具有指数图信息通信的大规模情境多智能体强化学习

李方昱 [1, 2] 刘金溢 [1, 2] 孙浩源 [1, 2] 韩红桂 [1, 2]

摘 要 多智能体强化学习 (MARL) 在协同任务中展现出卓越的性能. 然而, 在具有复杂交互关系的大规模多智能体系统 (MAS) 中, 传统的 MARL 算法由于缺乏高效的通信机制, 性能往往受到限制. 为提升 MARL 在大规模 MAS 中的性能, 本文提出一种具有指数图信息通信的情境 MARL 算法 (EMAGIC). 首先, 设计基于单点指数图的通信拓扑结构, 每个智能体在每个时间步仅与一个智能体进行通信, 消息通过循环通信链路传递给所有智能体. 其次, 构建图信息通信机制, 利用门控循环单元编码多个时间步的消息, 并通过最大化同一时间步不同智能体间消息的互信息来优化消息的编码特征. 最后, 构建独立情境记忆 (EM) 模块, 建立平均回报与全局状态的对应关系以构建记忆库, 利用 EM 目标与个体价值均值的误差来构建损失函数. 在多个大规模多智能体环境中的实验结果表明, EMAGIC 始终优于最先进的 MARL 基线方法.

关键词 多智能体系统; 强化学习; 多智能体通信; 指数图

引用格式 李方昱, 刘金溢, 孙浩源, 韩红桂. 具有指数图信息通信的大规模情境多智能体强化学习. 自动化学报, 2026, **52**(6): 1304−1318

DOI 10.16383/j.aas.c250633 **CSTR** 32138.14.j.aas.c250633

Large-scale Episodic Multi-agent Reinforcement Learning With Exponential Graph Information Communication

LI Fang-Yu[1, 2] LIU Jin-Yi[1, 2] SUN Hao-Yuan[1, 2] HAN Hong-Gui[1, 2]

Abstract Multi-agent reinforcement learning (MARL) demonstrates excellent performance in cooperative tasks. However, in large-scale multi-agent systems (MAS) with complex interaction relationships, traditional MARL algorithms perform poorly due to the lack of efficient communication mechanisms. To enhance the performance of MARL in large-scale MAS, this paper proposes an episodic MARL algorithm with exponential graph information communication (EMAGIC). First, this paper designs a one-peer exponential graph-based communication topology, where each agent communicates with only one other agent at each time step and transmits messages to all agents through a cyclic communication link. Second, this paper constructs a graph information communication mechanism that uses a gated recurrent unit to encode messages across multiple time steps and optimizes the encoded features of the messages by maximizing the mutual information between messages of different agents at the same time step. Finally, this paper builds an independent episodic memory (EM) module to establish the correspondence between average returns and global states for constructing a memory bank, and constructs the loss function by using the error between the EM target and the mean of individual values. Experimental results in multiple large-scale multi-agent environments show that EMAGIC consistently outperforms advanced MARL baseline methods.

Keywords multi-agent systems; reinforcement learning; multi-agent communication; exponential graph

Citation Li Fang-Yu, Liu Jin-Yi, Sun Hao-Yuan, Han Hong-Gui. Large-scale episodic multi-agent reinforcement learning with exponential graph information communication. *Acta Automatica Sinica*, 2026, **52**(6): 1304−1318

多智能体强化学习 (multi-agent reinforcement learning, MARL) 依托分布式协作决策优势, 在多机器人协同[1]、多无人机控制[2] 以及多无人车导航[3] 等复杂协同任务中展现出卓越性能[4]. 作为 MARL 的主流训练框架, 集中训练分散执行 (centralized training with decentralized execution, CTDE) 衍生出多种基准算法: 值分解网络 (value decomposition network, VDN)[5] 通过线性求和将全局 Q 值分解为个体 Q 值, 实现多智能体价值协

收稿日期 2025-11-17 录用日期 2026-01-30
Manuscript received November 17, 2025; accepted January 30, 2026
国家重点研发计划 (2023YFB3307300), 国家自然科学基金 (62373014, 92467205, 62522302, 62473011), 北京市科技新星项目 (20250484938), 北京市自然科学基金−小米创新联合基金 (L253010) 资助
Supported by National Key Research and Development Program of China (2023YFB3307300), National Natural Science Foundation of China (62373014, 92467205, 62522302, 62473011), Beijing Nova Program of Science and Technology (20250484938), and Beijing Natural Science Foundation-Xiaomi Innovation Joint Fund (L253010)
本文责任编委 罗彪
Recommended by Associate Editor LUO Biao
1. 北京工业大学信息科学技术学院 北京 100124 2. 数字社区教育部工程研究中心 北京 100124
1. School of Information Science and Technology, Beijing University of Technology, Beijing 100124 2. Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124

同建模; Q 值混合网络 (Q-mixing network, QMIX)[6] 引入非线性混合网络, 在满足单调性约束的同时捕捉智能体间非线性关联; 多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradient, MADDPG)[7] 利用各智能体的评价网络通过全局状态与联合动作评估策略价值; 多智能体双延迟深度确定性策略梯度 (multi-agent twin delayed deep deterministic policy gradient, MATD3)[8] 通过双评价网络裁剪机制降低过估计偏差, 利用延迟策略更新缓解训练震荡, 提升算法训练稳定性; 多智能体近端策略优化 (multi-agent proximal policy optimization, MAPPO)[9] 则通过优势函数约束策略更新步长. 此外, 独立深度 Q 网络 (independent deep Q-network, IDQN)[6] 采用独立学习范式, 将深度 Q 网络 (deep Q-network, DQN)[10] 扩展至多智能体场景中, 各智能体独立更新 Q 网络, 在奖励偏差较大的环境中表现突出. 随着现实场景需求升级, 智能体数量多、交互复杂的大规模多智能体系统 (multi-agent systems, MAS) 日益普遍[11], 在大规模 MAS 中, 传统的基准 MARL 算法面临严峻收敛挑战: 一方面, 智能体规模扩大导致联合动作与状态组合呈指数级增长, 有效经验样本占比极低, 需海量交互才能收集高价值样本; 另一方面, 传统经验回放机制的随机采样方式难以充分利用高价值样本, 样本效率低下[12].

情境记忆 (episodic memory, EM) 凭借提升样本效率的核心优势被引入 MARL 领域[13]. 在单智能体领域, EM 已得到广泛应用. 例如: 无模型情境控制 (model-free episodic control, MFEC)[14] 通过存储状态−动作对的历史最大回报来加速学习进程; 神经情境控制 (neural episodic control, NEC)[15] 采用可微分神经字典提升记忆泛化能力; 情境记忆 DQN (episodic memory DQN, EMDQN)[16] 将 EM 与 DQN 结合, 通过 EM 目标正则化时序差分 (temporal-difference, TD) 目标, 提升 DQN 的样本效率. 尽管 EM 在单智能体强化学习 (reinforcement learning, RL) 上已被成功应用, 然而其在 MARL 中的应用仍处于初步阶段. Ma 等[13] 提出基于状态的 EM (state-based EM, SEM), 仅基于全局状态构建记忆库并证明其收敛性. 若记忆库样本无法匹配任务动态需求, SEM 易因重复学习次优样本陷入次优策略. 因此, 当前将 EM 与 MARL 相结合的研究, 多依托探索机制展开[12, 17], 通过探索挖掘更多高奖励状态来丰富记忆库, 该思路在稀疏奖励环境中尤为常见. Zheng 等[12] 提出具有好奇心探索的情境 MARL (episodic MARL with curiosity-driven exploration, EMC), 使用个体 Q 值的预测误差作为内在奖励引导探索新的状态, 并通过 EM 模块保存高回报轨迹来加快协同策略的学习. Peng 等[17] 通过在 EMC 中添加图神经网络 (graph neural network, GNN) 以及进化算法 (evolutionary algorithm, EA), 将智能体间的信息交流与 EMC 结合, 其中 EMC 探索高回报轨迹并存储, GNN 建模智能体间的协作沟通, EA 优化联合策略参数. 尽管 EM 与探索机制结合展现出良好的性能, 但在大规模 MAS 中, 探索可能会加剧环境的非平稳性, 即单个智能体的探索动作会对其他智能体的观测与决策产生连锁影响, 降低学习稳定性[18].

环境非平稳性的原因之一是智能体间缺乏高效信息交互机制[19]. 构建高效通信机制可使智能体共享局部信息, 突破部分可观测的限制, 缓解环境的非平稳性问题[20]. 通信机制因能促进 MAS 协同并提升任务执行效果, 已逐渐成为 MARL 领域的研究热点. 其中通信拓扑作为信息交换的基础, 直接决定智能体之间的消息传递效率, 进而影响 MARL 的收敛性能及奖励获取能力. 部分研究致力于实现全局信息共享. 例如, Sukhbaatar 等[21] 提出的通信网络模型 (communication network model, CommNet) 采用全局通信拓扑, 所有智能体均与环境内其他智能体建立固定的通信链路, 无通信范围和邻居数量限制. CommNet 的通信链路数量随智能体数量呈平方级增长, 在大规模 MAS 中会产生高额通信开销. 为此, 图结构通信拓扑被引入到 MARL 领域中, 其将智能体抽象为节点、通信关系抽象为边, 仅允许边连接的邻居智能体进行通信, 降低了通信开销. Chu 等[22] 提出的神经通信协议 (neural communication protocol, NeurComm) 为大规模交通信号控制预设图结构拓扑, 智能体仅与预设邻居交互, 适配路口位置固定、交通流关系稳定的场景, 但在存在运动体的动态环境中, 固定拓扑无法实时调整边连接关系, 易出现无效通信或必要通信缺失的问题. Jiang 等[23] 提出的图卷积强化学习 (graph convolutional reinforcement learning, DGN) 则动态选择距离最近的 3 个智能体建立链路, 每隔一个时间步更新邻居集合, 通过多头注意力机制融合邻居消息特征, 但仅 3 个邻居的信息覆盖范围过小, 远距离智能体需多轮传递才能获取消息, 可能错过关键协同对象. Agarwal 等[24] 提出基于智能体−实体图的通信拓扑, 包含智能体与静态环境实体, 边由感知半径确定, 但需智能体在回合开始时感知所有实体位置, 不适用于存在遮挡或动态障碍物的场景. 进一步, Nayak 等[25] 提出的智能信息聚合 (intelligent information aggregation, InforMARL), 将障碍和目标纳入图节点, 根据距离感知半径选择

邻居, 建立智能体间双向通信、智能体与非智能体间单向通信, 但感知半径需人工设定, 难以在大规模 MAS 中兼顾通信开销与消息传递能力.

