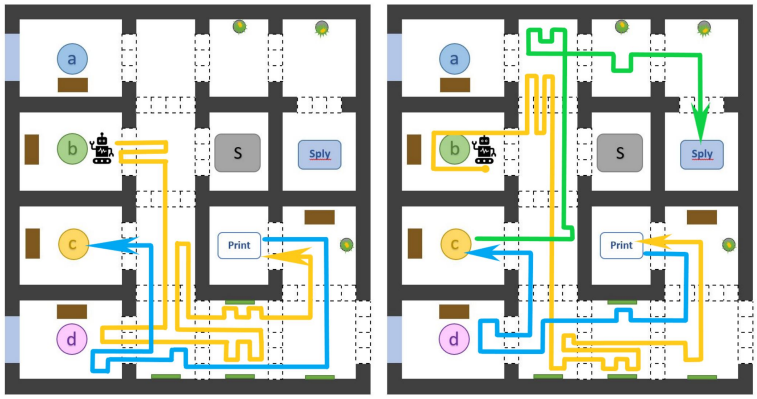
5.2.2 单方向观测

我们还考虑了另一种观测设置：TurtleBot 在每个状态下随机观测一个方向。因此，观测不确定性显著增加。如图 17所示，我们在 PyBullet 办公室环境中添加了一些物品，以促进策略收敛。因此，可观测空间包括“走廊”、“墙”、“门”、“窗户”、“桌子”、“墙上的画”和“墙边的花”。例如，在办公室“b”和“c”中，智能体分别以 50%、25% 和 25% 的概率观测到“墙”、“桌子”和“门”。

图 18分别展示了智能体完成第 5.2.1 节中相同任务时生成的路径。与图16中采用四个方向观测的路径相比，单个观测元素让智能体对当前状态的感知更少。图 18 还表明，智能体难以判断正确的移动方式。例如，图18 (a) 中的黄色路径显示，智能体在墙上有多幅画的走廊里来回移动了几次，之后才最终前往打印机室。这主要是由观测不确定性和运动不确定性共同造成的。



图 17 智能体只能观察一个方向时的改造后的办公室环境。



(a) (b)

图 18 智能体只能观察一个方向时，完成单轮指定任务的路径：（a）任务 1。（b）任务 2。

5.3 多智能体仓库仿真

我们还初步将所开发的无模型强化学习方法应用于一个多智能体问题。一个小型工厂仓库被建模为 8×8 网格世界，如图 19 所示。这里有两个智能体，每个智能体都必须反复将一个箱子从标记为蓝色“a”的某个位置运到标记为绿色“b”的任意位置，并将箱子放到传送带上。较深和较浅的灰色状态分别表示墙和其他不可移动的包裹。在这种情况下，每个智能体可以“向上”、“向下”、“向右”、“向左”和“停留”，但它们不能同时移动到同一位置。一旦两个智能体在一轮内都完成任务，位置“a”处就会有更多箱子等待搬运。每个智能体成功将箱子运到目标位置后都会获得奖励。此外，智能体的

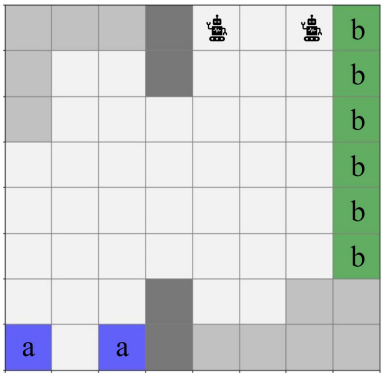


图 19 仓库环境。

初始位置如图 19所示。观测和转移概率的设置与第 2.3 节相同。

我们可以通过线性时序逻辑（LTL）将团队任务表述为 $\phi = \square\Diamond(a \wedge \Diamond b)$，与公式（13）类似。此外，我们还针对任务可行性实施了第 5.1.2 节提到的奖励重设计。多智能体强化学习（MARL）研究多个智能体如何在共同环境中交互。根据智能体的实际任务需求，MARL 可分为以下三大类 [36]:

- 合作型：智能体合作以实现团队目标，其中没有任何一个智能体能够独自完成整个任务。
- 竞争型：智能体彼此竞争。
- 混合型：智能体最大化效用，这可能需要合作和/或竞争。