除通信拓扑以外, 通信机制另一个重要的部分是构建智能体之间传递的消息特征. 消息特征将智能体分散的局部观测整合为全局协同所需的有效信息, 提高 MAS 的协同性能. DGN[23] 通过多层卷积从局部观测中提取信息, 借助注意力机制筛选关键特征实现协同; CommNet[21] 让智能体将经门控循环单元 (gated recurrent unit, GRU) 编码的局部观测隐藏状态发送给其他智能体, 接收端通过均值聚合完善环境观测. 为进一步优化消息特征, 部分研究引入额外损失函数: Lo 等[26] 提出通信对齐对比学习 (communication alignment contrastive learning, CACL), 将同一时间步不同智能体消息作为正样本、非当前时间步消息作为负样本, 通过对比学习优化编码, 增强去中心化场景下消息与全局语义的相关性. Li 等[11] 提出指数拓扑可扩展通信 (exponential topology-enabled scalable communication, ExpoComm), 采用 GRU 存储智能体历史观测轨迹, 将隐藏状态、自身消息、其他智能体消息分别作为注意力模块的查询、键、值生成新消息, 通过全局状态预测误差或对比损失优化消息全局相关性. 尽管以上研究能够使算法最终收敛至高值回报, 但在大规模 MAS 中, 通信机制的引入虽能使智能体获取其他智能体的局部信息, 但训练前期智能体缺乏有效协作策略, 可能导致算法收敛速度下降. 为此, 本文考虑将通信机制与 EM 模块结合, 通信机制通过提高 MAS 的协同性能来采集高价值样本, 并结合 EM 对这些高价值样本进行有效利用, 共同实现 MARL 在大规模 MAS 中快速收敛至高值回报. 本文提出具有指数图信息通信的情境 MARL 算法 (episodic MARL with exponential graph information communication, EMAGIC), 建立基于单点指数图的通信拓扑以权衡开销与消息传递, 设计最大化互信息的图信息通信方法以提升交互效率, 构建适配独立与集中学习的 EM 模块以优化收敛性, 实现大规模 MAS 的高效协同决策. 本文的贡献如下:

1) 设计基于单点指数图的通信拓扑结构, 在保证低通信成本的同时实现信息高效传播, 并从理论层面证明该拓扑结构的消息传递能力.

2) 提出图信息通信机制, 使智能体可基于自身观测与接收的消息获取环境全局信息, 提升 MAS 的协同性能以获取高价值奖励.

3) 开发适用于独立学习的 EM 模块, 建立全局状态与平均回报的关系; 通过高效利用通信机制获取的高价值样本, 提升 EMAGIC 的收敛性能.

1 预备知识

1.1 去中心化部分可观测马尔科夫决策过程

协同多智能体任务可形式化建模为去中心化部分可观测马尔科夫决策过程 (decentralized partially observable Markov decision process, DecPOMDP), 由元组 $\langle N, \mathcal{S}, \mathcal{A}, O, \mathcal{R}, T, \gamma\rangle$ 表示. 其中, N 为环境中的智能体总数; $\mathcal{S}$ 为 MAS 的全局状态空间; $\mathcal{A}$ 为单智能体的有限动作集合, 联合动作空间满足 $\boldsymbol{\mathcal{A}} = \mathcal{A}^N$; O 为从全局状态与智能体编号到个体观测空间 $\mathcal{O}$ 的观测函数; $\mathcal{R}$ 为从全局状态与联合动作到实数域的共享奖励函数; T 为单回合最大时间步数; $\gamma \in [0, 1)$ 为用于权衡即时奖励与长期回报的折扣因子. DecPOMDP 遵循部分可观测性假设: 在任意时间步 $t \in [0, T-1]$, 智能体通过观测函数得到个体观测 $o_{i,t} = O(s_t, i)$; 每个智能体 i 维护截至当前步的观测−动作历史轨迹 $\tau_{i,t} = \{o_{i,0}, a_{i,0}, \cdots, o_{i,t}, a_{i,t}\}$, 并基于该轨迹构造个体策略 π_i 以最大化 MAS 的期望回报. 智能体 i 根据策略 π_i 采样得到动作 $a_{i,t} \in \mathcal{A}$, 所有个体动作构成联合动作 $\boldsymbol{a}_t = [a_{i,t}]_{i=0}^{N-1} \in \boldsymbol{\mathcal{A}}$; 环境在接收到联合动作后, 将根据状态转移概率 $Pr(s_{t+1}|s_t, \boldsymbol{a}_t)$ 生成下一时间步全局状态 s_{t+1}, 同时系统通过奖励函数获得即时奖励 $r_t = \mathcal{R}(s_t, \boldsymbol{a}_t)$. MARL 的核心目标是求解由所有个体策略构成的联合策略 $\boldsymbol{\pi} = [\pi_i]_{i=0}^{N-1}$, 以最大化系统的期望累计折扣回报, 对应的联合价值函数与联合动作价值函数分别定义为 $V^{\boldsymbol{\pi}}(s) = \mathrm{E}_{\boldsymbol{\pi}, Pr}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k}|s_t = s\right]$ 和 $Q^{\boldsymbol{\pi}}(s, \boldsymbol{a}_t) = \mathcal{R}(s, \boldsymbol{a}_t) + \gamma \mathrm{E}_{s'\sim Pr(\cdot|s, \boldsymbol{a}_t)}\left[V^{\boldsymbol{\pi}}(s')\right]$, 二者满足马尔科夫决策过程的贝尔曼最优性方程.

1.2 情境多智能体强化学习

已有研究在 MARL 中引入 EM 模块, 通过存储全局状态的历史最优回报为策略训练提供最优价值参考, 该模块与 QMIX 结合后的架构如图 1 所示, B 为记忆库的大小. 采用从高斯分布中提取的固定随机矩阵 $\phi(s): \mathcal{S} \to \mathbf{R}^d$, 将全局状态 $s \in \mathcal{S}$ 转化为 d 维的投影特征; 记忆库保存投影特征−最大回报键值对信息: 以 $\phi(s_t)$ 表示键, 以 $H(\phi(s_t))$ 表示值, 其更新规则为

$$H(\phi(s_t)) = \begin{cases} \max\{H(\phi(\hat{s}_t)),\ R_t\}, & \|\phi(\hat{s}_t) - \phi(s_t)\|_2 < \delta \\ R_t, & \text{其他} \end{cases} \tag{1}$$

其中, $R_t = \sum_{t'=0}^{T-t-1}\gamma^{t'} r_{t+t'}$ 表示从时间步 t 开始的

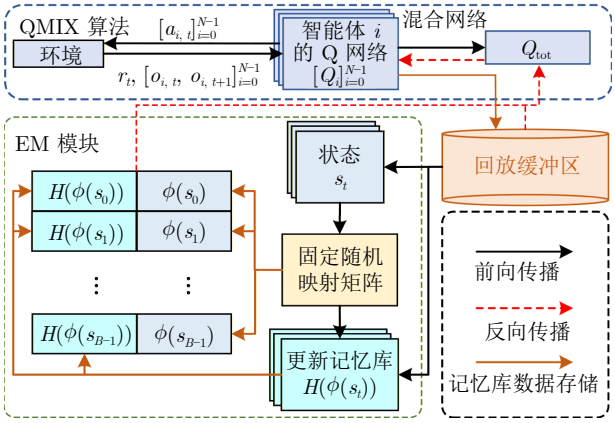


图 1　EM 模块与 QMIX 结合的架构

Fig. 1　Architecture of EM module combined with QMIX

累积回报; $\hat{s}_t$ 是记忆库中与当前状态 s_t 投影特征相似度最高的历史状态, $\phi(\hat{s}_t)$ 为其投影特征; δ 为投影特征相似度阈值 (一般设为 1×10^{-8} 的小实数). 若记忆库中无满足相似度阈值的历史状态, 则直接将当前状态回报 R_t 作为初始记忆回报存入.

目前的研究把 EM 与 QMIX 结合起来, 用 EM 的单步 TD 目标作为情境 MARL 算法损失的一部分. EM 模块为 QMIX 的训练提供每个状态的历史最优回报作为参照依据, 其 EM 单步 TD 目标为:

$$Q_{\text{EC}} = r_t + \gamma H(\phi(s_{t+1})) \tag{2}$$

其中, $H(\phi(s_{t+1}))$ 是时间步 $t+1$ 下全局状态 s_{t+1} 在记忆库中存储的最大回报. 情境 MARL 算法的总损失函数包含 QMIX 自身损失与 EM 损失两部分:

$$\begin{aligned} L_{\text{EC}} = & \underbrace{\mathrm{E}_{\boldsymbol{\tau},\,\boldsymbol{a}_t,\,r_t,\,\boldsymbol{\tau}'\in\mathcal{D}}\left[(y-Q_{\text{tot}}(\boldsymbol{\tau},\,\boldsymbol{a}_t;\,\theta))^2\right]}_{\text{QMIX 损失}} + \\ & \alpha\underbrace{\mathrm{E}_{\boldsymbol{\tau},\,\boldsymbol{a}_t,\,r_t,\,\boldsymbol{\tau}'\in\mathcal{D}}\left[(Q_{\text{EC}}-Q_{\text{tot}}(\boldsymbol{\tau},\,\boldsymbol{a}_t;\,\theta))^2\right]}_{\text{EM 损失}} = \\ & \frac{1}{T}\sum_{t=0}^{T-1}(y-Q_{\text{tot}}(\boldsymbol{\tau},\,\boldsymbol{a}_t;\,\theta))^2 + \\ & \alpha\frac{1}{T}\sum_{t=0}^{T-1}(Q_{\text{EC}}-Q_{\text{tot}}(\boldsymbol{\tau},\,\boldsymbol{a}_t;\,\theta))^2 \end{aligned} \tag{3}$$

其中, $\boldsymbol{\tau}=[\tau_{i,\,t}]_{i=0}^{N-1}$ 为所有智能体的联合观测–动作轨迹; $\mathcal{D}$ 为经验回放缓冲区; $Q_{\text{tot}}(\boldsymbol{\tau},\,\boldsymbol{a}_t;\,\theta)$ 是由参数 θ 构建的联合 Q 函数; α 为超参数;

$$y = r_t + \gamma\max_{\boldsymbol{a}_t\in\boldsymbol{\mathcal{A}}} Q_{\text{tot}}(\boldsymbol{\tau}',\,\boldsymbol{a}_t;\,\bar{\theta}) \tag{4}$$

为 QMIX 自身的单步 TD 目标, $\bar{\theta}$ 为目标网络参数, $\boldsymbol{\tau}'=[\tau_{i,\,t+1}]_{i=0}^{N-1}$ 为下一时间步样本轨迹. L_{EC} 的核心功能在于将记忆库中的历史最优回报纳入训练, 促进 QMIX 对高价值轨迹的学习进程.

2　指数图信息通信的情境多智能体强化学习

本节详细介绍 EMAGIC 的 3 个部分: 第 2.1 节提出基于单点指数图的通信拓扑; 第 2.2 节设计图信息通信机制; 第 2.3 节构建适用于独立学习范式的 EM 模块.

2.1　基于单点指数图的通信拓扑

首先定义 MAS 的通信拓扑及智能体编号.

定义 1 (智能体与通信拓扑). 设智能体集合为 $\mathcal{V}=\{0,\,1,\,\cdots,\,N-1\}$, 通信拓扑为有向图 $\{G_t\}_{t\geq0}$, 其中 $G_t=(\mathcal{V},\,\mathcal{E}_t)$, 边 $\mathcal{E}_t\subseteq\mathcal{V}\times\mathcal{V}$ 为通信链路.

本文采用稀疏型通信拓扑来降低通信开销, 借鉴分布式深度学习领域的指数图 (exponential graph) 设计思想[27], 将其引入 MARL 算法中来构建适配大规模环境的通信拓扑, 如图 2 所示.

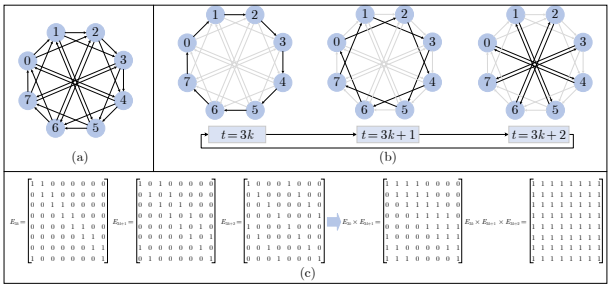


图 2　包含 8 个智能体的基于指数图的通信拓扑 ((a) 基于静态指数图的拓扑结构; (b) 基于单点指数图的通信拓扑结构; (c) 单点指数图单周期的布尔逻辑矩阵及矩阵相乘结果)

Fig. 2　Exponential graph-based communication topology with 8 agents ((a) Static exponential graph-based topology structure; (b) One-peer exponential graph-based communication topology structure; (c) Boolean logic matrix and matrix multiplication result of one-peer exponential graph in one cycle)

定义 2 (通信开销). 假设每条消息的传输产生固定的开销, 那么总的通信开销与总通信链路呈线性关系. 本文使用 $\mathrm{O}(\cdot)$ 来描述线性关系, 通信开销为 $\mathrm{O}(|\mathcal{E}_t|)$, 其中 $|\mathcal{E}_t|$ 是图中边的数量.

基于定义 2, 本文分析两种指数图的通信开销: 图 2(a) 的静态指数图中, 每个智能体与编号差为 $2^0,\,2^1,\,\cdots,\,2^{\lfloor\log_2(N-1)\rfloor}$ 跳的智能体建立通信链路 ($\lfloor\cdot\rfloor$ 表示向下取整), 每个时间步与 $\lceil\log_2 N\rceil$ 个邻居通信 ($\lceil\cdot\rceil$ 表示向上取整), 总开销为 $\mathrm{O}(N\lceil\log_2 N\rceil)$; 图 2(b) 的单点指数图将静态图分解为一系列单点图, 每个时间步每个智能体仅与 1 个邻居通信, 总成本降至 $\mathrm{O}(N)$. 虽然两者仿真时间复杂度一致, 但

单点指数图实际通信开销更低, 故采用其作为MAS的通信拓扑, 其通信矩阵定义如下.

定义 3 (通信拓扑矩阵). 时间步 t 的通信拓扑用布尔矩阵 $E_t = [e_{ij,\,t}]^{N\times N}$, $e_{ij,\,t}\in\{0,\,1\}$ 描述, 矩阵 E_t 中每个元素 $e_{ij,\,t}$ 的计算公式为:

$$e_{ij,\ t} = \begin{cases} 1, & i = j \\ 1, & \mathrm{mod}(j-i,\ N) = 2^{\mathrm{mod}(t,\,k)} \\ 0, & \text{其他} \end{cases} \tag{5}$$

其中, $(i,\,j)\in\mathcal{V}\times\mathcal{V}$ 为任意智能体的编号对, 消息从 $i\to j$; $k=\lceil\log_2 N\rceil$ 为通信周期; $\mathrm{mod}(t,\,k)$ 为 t 对 k 取余; $2^{\mathrm{mod}(t,\,k)}$ 为时间步 t 的通信跳数. 矩阵乘法遵循逻辑运算规则: 对任意智能体 i 和 j, 有

$$(E_{t_1}\times E_{t_2})_{i,\,j} = \vee_{x\in\mathcal{V}}(e_{ix,\,t_1}\wedge e_{xj,\,t_2}) \tag{6}$$

其中, $\vee$ 为逻辑或; $\wedge$ 为逻辑与. 式 (6) 的值为1表示智能体 j 能通过任意智能体 x 经时间步 $t_1\to t_2$, 借助通信链路 $e_{ix,\,t_1}=1$ 和 $e_{xj,\,t_2}=1$ 获取包含智能体 i 信息的消息特征. 因此, 本文将满足式 (6) 的情况定义为智能体 $i\to j$ 的消息传递.

单点指数图的权重矩阵序列乘积已被证明能实现周期精确平均[27], 但通信拓扑中权重矩阵转为布尔逻辑后, 原有理论不再直接适用. 本文对通信拓扑的消息传递能力进行理论分析.

定理 1 (单周期消息全局传递). 对于单点指数图, 任取起始时间步 $t_0\geq 0$, 在周期 $[t_0,\,t_0+k-1]$ 内, 布尔通信矩阵的乘积满足

$$\prod_{t=t_0}^{t_0+k-1} E_t = E_{t_0}\times E_{t_0+1}\times\cdots\times E_{t_0+k-1} = \mathbf{1}\mathbf{1}^{\mathrm{T}} \tag{7}$$

其中, $\mathbf{1}$ 为全1向量; $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ 为全1矩阵. 称该通信拓扑在单周期内实现消息全局传递.

证明. 由单点指数图的周期性原理可知, 任意时间步 t' 的通信跳数满足 $h_{t'+k}=h_{t'}$, 其中 $h_{t'}=2^{\mathrm{mod}(t',\,k)}$, 其取值集合为 $\{2^0,\,2^1,\,\cdots,\,2^{k-1}\}$. 对任意起始时间步 t_0, 周期 $[t_0,\,t_0+k-1]$ 内的跳数序列为 $\{h_{t_0},\,h_{t_0+1},\,\cdots,\,h_{t_0+k-1}\}$, 该序列为 $\{2^0,\,2^1,\,\cdots,\,2^{k-1}\}$ 的循环排列. 定义智能体 i 发出的消息在周期内经过的索引序列为 $\{i,\,x_{t_0},\,x_{t_0+1},\,\cdots,\,x_{t_0+k-1}\}$, 其中 $e_{ix_{t_0},\,t_0}=1$, 且对任意 $m\in[t_0,\,t_0+k-1]$, x_m 与 x_{m+1} 在时间步 m 存在通信链路, x_m 的消息包含自身获取和向编号差为 h_m 的智能体传递两条路径. 引入系数 $\{b_{t_0},\,\cdots,\,b_{t_0+k-1}\}\in\{0,\,1\}$ 控制路径选择: $b_m=1$ 时, 消息从 x_m 向 x_m+h_m 跳传输, 即 $e_{x_m(x_m+h_m),\,m}=1$; $b_m=0$ 时, x_m 获取自身消息, 即 $e_{x_mx_m,\,m}=1$. 单周期内智能体索引递推关系为

$$x_{m+1} = \mathrm{mod}(x_m + b_m h_m,\ N) \tag{8}$$

以 $N=8$, 即 $k=3$ 为例, 图 2(c) 展示了以 $t_0=3k$ 为起点的示例. 将式 (8) 从 i 递推至 x_{t_0+k-1} 得

$$x_{t_0+k-1} = \mathrm{mod}\left(i+\sum_{m=t_0}^{t_0+k-1} b_m h_m,\ N\right) = \mathrm{mod}\left(i+\sum_{m=t_0}^{t_0+k-1} b_m 2^{\mathrm{mod}(m,\,k)},\ N\right) \tag{9}$$

由单点指数图的周期性可知, $\mathrm{mod}(m,\,k)$ 在 $m\in[t_0,\,t_0+k-1]$ 内遍历 $\{0,\,1,\,\cdots,\,k-1\}$, 求和项可改写为

$$b_{t_0}2^{a_0} + b_{t_0+1}2^{a_1} + \cdots + b_{t_0+k-1}2^{a_{k-1}} \tag{10}$$

其中, $\{a_0,\,\cdots,\,a_{k-1}\}$ 为 $\{0,\,1,\,\cdots,\,k-1\}$ 的循环排列. 根据二进制数特性, 式 (10) 可生成 $[0,\,2^k-1]$ 内所有整数. 由定义 3 中 $k=\lceil\log_2 N\rceil$ 可知 $N\leq 2^k$, 即 $N-1\leq 2^k-1$; 故对于任意智能体 j, 均有 $\mathrm{mod}(j-i,\,N)\in[0,\,2^k-1]$, 存在系数 $\{b_{t_0},\,\cdots,b_{t_0+k-1}\}$ 使得

$$\sum_{m=t_0}^{t_0+k-1} b_m 2^{\mathrm{mod}(m,\,k)} = \mathrm{mod}(j-i,\ N) \tag{11}$$

代入式 (9) 并结合 $i,\,j\in\{0,\,\cdots,\,N-1\}$ 得

$$\begin{aligned} x_{t_0+k-1} &= \mathrm{mod}(i+\mathrm{mod}(j-i,\ N),\ N) = \\ &\mathrm{mod}(\mathrm{mod}(i,\ N)+\mathrm{mod}(j-i,\ N),\ N) = \\ &\mathrm{mod}(i+j-i,\ N) = j \end{aligned} \tag{12}$$

因此, 存在索引序列 $\{i,\,x_{t_0},\,\cdots,\,x_{t_0+k-1}=j\}$, 使消息能在单周期 $[t_0,\,t_0+k-1]$ 内从智能体 i 传递到智能体 j, 即 $e_{x_{t_0+k-2}j,\,t_0+k-1}=1$, 任意智能体对 $(i,\,j)$ 可实现消息传递, 故存在中间智能体序列使下式成立:

$$e_{ix_{t_0},\,t_0}\wedge e_{x_{t_0}x_{t_0+1},\,t_0+1}\wedge\cdots\wedge e_{x_{t_0+k-2}j,\,t_0+k-1} = 1 \tag{13}$$

进一步, 乘积矩阵的任意元素满足:

$$\left(\prod_{t=t_0}^{t_0+k-1} E_t\right)_{ij} = (E_{t_0}\times E_{t_0+1}\times\cdots\times E_{t_0+k-1})_{ij} = \vee_{x\in\mathcal{V}}\left(e_{ix_{t_0},\,t_0}\wedge\cdots\wedge e_{x_{t_0+k-2}j,\,t_0+k-1}\right) = 1 \tag{14}$$

因 $(i,\,j)\in\mathcal{V}\times\mathcal{V}$ 是布尔通信矩阵 E_t 中的任意元素, 故 $\prod_{t=t_0}^{t_0+k-1}E_t = \mathbf{1}\mathbf{1}^{\mathrm{T}}$. □

基于上述理论分析, 得出如下结论: 1) 任意两个不同智能体之间的消息传播至多需要 $k-1=$

$\lceil \log_2 N \rceil - 1$ 个时间步, 确保部分可观测环境下信息的及时获取; 2) 基于单点指数图的通信拓扑的通信开销为 O(N), 通信开销与智能体数量呈线性关系, 适配大规模 MAS. 综上, 基于单点指数图的通信拓扑具有很好的平衡通信开销与消息传递的能力.