此外，多智能体系统还可以采用集中式或去中心化框架进行建模，智能体可以分别学习一个集中式策略或多个独立策略 [36], 。

在这种情况下，智能体在同一环境中朝着共同目标协同工作。这个多智能体强化学习问题可归类为合作型问题。我们将该问题建模为一个简化的去中心化部分可观测马尔可夫决策过程（Dec-POMDP）框架。每个智能体都有一组动作 A^i 和一组观测 O^i，其中 i 表示智能体的索引：$i \in \{1, 2\}$。然而，由于它们是在相同状态空间中的同类智能体，我们定义$A^1 = A^2$ 和 $O^1 = O^2$。在这种情况下只考虑静态事件。

已有大量研究从不同角度解决多智能体强化学习问题。受 Zhou 等人 [37], 的工作启发，我们在智能体之间建立合作型通信，并将其应用于我们的无模型强化学习算法中。我们假设智能体完全了解任务。为每个智能体初始化两个独立的 Q 网络。每个 Q 网络的输入由观测序列、 q 状态序列以及 q 另一个智能体的状态序列构成。智能体通过同一通信网络中的消息共享任务识别信息。相应的 Q 网络如图20所示。由于两个 Q 网络经过独立训练，分别用于估计每个智能体的 Q值，因此我们将这种方法视为去中心化训练与执行。

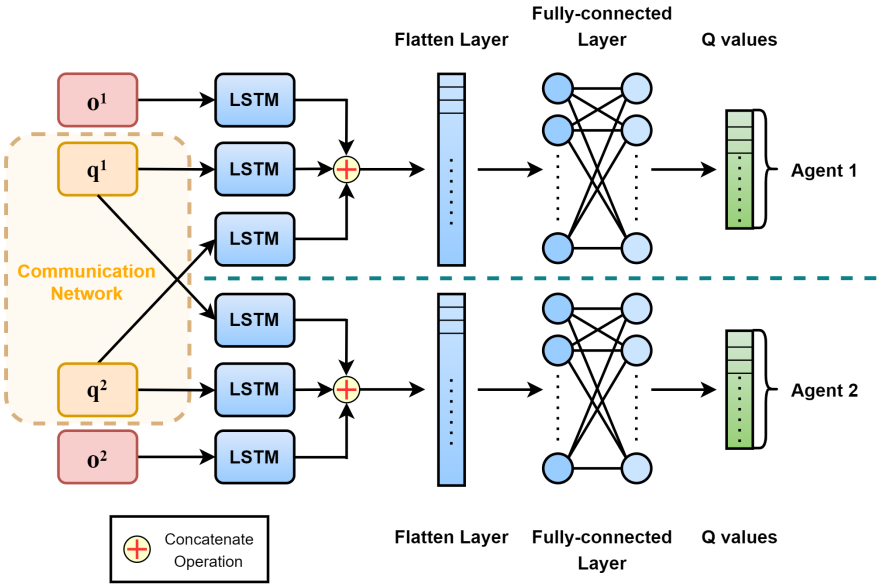


图 20 多智能体强化学习问题的 Q网络架构。

仿真共进行了 30,000 个回合，每个回合 300 步。图21展示了累积奖励随回合变化的趋势。

图 22 展示了两个智能体生成的路径，分别以红色和黄色路径表示。在图 22 （a）中任务开始时，两个智能体都停留在原始位置，以收集足够的周围环境观测用于决策。随后不久，它们前往蓝色装载区，黄色智能体在前往装载区的路径上停留在某一状态，以避免与另一个智能体（红色）发生碰撞。包裹装载完毕后，智能体朝绿色区域行进，并将包裹放到传送带上，如图 22 （b）所示。第二次运行在任务完成后立即开始，如图 22 （c）所示。需要注意的是，黄色智能体数次选择“停留”动作，以等待另一个智能体通过，因为我们优先考虑红色智能体。生成路径表明，即使在有限通信条件下为智能体分别训练策略，期望的协作任务仍然得以完成。

6 讨论

本研究将部分可观测环境中自主智能体的运动规划建模为 PL-POMDP 问题。我们还通过线性时序逻辑表达高层复杂任务，并将其转换为 LDGBA，以进行模型检测。因此，这类运动规划问题等价于在 PL-POMDP 与其诱导的 LDGBA 的乘积上寻找最优策略，并据此提出了一种无模型强化学习方法。我们采用 LDGBA 进行模型检测，因为与相应的确定性再生自动机相比，它们具有更小的自动机状态空间。此外，它们具有多个接受集，使智能体能够无限次访问所有接受