2.2　图信息通信

智能体间通信链路搭建完毕后, 需要将有价值的信息在各智能体之间发送, 实现高效的协同决策. 本文在每个智能体的个体 Q 网络上构建图信息通信机制, 如图 3 所示. 具体而言, 对任意智能体 i, 采用结合 GRU 的 DQN 来处理时序观测和历史决策数据, 挖掘出环境变化及与智能体交互的时间序列特点. 在时间步 t, 智能体 i 的个体 Q 网络输入为当前观测 $o_{i,\,t}$ 与上一时间步动作 $a_{i,\,t-1}$. 首先使用多层感知机 (multi-layer perceptron, MLP) 对其做初步特征编码, 再将其送入 GRU, 并结合 GRU 上一时间步 $t-1$ 的隐藏状态 $hs_{i,\,t-1}$ 生成当前时间步隐藏状态 $hs_{i,\,t}$, 整合从初始时间步开始一直累积到时间步 t 的全部历史时序信息, 避免出现因观测瞬时性而导致决策短视的问题. GRU 的输出 $hs_{i,\,t}$ 以及此时自身要发送的消息 $m_{i,\,t}$ 组成向量 $[hs_{i,\,t},\ m_{i,\,t}]$, 输入到 MLP 中获取个体 Q 值 $Q_i(\tau_{i,\,t},\ \cdot)$; 按照 ϵ-贪婪策略 π_i 根据 Q_i 选取当前动作 $a_{i,\,t}$, 并利用 Q_i 实现策略更新.

为实现交互信息的有效传递, 本文构建适配于多智能体场景的消息编码模块. 考虑到大规模 MAS 的计算约束情况, 本文不采用计算较为复杂的 GNN 类方法, 而是选择适合于处理序列问题的 GRU, GRU 是一种常见的能够用于解决大规模 MAS 的消息编码方案[11, 20]. 假定智能体 n 与 i 之间在时间步 $t-1$ 有通信链路 $e_{ni,\,t-1}=1$, 则智能体 i 可以在该时间步接收到 n 发出的消息 $m_{n,\,t-1}$. 智能体 i 在时间步 t 生成消息 $m_{i,\,t}$ 的流程为: 首先拼接个体 Q 网络 GRU 的隐藏状态 $hs_{i,\,t}$ 和消息 $m_{n,\,t-1}$ 得到 $[hs_{i,\,t},\ m_{n,\,t-1}]$, 通过线性层进行编码; 然后将编码后的结果输入消息模块的 GRU 中, 并把其自身上一时间步的消息 $m_{i,\,t-1}$ 也输入 GRU 中, 得到该时间步要向其他智能体传递的消息 $m_{i,\,t}$; 最后, 消息模块的输出还要反馈给个体 Q 网络 GRU 中, 以便为智能体 i 决策提供交互信息.

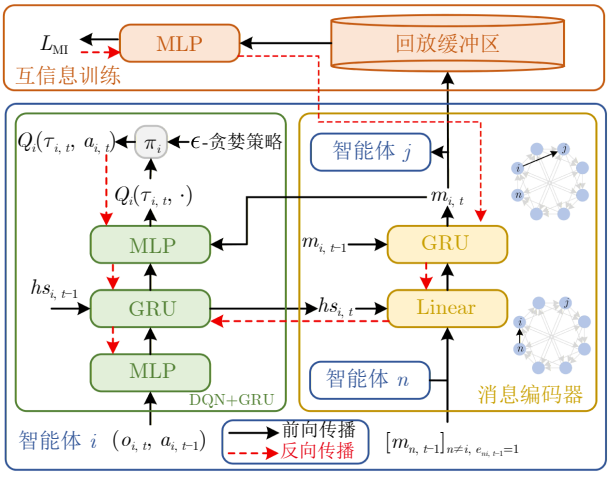


图 3　图信息通信架构

Fig. 3　Architecture of graph information communication

虽然消息中已经囊括了智能体的历史观测, 但由于受部分可观测性限制, 该消息只由发出方的局部观测信息组成, 接收方可能因观测偏差误解消息, 影响到 MAS 的协同效率, 而解决该问题的方法就是让智能体学会具备互理解性的消息, 从而实现更好的协作. 本文借鉴 Lo 等[26] 的思想: 每个智能体生成的消息本质上是对环境全局状态的不完整视图编码, 如果消息之间的相关性较高, 则表示不同的智能体都表达了对同一个全局状态的一致描述; 这种情况下可以将多个分散的局部视图进行互补整合, 从而还原完整全局状态. 本文通过最大化消息间互信息的下界提升相关性, 采用最先进的互信息下界估计方法平滑互信息下界估计 (smoothed mutual information lower-bound estimator, SMILE)[28], 建立同一时间步下任意两个不同智能体消息间的相关性. EMAGIC 是首次将 SMILE 引入到 MARL 中的应用研究, 消息对采样流程如下:

- 随机选取时间步 $t \in \{0,\ 1,\ \cdots,\ T-1\}$, 从 N 个智能体中随机选取任意智能体 $i \in \{0,\ 1,\ \cdots,\ N-1\}$ 作为基准;
- 获取编号不同于智能体 i 的智能体 i^c: 随机生成整数 $n \in [0,\ N-2]$, 按如下规则获取 i^c 索引

$$i^c = \begin{cases} n, & n < i \\ n+1, & n \geq i \end{cases} \tag{15}$$

此时 $i^c \in \{0,\ 1,\ \cdots,\ N-1\}$ 且 $i^c \neq i$;

- 按智能体编号 i、i^c 及时间步 t 从回放缓冲区内采集消息对 $(m_{i,\,t},\ m_{i^c,\,t})$, 目标是最大化该消息对的互信息.

主流互信息下界估计方法 (信息噪声对比估计 (information noise-contrastive estimation, InforNCE)[29]、Nemenman-Wang-Jordan (NWJ)[30] 以及互信息神经估计 (mutual information neural estimation, MINE)[31]) 采用对比学习的逻辑思路, 分别生成正/负样本, 在正样本中求解最大值, 使目标变量信息之间关联度最大, 即获得最有效的信息耦合; 同时在负样本中求解最小值, 将假相关性降到

最低, 以此达到逼近互信息下界的目的. 基于该思路, 本文定义消息传递过程的正/负样本:

定义 4 (正/负样本). 同一时间步 t 下, 不同智能体的消息对 $(m_{i,\,t},\ m_{i^c,\,t})$ 定义为正样本. 正样本可视为同一全局状态 s_t 的不完整编码, 其在回放缓冲区中的采样服从联合分布 P. 超出时间步 t 一个通信周期以上的任意智能体消息 $m_{i,\,t'}$ 与 $m_{i,\,t}$ 组成的消息对 $(m_{i,\,t},\ m_{i,\,t'})$ 定义为负样本, 其中 $t' \in [0,\ t-k) \cup (t+k,\ T-1]$. 假设时间间隔超过通信周期后, 环境从 s_t 变为特征差异大的新状态 $s_{t'}$, 因此, 负样本为差异显著的不同全局状态编码, 从回放缓冲区采样服从边缘乘积分布 $P' = P(m_{i,\,t}) \times P(m_{i,\,t'})$.

基于上述正/负样本, 结合 SMILE 构建多智能体消息的互信息下界估计公式:

$$I_{\text{SMILE}} = \underbrace{\mathrm{E}_{(m_{i,\,t},\,m_{i^c,\,t})\sim P}\left[F_\psi(m_{i,\,t},\ m_{i^c,\,t})\right]}_{\text{正样本得分}} - \underbrace{\ln\left(\mathrm{E}_{(m_{i,\,t},\,m_{i,\,t'})\sim P'}\left[\exp\left(F_\psi^\xi(m_{i,\,t},\ m_{i,\,t'})\right)\right]\right)}_{\text{负样本得分}} \tag{16}$$

其中, $F_\psi : \mathbf{R}^D \times \mathbf{R}^D \to \mathbf{R}$ 是 ψ 参数化的评分网络, 具体采用 1 个 3 层 MLP f_ψ 编码消息获取特征, 再通过点积获取样本得分:

$$F_\psi(m_{i,\,t},\ m_{i^c,\,t}) = f_\psi(m_{i,\,t}) f_\psi^{\mathrm{T}}(m_{i^c,\,t}) \tag{17}$$

D 为消息特征维度; $\mathrm{E}_{(m_{i,\,t},\,m_{i^c,\,t})\sim P}[\cdot]$ 量化正样本平均关联度; $\mathrm{E}_{(m_{i,\,t},\,m_{i,\,t'})\sim P'}[\cdot]$ 抑制负样本虚假相关性. 为解决传统方法在真实互信息增大时方差呈指数级增长的问题, SMILE 使用 clip$(\cdot)$ 函数, 用于裁剪评分网络输出:

$$F_\psi^\xi(m_{i,\,t},\ m_{i,\,t'}) = \mathrm{clip}(F_\psi(m_{i,\,t},\ m_{i,\,t'}),\ -\xi,\ \xi) \tag{18}$$

其中, ξ 为裁剪阈值, 用于限制异常负样本的影响, 避免该异常值主导损失函数更新. 本文参考 Song 等[28] 的实验结果, 取 $\xi = 5$ 作为裁剪阈值.

本文的目标是最大化消息间互信息, 故将式 (16) 的负值作为互信息损失函数 L_{MI}, 通过梯度下降最大化真实互信息下界. 结合 MARL 批量训练特性, L_{MI} 具体形式为

$$L_{\text{MI}} = -I_{\text{SMILE}} = -\frac{1}{N}\sum_{i=0}^{N-1} F_\psi(m_{i,\,t},\ m_{i^c,\,t}) + \ln\left(\frac{1}{N(T-2k)}\sum_{i=0}^{N-1}\sum_{t'}\exp\left(\mathrm{clip}\left(F_\psi(m_{i,\,t},\ m_{i,\,t'})\right)\right)\right) \tag{19}$$

其中, $T-2k$ 为单条轨迹中符合负样本时间步约束的总数. 式 (19) 右边两项分别对应式 (16) 的正/负样本得分.

2.3 独立情境记忆模块

EM 与 QMIX 结合虽性能优异, 但仍存在局限: QMIX 需通过混合网络将个体 Q 值聚合为联合 Q 值, 该过程依赖共享奖励更新策略与记忆库, 导致 QMIX 和传统情境多智能体强化学习 (episodic multi-agent reinforcement learning, EMARL) 仅适用于共享奖励场景. 已有研究证明在一些智能体间不共享奖励的多智能体环境中, 独立学习范式优于 CTDE 方法[32]. 因此, 本文构建情景记忆与 IDQN 相结合的方法, 提高 EM 的适用性. 然而, 若为每个智能体独立构建专属记忆库, 会产生较大的计算资源消耗, 难以适配大规模 MAS. 为此, 本文提出一种简单有效的方法: 非共享奖励场景下建立全局状态与平均回报的对应关系, 所有智能体共享一个记忆库, 避免单独构建记忆库的高计算需求. 尽管该场景中智能体间不共享奖励, 但 MARL 的优化目标仍是最大化 MAS 的全局回报[33]. 记忆库中每个状态索引 $\phi(s_t)$ 存储对应全局状态 s_t 的平均回报 $\bar{R}_t = \frac{1}{N}\sum_{i=0}^{N-1}\sum_{t'=0}^{T-t-1}\gamma^{t'} r_{i,\,t+t'}$, 记忆库更新仍遵循式 (1), 只是将 R_t 替换为 $\bar{R}_t$. 进一步, 本文提出适用于非共享奖励场景的 EM 单步 TD 目标 Q_{EC}:

$$Q_{\text{EC}} = \frac{1}{N}\sum_{i=0}^{N-1} r_{i,\,t} + \gamma H(\phi(s_{t+1})) \tag{20}$$

基于式 (20) 和 IDQN 自身损失函数, 推导非共享奖励场景下 EMAGIC 的 EM 模块损失函数:

$$L_{\text{EC}} = \underbrace{\mathrm{E}_{\mathcal{D}}\left[(y_i - Q_i(\tau_{i,\,t},\ a_{i,\,t};\ \theta_i))^2\right]}_{\text{IDQN 损失}} + \underbrace{\alpha\,\mathrm{E}_{\mathcal{D}}\left[\left(Q_{\text{EC}} - \frac{1}{N}\sum_{i=0}^{N-1} Q_i(\tau_{i,\,t},\ a_{i,\,t};\ \theta_i)\right)^2\right]}_{\text{EM 损失}} = \frac{1}{TN}\sum_{i=0}^{N-1}\sum_{t=0}^{T-1}(y_i - Q_i(\tau_{i,\,t},\ a_{i,\,t};\ \theta_i))^2 + \alpha\frac{1}{T}\sum_{t=0}^{T-1}\left(Q_{\text{EC}} - \frac{1}{N}\sum_{i=0}^{N-1} Q_i(\tau_{i,\,t},\ a_{i,\,t};\ \theta_i)\right)^2 \tag{21}$$

其中,

$$y_i = r_{i,\,t} + \gamma \max_{a_{i,\,t}\in\mathcal{A}} Q_i(\tau_{i,\,t+1},\ a_{i,\,t};\ \bar{\theta}_i) \tag{22}$$

为 IDQN 的标准单步 TD 目标. 式 (21) 通过最小化 Q_{EC} 与个体 Q 值均值的误差, 结合 IDQN 自身损失, 在利用个体奖励更新策略的同时添加全局优化约束, 平衡大规模环境下的个体学习与全局协同. EMAGIC 是首个讨论 EM 与独立学习范式结合的研究. 根据环境奖励设置动态选择 L_{EC} 计算方式: 共享奖励场景采用式 (3), 非共享奖励场景采用式 (21). 结合式 (19), EMAGIC 的总损失函数为

$$L = L_{\mathrm{EC}} + \beta L_{\mathrm{MI}} \tag{23}$$

其中, β 为控制互信息损失权重的超参数; L_{EC} 用于提升样本利用效率; L_{MI} 用于提升 MAS 的协同性能. EMAGIC 具体实现流程如算法 1 所示.

算法 1. EMAGIC

1. 初始化 Q 网络、目标网络及消息编码网络;
2. 初始化回放缓冲区和记忆库;
3. 初始化状态投影函数 $\phi(\cdot)$;
4. 初始化环境;
5. **for** $n = 0,\ 1,\ \cdots,$ 训练步数 **do**
6. **for** $t = 0,\ 1,\ \cdots,\ T-1$ **do**
7. 收集当前观测和全局状态, 按式 (5) 构建通信拓扑;
8. **for** 智能体 $i = 0,\ 1,\ \cdots,\ N-1$ **do**
9. 传递消息并基于 ϵ-贪婪策略选择动作;
10. **end for**
11. 执行联合动作, 获取各智能体局部/共享奖励、下一时间步观测值及全局状态;
12. **end for**
13. 计算本回合累积回报, 按式 (1) 更新记忆库;
14. 将本回合观测、动作、消息、奖励及全局状态存入回放缓冲区;
15. **if** $n \geq$ 批次大小 **then**
16. 从回放缓冲区随机抽取大小为 B 的样本;
17. **if** 基线算法为 QMIX **then**
18. 按式 (4)、式 (2)、式 (3) 计算 QMIX 单步 TD 目标、EM 单步 TD 目标及 EM 损失;
19. **else**
20. 按式 (22)、式 (20)、式 (21) 计算 IDQN 单步 TD 目标、EM 单步 TD 目标及 EM 损失;
21. **end if**
22. 随机选时间步 t, 按式 (15) 构建正样本对, 选超过当前时间步一个通信周期的消息作为负样本;
23. 按式 (19) 计算互信息损失;
24. 按式 (23) 计算 EMAGIC 总损失;
25. 更新 Q 网络及消息编码网络参数;
26. **if** 满足目标网络更新间隔 **then**
27. 更新目标网络参数.
28. **end if**
29. **end if**
30. **end for**

3　实验结果与分析

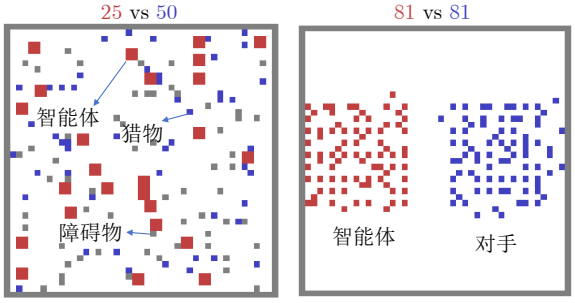
本文首先介绍用于测试 EMAGIC 性能的多智能体环境任务, 其次阐明环境、EMAGIC 以及对比基线的参数设置, 然后展示 EMAGIC 与基线的对比实验, 最后给出 EMAGIC 各组成部分的消融实验.

3.1　多智能体环境任务

本文采用两种大规模多智能体环境验证 EMAGIC 的性能, 环境的基础配置分别遵循 Zheng 等[32] 为 MAgent 环境提供的标准设置以及 Leroy 等[34] 为 IMP 环境提供的标准设置.

3.1.1　MAgent

本文采用 MAgent 环境的两个任务场景来测试 EMAGIC 性能, 分别为对抗性追击 (adversarial pursuit, AdvPursuit) 和战斗 (Battle), 如图 4 所示. AdvPursuit 的默认设置为 25 个捕食者 (智能体) 同时去追捕 50 个猎物, 智能体的任务是要躲避障碍, 把每只猎物都标记好才算完成任务; 而猎物的任务是躲避被标记. 由于智能体跑得慢而且体积大, 独自很难标记猎物, 要想做到逐个标记猎物就需要智能体协同围堵, 使猎物无法逃跑. Battle 任务中红方智能体和蓝方智能体数量一致, 回合结束时红方存活数大于蓝方则获胜. MAgent 环境采用智能体间不共享奖励的设置, 智能体的奖励仅与自身有关, 与其他智能体无关, 共享奖励有可能造成信用分配问题, 这也是本文要设计独立 EM 模块的原因. 在 MAgent 环境中, EMAGIC 控制红色的智能体, 蓝色的智能体由 IDQN 预训练的策略控制.



(a) AdvPursuit 任务　(b) Battle 任务
(a) AdvPursuit task　(b) Battle task

图 4　MAgent 环境下两个任务的示意图
Fig. 4　Schematic diagrams of two tasks in the MAgent environment

3.1.2 IMP

IMP 的环境任务如图 5 所示, 该环境中系统故障风险被定义为组件损坏情况概率分布的函数, 组件在每个时间步 t 可执行检查、维修及不操作 3 种动作, 目标是通过平衡系统故障风险奖励 $r_{f,\,t}$、检查奖励 r_{ins} 及维修奖励 r_{rep} (均为负奖励), 实现组件维护成本与系统失效风险的平衡. 当组件未被检查或维修时, 其损坏概率按劣化过程演变, 检查后更新损坏概率时会考虑检查信息, 维修后则损坏概率重置为初始分布, 图 5 展示了时间步 t 具有相同损坏概率的 3 个组件. 本文采用 IMP 环境中的 k-out-of-n (kn)、Correlated k-out-of-n (Ckn) 及Off-shore wind farm (Owf) 3 个任务作为 EMAGIC 的测试环境: kn 任务由 n_comp 个组件构成, 系统正常运行需要环境中至少有 k_comp 个组件处于正常状态, 当 n_comp-k_comp+1 个组件损坏时系统故障, 每个智能体负责一个组件; Ckn 任务在 kn 任务基础上, 组件初始损伤分布存在关联, 这一关联由相关性因子调控, 相关性因子是表征组件间初始损伤关联程度的全局共享参数, 智能体需结合该因子推断关联组件风险并协作, kn 和 Ckn 两个任务中 n_comp 取值为 50 或 100, k_comp 取值为 48 或 95[26]; Owf 任务模拟海上风电场维护场景, 每个风电机包含可查修的顶部、中部组件及不可查修的泥线组件, 每个风机分配 2 个智能体分别负责顶部和中部组件. IMP 环境下的 3 个任务均通过执行检查、维修或不操作动作来平衡维护成本与系统失效风险.