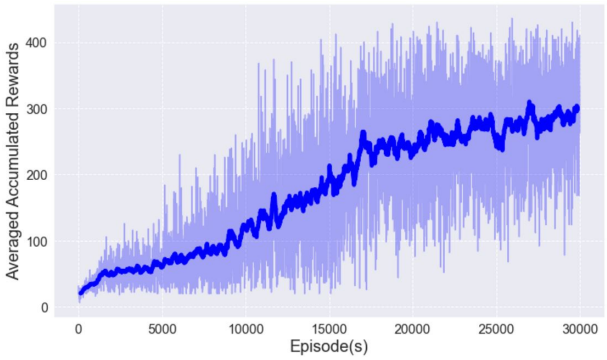
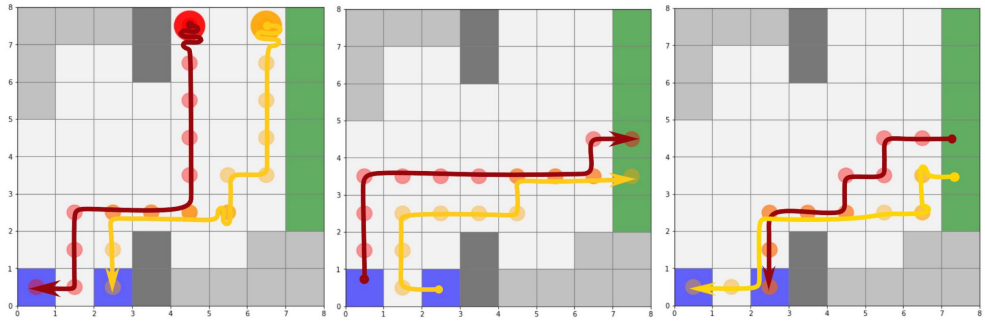


图 21 多智能体强化学习案例的累积奖励。



(a) (b) (c)

图 22 红色和黄色智能体为完成指定任务而生成的路径。

集。通过实施奖励重设计，修改后的 LDGBA 能够解决 LDBA 在强化学习运动规划中造成的奖励稀疏问题 [28]。未来，我们计划利用[33] 提出的跟踪前沿函数，跟踪 LDGBA 中尚未访问的接受集。当部分可观测环境中需要执行更复杂的无人机监视任务时，这一方法尤其有用。通过进一步完善该函数，我们希望提升所提方法的性能和适用性。

所提出的无模型强化学习方法采用深度Q学习，在乘积PL-POMDP上学习最优策略。具体而言，该方法将长短期记忆网络融入Q网络架构，以处理智能体的观测历史并进行任务识别。使用诱导的自动机状态或感知到的标签历史来表示任务识别，取决于智能体是否完全了解任务。仿真示例表明，在具有复杂任务的部分可观测环境中，这两种表示方式都能提升智能体的学习效果。我们还研究了采用不同长度的输入序列时Q网络的表现。序列过长或过短都可能导致训练过程不稳定。本研究采用的序列长度能够为智能体提供充足的信息，以便

进行决策，并避免在开始动作选择和遵循训练策略之前出现冗长的延迟。需要注意的是，长度选择可能取决于所解决的问题。未来的工作可以轻松地将所提出的方法更新为其他基于价值或基于策略的深度强化学习（DRL）方法。

在本研究中，我们仅将所提出的无模型强化学习方法应用于离散环境中自主智能体的运动规划。未来的工作将把该方法扩展到具有连续状态与动作空间的部分可观测马尔可夫决策过程问题。此外，一般的 LDGBA 可以由非确定性集合和确定性集合组成。尽管与我们仿真实例中所考虑任务相对应的大多数 LDGBA 仅包含确定性集合，但其中一个实例表明，所提出的方法能够处理 ϵ-转移，而这类转移具有非确定性。即使初始状态位于非确定性集合中，在发生 ϵ-转移后，自动机状态转移也会被限制在确定性集合中。由于接受状态仅位于确定性集合中，所提出的方法仍然能够获得最优策略。

我们还对一个协作式多智能体系统进行了初步仿真，以展示所提出方法的多功能性和适应性。此外，通过实现智能体之间的任务识别通信，所获得的策略表现出高度的协作性，从而成功完成任务。尽管当前多智能体仿真采用去中心化训练与执行框架，通过简化场景来开展研究，而该框架通常可能具有非平稳性 [36]，但它有潜力用于处理由多个无人机组成的协作任务系统。未来，使用能够处理连续状态空间的集中式训练与去中心化执行（CTDE）框架可能更为合适。此外，智能体之间的任务识别通信依赖于它们对任务的完整了解。另一种方法可以利用来自其他相邻智能体的局部观测，因此需要额外的神经网络来近似该通信机制。

致谢。 Li 和 Xiao 谨此感谢美国教育部（ED#P116S210005）和美国国家科学基金会（#2226936）对本研究的支持。

声明

•资助：本研究由美国教育部（ED#P116S210005）和美国国家科学基金会（#2226936）资助。

• 利益冲突：作者声明不存在任何利益冲突。
•代码可用性：代码文件可在以下网址获取：https://github.com/JunchaoLi001/ 基于 LDGBA 的部分可观测马尔可夫决策过程模型无关深度强化学习长短期记忆网络

参考文献

[1]Kurniawati, H.：部分可观测马尔可夫决策过程与机器人学。《控制、机器人与自主系统年度评论》**5** （2022年）https://doi.org/10.1146/annurev-control-042920-092451

[2] Kaufman, H., Howard, R.A.：动态规划与马尔可夫过程。《美国数学月刊》68（1961年）https://doi.org/10.2307/2312519[3] Perez, A., Platt, R., Konidaris, G., Kaelbling, L., Lozano-Perez, T.：Lqr-rrt*：具有自动推导扩展启发式的最优基于采样的运动规划。载于：2012年IEEE机器人与自动化国际会议，第2537–2542页（2012年）。IEEE[4] Sutton, R.S., Barto, A.G.：强化学习：导论。麻省理工学院出版社，剑桥，马萨诸塞州（2018年）。第1、2、3章[5]Shani, G., Pineau, J., Kaplow, R.：基于点的POMDP求解器综述。《自主智能体与多智能体系统》27（2013年）https://doi.org/10. 1007/s10458-012-9200-2[6]Pineau, J., Gordon, G., Thrun, S.：基于点的价值迭代：一种用于POMDP的任意时刻算法。（2003年）[7] SARSOP：通过近似最优可达信念空间实现高效的基于点的POMDP规划，第4卷（2009年）。https://doi.org/10.15607/rss.2008.iv.009[8] Icarte, R.T., Waldie, E., Klassen, T.Q., Valenzano, R., Castro, M.P., McIlraith, S.A.：为部分可观测强化学习学习奖励机，第32卷（2019年）[9] Mnih, V., Silver, D., Riedmiller, M.：使用深度Q学习玩雅达利。神经信息处理系统大会（2013年）[10] 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.：通过深度强化学习实现人类水平控制。《自然》518（2015年）https://doi.org/10.1038/nature14236[11] Krizhevsky, A., Sutskever, I., Hinton, G.E.：使用深度卷积神经网络进行ImageNet分类。《美国计算机学会通讯》60（2017年）https: //doi.org/10.1145/3065386[12] Hausknecht, M., Stone, P.：针对部分可观测马尔可夫决策过程的深度循环Q学习，第FS-15-06卷（2015年）[13] Hochreiter, S., Schmidhuber, J.：长短期记忆网络。《神经计算》9（1997年）https://doi.org/10.1162/neco.1997.9.8.1735[14] Jakob N. Foerster, N.d.F. Yannis M. Assael, Whiteson, S.：使用深度分布式循环Q网络学习通过交流解决谜题（2016年）[15] Pengfei Zhu, P.P. Xin Li, Miao, G.：关于改进深度强化学习