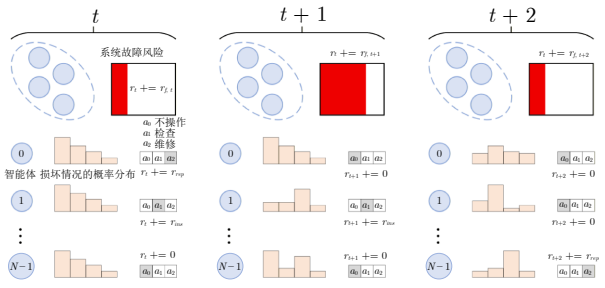


图 5 IMP 环境任务示意图

Fig. 5 Schematic diagram of the IMP environment tasks

3.2 实验设置

EMAGIC 的基线算法选择严格匹配环境奖励特性: 针对 MAgent 环境的智能体间不共享奖励设置, 采用 IDQN 作为基线; 针对 IMP 环境的共享奖励设置, 采用 QMIX 作为基线. 所有算法均基于统一代码库 EPyMARL[35] 实现, 采用相同的超参数进行配置, 本文仅展示 EMAGIC 新增超参数, 如表 1 所示. 为降低随机误差对结果的影响, 每种算法的回报曲线均通过多组不同随机种子的重复实验取平均得到.

表 1 EMAGIC 的重要超参数

Table 1 Important hyperparameters of EMAGIC

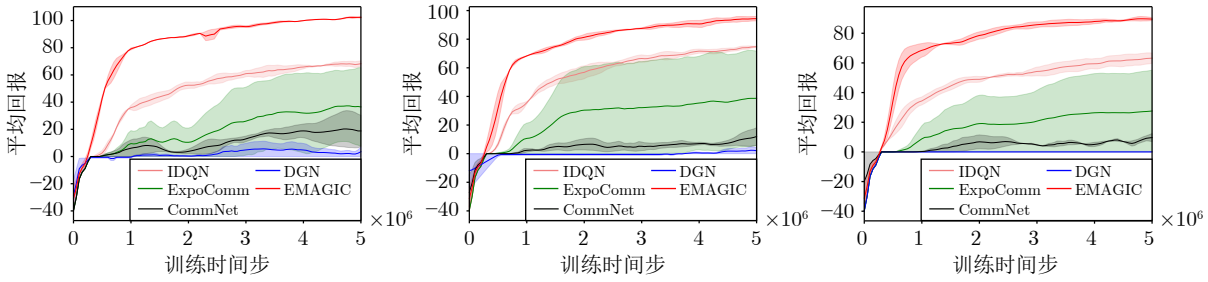
消息特征维度 D	α	β	记忆库大小 B	嵌入特征维度 d
64	0.1	0.1	1×10^6	4

3.3 对比实验

由于 MARL 以获得最高累积回报作为目标, 所以本文将所有智能体单回合的平均回报 $\bar{R}=\frac{1}{N}\sum_{i=0}^{N-1}\sum_{t=0}^{T-1}r_{i,\,t}$ 和 $\bar{R}=\sum_{t=0}^{T-1}r_t$ 分别作为衡量每种算法在 MAgent 和 IMP 环境中性能表现的评价指标. 本文采用先进的基于通信的 MARL 算法 (ExpoComm、CommNet 和 DGN)、基于值函数的 MARL 算法 (QMIX、VDN 和 IDQN) 及 EMARL 算法 (EMC) 作为 EMAGIC 的对比算法. 其中, ExpoComm 和 CommNet 分别基于静态指数图和全连接图构建通信拓扑, 因此, 这两种算法可适用于 MAgent 和 IMP 环境中的所有任务场景; 而 DGN 基于距离的图结构构建通信拓扑, 仅适用于具有空间信息的 MAgent 环境. 此外, 因 MAgent 环境与 IMP 环境的奖励设置不同, VDN 与 QMIX 适用于 IMP 环境, IDQN 适用于 MAgent 环境. 本文的所有实验结果图中, 实线为重复实验下的平均回报平滑曲线, 阴影区域则代表实验结果的波动范围.

3.3.1 MAgent 环境下实验结果与分析

为验证 EMAGIC 在大规模 MAS 中的性能表现, 本文在 MAgent 环境的不同任务及不同智能体规模下对其展开测试, 并将 EMAGIC 的平均回报曲线与各对比算法进行直观对比, 实验结果如图 6 和图 7 所示. 图 6(a) ~ 图 6(c) 展示了几种算法在 AdvPursuit 任务中的性能, 随着智能体−猎物数量从 25 与 50 增至 61 与 122, 所有算法的回报值均出现不同程度下降, 但 EMAGIC 始终保持最高回报水平. 即使在 61 与 122 的最大规模下, 仍能稳定收敛至 90 左右的平均回报值, 且相较于 25 与 50 的规模性能无明显衰减, 意味着 EMAGIC 具有扩展至大规模 MAS 的潜力. 同时, EMAGIC 的平均回报曲线上升速度明显快于其他算法, 曲线方差更小. 尽管 IDQN 在 3 种不同智能体规模下的性能表现仅次于 EMAGIC, 但其收敛后的平均回报值较 EMAGIC 低约 30, 二者存在显著性能差距. ExpoComm 虽展现出良好的性能, 但回报值波动极大, 甚至在部分实验中长期围绕 0 波动. DGN 与 CommNet 二者表现均不理想, 特别是 DGN, 多次实验的



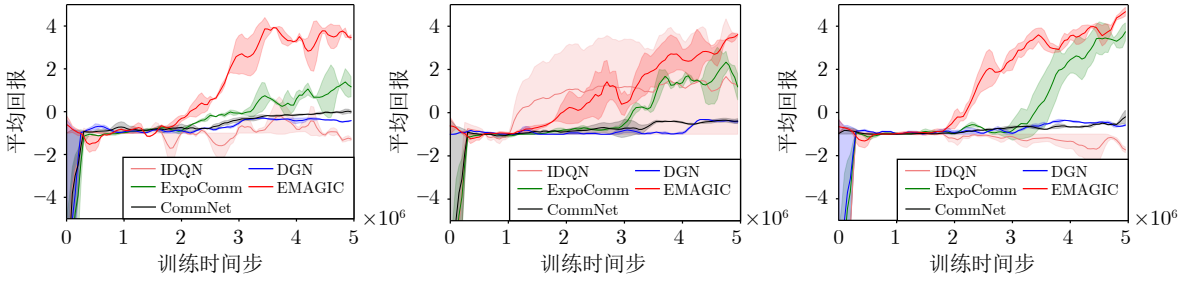
(a) 25 个智能体与 50 个猎物
(a) 25 agents and 50 prey

(b) 45 个智能体与 90 个猎物
(b) 45 agents and 90 prey

(c) 61 个智能体与 122 个猎物
(c) 61 agents and 122 prey

图 6　MAgent 环境下 AdvPursuit 任务中的对比实验

Fig. 6　Comparative experiments on AdvPursuit tasks in the MAgent environment



(a) 20 个智能体与 20 个对手
(a) 20 agents and 20 opponents

(b) 42 个智能体与 42 个对手
(b) 42 agents and 42 opponents

(c) 64 个智能体与 64 个对手
(c) 64 agents and 64 opponents

图 7　MAgent 环境下 Battle 任务中的对比实验

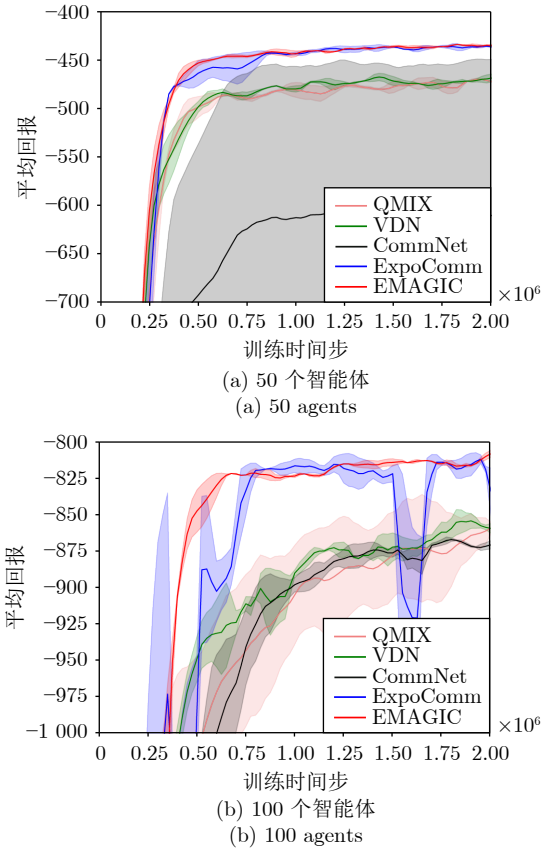
Fig. 7　Comparative experiments on Battle tasks in the MAgent environment

平均回报值几乎都在 0 附近. 本文认为这是二者的通信机制造成的: DGN 只能与最近的 3 个智能体通信, 消息难以传播至远距离智能体; 而 CommNet 与所有智能体进行通信, 虽然可以在一定程度上促进智能体间协同, 但是无视通信对象进行盲目的通信可能会产生大量冗余信息, 阻碍算法收敛, 这也是二者的表现不如 IDQN 和 ExpoComm 的原因. 相比之下, EMAGIC 中每个智能体每个时间步仅与 1 个智能体通信, 且消息经单个周期即可传递给所有智能体, 避免了通信受限和信息冗余的问题. 综上, EMAGIC 在 AdvPursuit 任务中的收敛性能以及平均回报值都优于其余算法.

Battle 任务中回报曲线如图 7(a) ~ 图 7(c) 所示. 在所有智能体与对手数量的规模下, EMAGIC 均能取得最高的平均回报; 尤为关键的是, EMAGIC 在 64 个智能体与 64 个对手规模下的最终回报值, 反而高于 20 个智能体与 20 个对手的规模, 直接验证其在大规模 MAS 中的应用潜力. ExpoComm 虽能达到较高回报, 但其需要静态指数图通信的方式, 这种设置在实际大规模 MAS中难以满足. 虽然在 42 个智能体规模下 IDQN 算法在前期曲线的上升速度方面优于 EMAGIC, 但 EMAGIC 最终仍能收敛至最优性能, 且 IDQN 的波动很大, 甚至在部分实验中回报值一直在 0 附近波动, 相比之下, EMAGIC 的方差就小得多, 说明 EMAGIC 性能稳定, 不像传统的 EMARL 依赖于投影矩阵的初始化好坏. 与 AdvPursuit 类似, DGN 与 CommNet 仍受限于二者的通信机制, 在 Battle 任务中性能较差. 经上述实验分析, EMAGIC 在 Battle 任务中仍展现最优性能, 这是由于 EMAGIC 的通信机制不仅能获取高值回报, 而且 EM 模块还能对高价值样本进行充分利用, 二者相辅相成, 共同实现 EMAGIC 优异的性能表现.