[16] Nicolas Heess, T.P.L. Jonathan J Hunt, Silver, D.：使用循环神经网络的基于记忆的控制（2015）[17] Meng, L., Gorbet, R., Kulic, D.：面向部分可观测马尔可夫决策过程的基于记忆的深度强化学习。（2021）。https://doi.org/10.1109/IROS51168.2021.9636140[18] Baier, C., Katoen, J.-P.：模型检验原理，第950卷，（2008）[19] Kˇret´ınsk´y, J., Meggendorfer, T., Sickert, S.：Owl：用于 ω-词、自动机和线性时序逻辑的库，第11138卷 LNCS（2018）。https://doi.org/10.1007/978-3-030-01090-4 34[20] Chatterjee, K., Chmel´ık, M., Gupta, R., Kanodia, A.：具有时序逻辑规范的部分可观测马尔可夫决策过程的定性分析及其在机器人应用中的应用，第2015年6月卷（2015）。https://doi.org/10.1109/ICRA.2015.7139019[21] Sharan, R., Burdick, J.：具有线性时序逻辑规范的部分可观测马尔可夫决策过程的有限状态控制。（2014）。https://doi.org/10.1109/ACC.2014.6858909[22] Bouton, M., Kochenderfer, M.J.：部分可观测马尔可夫决策过程中的模型检测基于点的方法。（2020）。https://doi.org/10.1609/aaai. v34i06.6563[23] Ahmadi, R.. M.; Sharan, Burdick, J.W.：具有线性时序逻辑规范的部分可观测马尔可夫决策过程的随机有限状态控制（2020）[24]Carr, S., Jansen, N., Wimmer, R., Serban, A., Becker, B., Topcu, U.：使用循环神经网络对部分可观测马尔可夫决策过程进行反例引导的策略改进，第2019年8月卷（2019）。https://doi.org/10.24963/ijcai.2019/768[25] Carr, S., Jansen, N., Topcu, U.：时序逻辑约束下部分可观测马尔可夫决策过程的可验证的基于循环神经网络的策略，第2021年1月卷（2020）。https://doi.org/10.24963/ijcai.2020/570[26]Hahn, E.M., Perez, M., Schewe, S., Somenzi, F., Trivedi, A., Wojtczak, D.：无模型强化学习中的ω-正则目标，第11427卷 LNCS（2019）。https://doi.org/10.1007/978-3-030-17462-0 27[27] Hasanbeig, M., Kantaros, Y., Abate, A., Kroening, D., Pappas, G.J., Lee, I.：具有概率满足度保证的时序逻辑控制综合强化学习，第2019年12月卷（2019）。https://doi.org/10.1109/CDC40024.2019.9028919[28]Oura, R., Sakakibara, A., Ushio, T.：使用极限确定性广义 Büchi 自动机学习线性时序逻辑规范的控制策略。IEEE Control Systems Letters 4（2020）https://doi.org/10.1109/ LCSYS.2020.2980552

[29]基于强化学习、规划和教学的自我改进反应式智能体。Machine Learning **8**（1992年）https://doi.org/10.1007/bf00992699[30] 使用无模型强化学习从线性时序逻辑规范进行控制综合。（2020年）。https://doi.org/10.1109/ICRA40945.2020.9196796[31]Sickert，S.，Esparza，J.，Jaax，S.，Kˇret´ınsk´y, J.：用于线性时序逻辑的极限确定性 B¨uchi 自动机，第9780卷（2016年）。https://doi.org/10.1007/978-3-319-41540-6 17[32] Cai, M., Hasanbeig, M., Xiao, S., Abate, A., Kan, Z.：结合时序逻辑的连续运动规划模块化深度强化学习。IEEE Robotics and Automation Letters **6** （2021年）https://doi.org/10.1109/LRA.2021.3101544[33] Cai, M., Xiao, S., Li, B., Li, Z., Kan, Z.：基于强化学习的具有最大概率满足度的时序逻辑控制，第2021年5月卷（2021年）。https://doi.org/10.1109/ICRA48506.2021.9561903[34] Mingyu, C., Xiao, S., Li, Z., Kan, Z.：具有潜在不可满足线性时序逻辑约束的最优概率运动规划。IEEE Transactions on Automatic Control（2021年）https://doi.org/10.1109/TAC.2021.3138704[35] Coumans, E., Bai, Y.：PyBullet，一个用于游戏、机器人学和机器学习物理仿真的 Python 模块。http://pybullet.org（2016年至2021年）[36] Oroojlooy, A., Hajinezhad, D.：合作型多智能体深度强化学习综述。Appl Intell（2022年）https://doi.org/10.1007/s10489-022-04105-y[37] Zhou, W., Li, J., Zhang, Q.：深度强化学习在无人机群多目标跟踪中的联合通信与动作学习。Drones **6**（11），339（2022年）