3.3.2　IMP 环境下实验结果与分析

为验证 EMAGIC 在共享奖励设置下的性能表现, 本文还在 IMP 环境下对其进行测试, 实验结果如图 8 ~ 图 10 所示. Owf 任务中的回报曲线如图 8(a) 和图 8(b) 所示, 无论智能体数量为 50 还是 100, EMAGIC 均能收敛至最优回报值. 特别是在 100 个智能体的更大规模场景中, EMAGIC 的平均回报值明显优于 QMIX、VDN 与 CommNet. ExpoComm 虽能收敛至与 EMAGIC 相近的回报值, 但收敛后回报波动大, 且曲线上升速度慢, 收敛性能不及 EMAGIC. Ckn 任务中的回报曲线如图 9(a) 和图 9(b) 所示, 在 50 个智能体的规模下, EMAGIC 展现出的性能最优. 在 100 个智能体的场

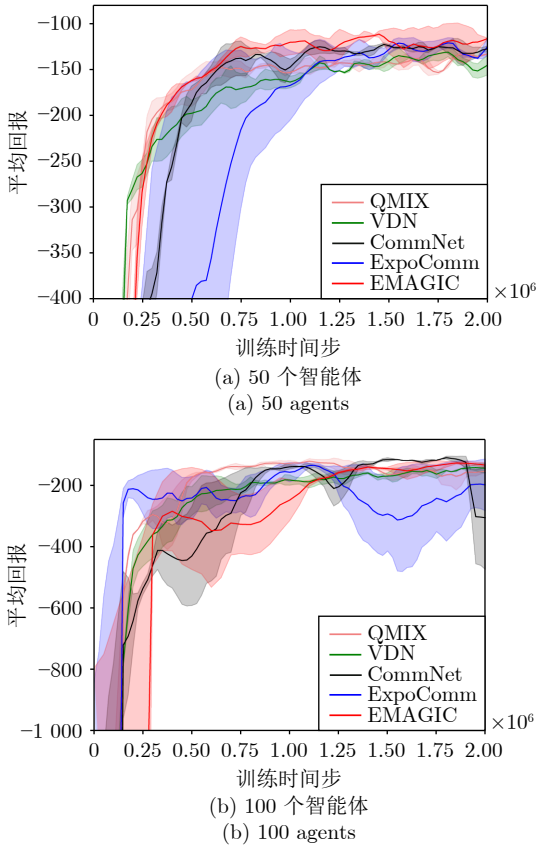


(a) 50 个智能体
(a) 50 agents

(b) 100 个智能体
(b) 100 agents

图 8 IMP 环境下 Owf 任务中的对比实验

Fig. 8 Comparative experiments on Owf tasks in the IMP environment



(a) 50 个智能体
(a) 50 agents

(b) 100 个智能体
(b) 100 agents

图 9 IMP 环境下 Ckn 任务中的对比实验

Fig. 9 Comparative experiments on Ckn tasks in the IMP environment

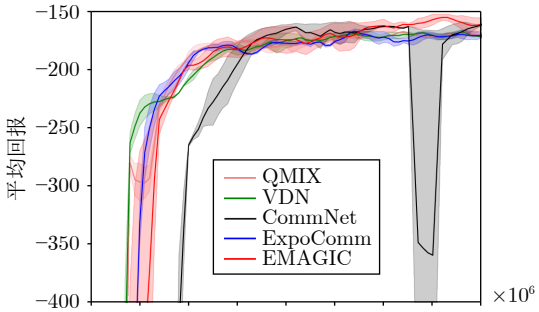
景下, CommNet 虽能更快地收敛至较高回报值, 但存在明显缺陷: 在时间步 1.9×10^6 处出现剧烈波动, 导致其最终平均回报值不如 EMAGIC; 且在 50 个智能体规模下, CommNet 的回报值与收敛速度均弱于 EMAGIC. 因此, EMAGIC 在 Ckn 任务中的性能优于其余算法. kn 任务中的性能对比如图 10(a) 和图 10(b) 所示. 在智能体规模数为 50 的场景下, EMAGIC 能够收敛至最高的回报值. 在 100 个智能体规模下, CommNet 的最终回报值略优. 但 CommNet 存在两点不足: 一是 50 个智能体规模下, CommNet 的平均回报值在时间步 1.6×10^6 处出现极大波动, 稳定性差; 二是 CommNet 的收敛速度远慢于 EMAGIC. 此外, CommNet 的全局通信在实际大规模基础设施中也难以实现. 综合 IMP 环境下每个任务中的算法性能表现, QMIX 与 VDN 缺乏通信机制, 无法充分整合各智能体的局部观测, 导致二者性能不佳. ExpoComm 虽能保证信息传递, 但其在一些任务中性能出现较大波动, 且收敛较慢, 本文认为这是缺乏数据高效利用机制造成的. CommNet 虽能传递全局信息, 但计算开销大的同时易引入冗余信息, 可能影响维护决策的稳定性. 相比之下, EMAGIC 则优于对比算法, 通过图信息通信机制提升 MAS 的协同性, 获取高平均回报值, 结合 EM 模块对具有高回报值的状态的重复利用, 提升样本效率. 同时, 基于单点指数图的通信拓扑也使 EMAGIC 具备向实际大规模场景中扩展的潜力.

本文在表 2 中展示了 EMC 与 EMAGIC 在

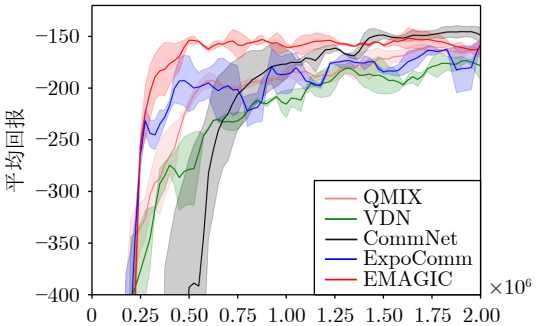
表 2 EMC 和 EMAGIC 在 IMP 环境下收敛后的平均回报值

Table 2 Converged average return values of EMC and EMAGIC in the IMP environment

算法	Owf 任务		Ckn 任务		kn 任务	
	50 个智能体	100 个智能体	50 个智能体	100 个智能体	50 个智能体	100 个智能体
EMC	−2892.53	−5785.06	−1155.38	−1743.98	−2047.55	−1788.10
EMAGIC	−432.77	−801.71	−119.79	−117.66	−168.93	−165.59



(a) 50 个智能体
(a) 50 agents



(b) 100 个智能体
(b) 100 agents

图 10　IMP 环境下 kn 任务中的对比实验

Fig. 10　Comparative experiments on kn tasks in the IMP environment

IMP 环境下收敛后的回报值, 可以看出 EMC 的性能远不如 EMAGIC. 由于 EMAGIC 在 IMP 环境下采用第 1.2 节的方法, 因此 EM 模块的设计以及参数的选择和 EMC 是完全一样的. 不同的是, EMC 将 EM 和好奇心驱动探索相结合, 而 EMAGIC 则是将 EM 与通信机制结合起来, 表明探索难以在大规模 MAS 中使用, 实验结果也验证了本文在前文中提出的观点, 即探索可能会加剧大规模环境的非平稳性.

3.3.3　消息传播能力与通信开销

本文通过一个简单例子比较多种通信拓扑在 256 个智能体下的消息传播效率, 如图 11 所示. 时间步 $t=0$ 时任意智能体获得观测信息, $t=1$ 时该智能体传递消息. 全连接图传播最快, $t=1$ 即可实现全局传播; 单点指数图在 $t=8$ 时实现全局传播, 这一点也与本文定理 1 的结论一致, 即单周期 $[1,\ 1+\lceil\log_2 256\rceil-1]=[1,\ 8]$ 内即可实现消息全局传递. 此外, 尽管静态指数图在前期消息传递速度较快, 但仍需一个周期才能实现消息全局传递. 而基于距离的图传播缓慢, $t=14$ 时仅覆盖 44 个智能体. 当回合长度超过通信周期时, 单点指数图能实现与全连接图和静态指数图相当的全局传播能力. 在本文的实验中, MAgent 环境和 IMP 环境一个回合最大时间步默认值分别为 200 和 20, 高于通信周期的长度. 因此, 基于单点指数图的通信拓扑能够将消息传播给所有智能体, 具有良好的消息传递能力.

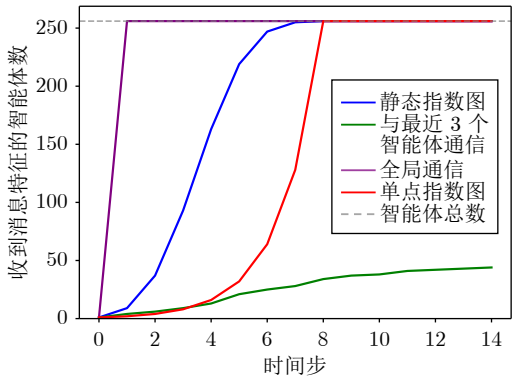


图 11　几种通信拓扑的消息传播速度比较

Fig. 11　Message propagation speed comparison among several communication topologies

考虑到实际大规模 MAS 中通信成本的约束, 本文进一步分析不同通信拓扑的通信开销. 结合第 2.1 节的分析结果, 基于单点指数图与静态指数图的通信成本分别为 $O(N)$ 和 $O(N\lceil\log_2 N\rceil)$. 同时, 本文补充全局通信以及与最近的 3 个智能体通信的通信开销对比, 如表 3 所示. 其中, 基于单点指数图的通信拓扑开销最低, 具备应用于实际大规模 MAS 的潜力. 综合通信开销与消息传递能力两种指标, 本文提出的通信拓扑以最小的通信开销实现第 2 的消息传播能力, 实现了两个指标的平衡.

表 3　几种通信拓扑的通信成本对比, 后 3 种通信拓扑分别对应 ExpoComm、DGN 和 CommNet

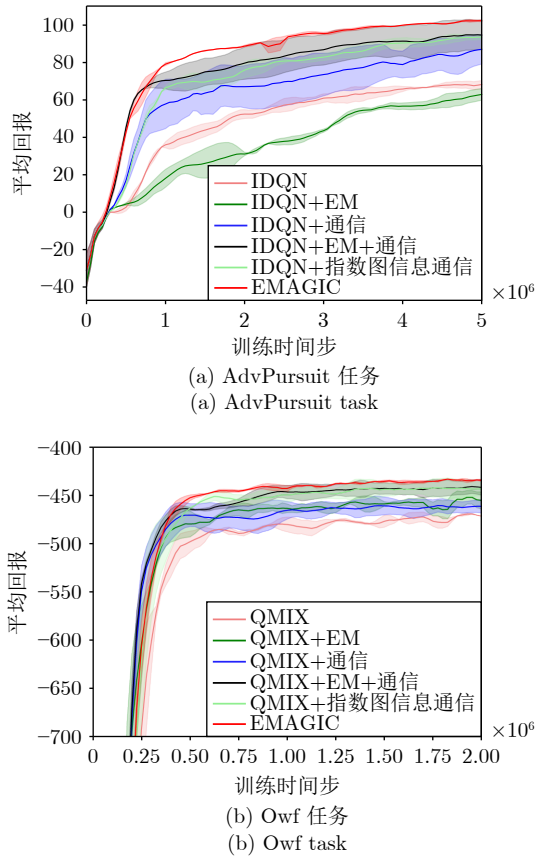
Table 3　Comparison of communication costs of several communication topologies, where the latter 3 communication topologies correspond to ExpoComm, DGN, and CommNet, respectively

通信拓扑	成本
单点指数图 (EMAGIC)	$O(N)$
静态指数图[26]	$O(N\lceil\log_2 N\rceil)$
与最近 3 个智能体通信[23]	$O(3N)$
全局通信[21]	$O(N^2)$

3.4　消融实验

本文通过消融实验展示每个组成部分对于 EMAGIC 性能的提升, 如图 12(a) 和图 12(b) 所示. 本文将 EMAGIC 各部分分为以下几组: 1) 本文的

基线算法 IDQN/QMIX; 2) 基线算法+EM; 3) 基线算法+通信, 注意该组别仅包含基线和基于单点指数图的通信拓扑, 智能体之间传递未经互信息编码后的消息; 4) 基线算法+EM+通信; 5) 基线+指数图信息通信, 包含基线、基于单点指数图的通信拓扑和图信息通信; 6) 完整的 EMAGIC. 由图 12(a) 中组 1) 和组 2) 可知, 在 IDQN 的基础上直接添加 EM 模块可能导致算法性能下降, 本文认为这是由于 IDQN 的独立学习机制而使 MAS 缺乏协同性能, 采集的样本数据缺乏高奖励状态, EM 又使用这些次优样本数据进行重复学习利用, 使得算法陷入次优策略, 这一点也验证了前文中的观点, 即 EM 易因重复学习次优样本而导致算法陷入次优策略; 相比之下, 图 12(b) 中的 QMIX 添加 EM 后算法性能有所提升, 这是由于 QMIX 本身具有混合网络利用全局状态来提升 MAS 的协同性能, 而 QMIX 本身因其集中训练机制导致其收敛较慢, EM 刚好可以弥补这一缺陷. 尽管将 IDQN 与 EM 直接结合会使算法性能下降, 但组 3) 和组 4) 以及组 5) 和组 6) 的结果说明, 通过通信机制提高算法对高价值样本的获取能力, 并结合情景记忆对这些高价值样本高效利用, 能够提升算法的性能. 此外, 组 3) 和组 5) 也进一步说明利用互信息优化消息特征可以提升算法的性能. 消融实验验证了各组成部分并非孤立作用, 而是共同支撑 EMAGIC 在大规模 MAS 中的优异性能.



(a) AdvPursuit 任务
(a) AdvPursuit task

(b) Owf 任务
(b) Owf task

图 12 EMAGIC 的消融实验

Fig. 12 Ablation experiments of EMAGIC

4 结束语

本文提出 EMAGIC 来提升 MARL 在大规模 MAS 中的性能, 设计基于单点指数图的通信拓扑来平衡通信开销与消息传播能力, 并在此基础上建立图信息通信机制, 最大化同一时间步上不同智能体间消息的互信息来对消息编码特征进行优化, 使智能体能在接收到的信息与自身观测条件下还原全局信息, 突破因部分可观测性对算法性能所带来的限制. 同时本文还引入 EM 模块提高 EMAGIC 在大规模 MAS 下的收敛性能. EMAGIC 的出色性能表明, 将 EM 和通信机制结合能使算法获得更好的回报并加快学习进程. 未来的研究将优化智能体间的通信机制, 进一步提升算法在大规模 MAS 中的性能.

参考文献

1 Peng Z, Wu G H, Luo B, Wang L. Multi-UAV cooperative pursuit strategy with limited visual field in urban airspace: A multi-agent reinforcement learning approach. *IEEE/CAA Journal of Automatica Sinica*, 2025, **12**(7): 1350−1367

2 Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623
(施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623)

3 Ou W, Luo B, Xu X D, Feng Y, Zhao Y Q. Reinforcement learned multiagent cooperative navigation in hybrid environment with relational graph learning. *IEEE Transactions on Artificial Intelligence*, 2025, **6**(1): 25−36

4 Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, Huang Ting-Wen, Yang Chun-Hua, Gui Wei-Hua. Survey on multi-agent reinforcement learning for control and decision-making. *Acta Automatica Sinica*, 2025, **51**(3): 510−539
(罗彪, 胡天萌, 周育豪, 黄廷文, 阳春华, 桂卫华. 多智能体强化学习控制与决策研究综述. 自动化学报, 2025, **51**(3): 510−539)

5 Sunehag P, Lever G, Gruslys A, Czarnecki W M, Zambaldi V, Jaderberg M, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018. 2085−2087

6 Rashid T, Samvelyan M, de Witt C S, Farquhar G, Foerster J, Whiteson S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 2020, **21**(1): Article No. 178

7 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-

agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6382−6393

8 Ackermann J, Gabler V, Osa T, Sugiyama M. Reducing overestimation bias in multi-agent domains using double centralized critics. arXiv preprint arXiv: 1910.01465, 2019.

9 Yu C, Velu A, Vinitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1787

10 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

11 Li X R, Wang X L, Bai C J, Zhang J. Exponential topology-enabled scalable communication in multi-agent reinforcement learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025. 1−25

12 Zheng L L, Chen J R, Wang J H, He J M, Hu Y J, Chen Y F, et al. Episodic multi-agent reinforcement learning with curiosity-driven exploration. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 287

13 Ma X, Li W J. State-based episodic memory for multi-agent reinforcement learning. *Machine Learning*, 2023, **112**(12): 5163−5190

14 Blundell C, Uria B, Pritzel A, Li Y Z, Ruderman A, Leibo J Z, et al. Model-free episodic control. arXiv preprint arXiv: 1606.04460, 2016.

15 Pritzel A, Uria B, Srinivasan S, Badia A P, Vinyals O, Hassabis D, et al. Neural episodic control. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org, 2017. 2827−2836

16 Zhao Y Y, Wang Z Y, Zhu C X, Wang S H. Efficient dialogue complementary policy learning via deep Q-network policy and episodic memory policy. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021. 4311−4323

17 Peng K X, Li P Y, Hao J Y. Enhancing graph-based coordination with evolutionary algorithms for episodic multi-agent reinforcement learning. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. Detroit, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2025. 1623−1631

18 Li Yi-Chun, Liu Ze-Jiao, Hong Yi-Tian, Wang Ji-Chao, Wang Jian-Rui, Li Yi, et al. Multi-agent reinforcement learning based game: A survey. *Acta Automatica Sinica*, 2025, **51**(3): 540−558 (李艺春, 刘泽娇, 洪艺天, 王继超, 王健瑞, 李毅, 等. 基于多智能体强化学习的博弈综述. 自动化学报, 2025, **51**(3): 540−558)

19 Chai J J, Zhu Y H, Zhao D B. NVIF: Neighboring variational information flow for cooperative large-scale multiagent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(12): 17829−17841

20 Li X S, Li J C, Shi H B, Hwang K S. A decentralized communication framework based on dual-level recurrence for multiagent reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems*, 2024, **16**(2): 640−649

21 Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016. 2252−2260

22 Chu T S, Wang J, Codecà L, Li Z J. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2020, **21**(3): 1086−1095

23 Jiang J C, Dun C, Huang T J, Lu Z Q. Graph convolutional reinforcement learning. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020. 1−13

24 Agarwal A, Kumar S, Sycara K, Lewis M. Learning transferable cooperative behavior in multi-agent teams. In: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020. 1741−1743

25 Nayak S, Choi K, Ding W Q, Dolan S, Gopalakrishnan K, Balakrishnan H. Scalable multi-agent reinforcement learning through intelligent information aggregation. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 25817−25833

26 Lo Y L, Sengupta B, Foerster J N, Noukhovitch M. Learning multi-agent communication with contrastive learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024. 1−18

27 Ying B C, Yuan K, Chen Y M, Hu H B, Pan P, Yin W T. Exponential graph is provably efficient for decentralized deep training. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1071

28 Song J M, Ermon S. Understanding the limitations of variational mutual information estimators. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020. 1−18

29 van den Oord A, Li Y Z, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748, 2018.

30 Nguyen X, Wainwright M J, Jordan M I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010, **56**(11): 5847−5861

31 Belghazi M I, Baratin A, Rajeshwar S, Ozair S, Bengio Y, Courville A, et al. Mutual information neural estimation. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 531−540

32 Zheng L M, Yang J C, Cai H, Zhou M, Zhang W N, Wang J, et al. MAgent: A many-agent reinforcement learning platform for artificial collective intelligence. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 8222−8223

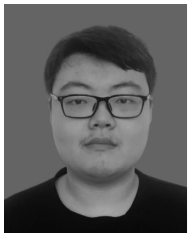
33 Chen D, Zhang K X, Wang Y Q, Yin X Y, Li Z J, Filev D. Communication-efficient decentralized multi-agent reinforcement learning for cooperative adaptive cruise control. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(10): 6436−6449

34 Leroy P, Morato P G, Pisane J, Kolios A, Ernst D. IMP-MARL: A suite of environments for large-scale infrastructure management planning via marl. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New

Orleans, USA: Curran Associates Inc., 2023. Article No. 2329

35 Papoudakis G, Christianos F, Schäfer L, Albrecht S V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 113

李方昱 北京工业大学信息科学技术学院教授. 主要研究方向为智能感知与建模, 复杂系统分析与控制和分布式智能检测优化. 本文通信作者.
E-mail: fangyu.li@bjut.edu.cn
(**LI Fang-Yu** Professor at the School of Information Science and Technology, Beijing University of Technology. His research interests include intelligent perception and modeling, complex system analysis and control, and distributed intelligent detection and optimization. Corresponding author of this paper.)



刘金溢 北京工业大学信息科学技术学院硕士研究生. 主要研究方向为复杂环境下多智能体强化学习方法设计.
E-mail: JinyiLiu@emails.bjut.edu.cn
(**LIU Jin-Yi** Master student at the School of Information Science and Technology, Beijing University of Technology. His main research interest is multi-agent reinforcement learning method design in complex environments.)

孙浩源 北京工业大学教授. 主要研究方向为复杂动态系统智能鲁棒控制.
E-mail: sunhaoyuan@bjut.edu.cn
(**SUN Hao-Yuan** Professor at Beijing University of Technology. His main research interest is intelligent robust control of complex dynamic systems.)

韩红桂 北京工业大学教授. 主要研究方向为典型资源循环利用过程智能优化控制, 神经网络结构设计与优化.
E-mail: rechardhan@bjut.edu.cn
(**HAN Hong-Gui** Professor at Beijing University of Technology. His research interests include intelligent optimization and control of typical resource recycling processes, and structure design and optimization of neural networks.)