用于改善开放式多智能体系统中强化学习的潜在交互式 A2C

Keyang He[1] **Prashant Doshi**[1] **Bikramjit Banerjee**[2]

[1]佐治亚大学计算学院，美国佐治亚州雅典[2]南密西西比大学计算科学与计算机工程学院，美国密西西比州哈蒂斯堡，

arXiv:2305.05159v1 [cs.LG] 9 May 2023

摘要

目前，多智能体强化学习（MARL）方法普遍采用集中式训练。然而，这类方法需要获取其他智能体的各种信息，在竞争或对抗环境中可能难以实现。近期提出的一种方法——交互式优势行动者-评论家（IA2C）——采用分散式训练和分散式执行，旨在根据可能含噪的观测来预测其他智能体的动作。本文提出潜在 IA2C，利用编码器-解码器架构学习隐藏状态和其他智能体动作的潜在表征。在两个各自包含大量智能体的领域中进行的实验表明，潜在 IA2C 通过降低方差、加快收敛，显著提高了样本效率。此外，我们还引入了这些领域的开放版本，其中智能体群体可能随时间变化，并在这些实例上进行了评估。

1 引言

在机器学习领域，强化学习（RL）是指智能体在主动探索环境的过程中，根据收到的奖励，自主学习在各种情境下应采取何种行动。它必须在探索与利用之间权衡，而其动作往往会产生较长期的影响（延迟奖励）。强化学习方法 [Kael- bling 等人， 1996] 颇具吸引力，因为智能体能够自主学习控制策略，且可能无需访问任何关于环境的先验规范。

近期的多智能体强化学习（MARL）技术，要么将策略优化推广到多智能体环境，要么将行动者—评论家（AC）架构推广到多智能体环境，同时主要采用集中式训练与分散式执行（CTDE）。这些技术维护联合动作价值函数或优势函数，而行动者则学习个体策略，将每个智能体自身的观测映射为其动作。例如，多智能体近端策略优化（MA-PPO） [Yu 等人， 2022] 通过使用所有智能体的动作—观测输入更新集中式价值函数，并利用该价值函数为每个智能体学习改进后的策略。基于 AC 的方法，如多智能体深度确定性策略梯度（MADDPG） [Lowe 等人， 2017],学习对方学习算法 [Foerster 等人， 2018年a], 以及反事实多智能体策略梯度（COMA） Foerster 等人[2018b]，都采用集中训练分散执行（CTDE），由智能体交换不同类型的信息，以促进集中式评论家的运作。然而，在竞争性和对抗环境中，这类信息交换可能不可行。这促使人们采用分散式训练与分散式执行 （DTDE）方法；基于 AC 的方法（如 IA2C） [He 等人， 2021],便采用了这种方法，通过对其他智能体进行建模，从主观视角学习和处理交互。这类分散式方法也更符合追求最优性的目标，即人工智能立场：设计出能够在其环境中有效运行的最优学习智能体 [Shoham 等人， 2007]。

IA2C 最近被扩展到多智能体系统中，采用的是使用狄利克雷-多项式建模 [He 等人， 2022]。这 该方法标记为 IA2C^{++}，更新狄利克雷-多项分布。使用智能体的私有观测对模型进行建模，但分别来自 AC。虽然已证明 IA2C^{++} 能够收敛到较高的... 混合动机环境中的高质量策略，我们的实验表明该方法存在较高的方差，尤 其是在某些状态下只有一个最优动作的领域中，而狄利克雷-多项式模型无法预测这一点；此外，模型的预测结果 在样本量起初较小时往往会发生剧烈变化 low.

为了提高样本效率，我们提出了一种 IA2C^{++} 变体，将狄利克雷-多项式模型及其更新机制通过编码器-解码器架构集成到我们的神经网络流程中，这也使该 DTDE 技术形成了一种端到端的神经网络架构。编码器-解码器中隐含的是一种对隐-

den 状态和其他智能体的均场动作，该均场动作作为输入传入 AC 中的优势值函数。在评论家中使用该潜变量还具有正则化效果，可降低方差并提高样本复杂度，这一点已在近期研究中得到证实 [Lee 等人，2020, Xu 等人，2022]。

我们在两个领域评估这种新方法：Org [He 等人，2021] 以及MAgent [Zheng 等人， 2018],每个领域最多包含一百个智能体。Org 模拟了一个典型的商业组织，其中既有为提升组织整体表现而进行的合作，也有员工个人竞争。本文将Org 设为一个开放的大量智能体系统，在该系统中，监督智能体可以增减员工；因此，智能体是否加入或退出由系统内的智能体控制，而非由外部因素驱动。MAgent 是一款战场模拟游戏，包含两个智能体群体；我们也为其引入了一个开放变体。在这两个领域中，我们发现，潜在 IA2C 的方差更小，样本效率也显著高于相关基线方法。

2 背景：部分可观测性下的强化学习

非平稳环境是多智能体强化学习面临的主要挑战。在合作环境中，智能体可以直接与其他智能体共享策略。在竞争环境中，这可能行不通，因此根据其他智能体过去的动作历史推断其动作是一种常见的替代方案。在现实问题中，智能体之间无法通信，且其他智能体的动作并非完全可观测，频率学派推断方法（如最大似然估计）往往难以提供准确的预测。

一种近期的强化学习方法——交互式A2C（IA2C）——在部分可观测多智能体环境中采用分散训练分散执行（DTDE）。与先前主要采用集中训练分散执行（CTDE）的多智能体强化学习方法 [Foerster 等人，2018b，Lowe 等人，2017，Foerster 等人，2018a] 相比，这种强化学习训练视角更适用于混合合作-竞争环境。采用CTDE的方法需要可能相互竞争的智能体之间交换多种类型的信息，但这些信息交换未必可行。下面我们简要回顾IA2C，并讨论如何将其扩展到与若干智能体共享的环境中执行强化学习。

2.1 交互式A2C

交互式优势演员评论家（IA2C） [He 等人， 2021] 是一种分散式演员评论家方法，旨在其他智能体共享的部分可观测马尔可夫环境中进行自我中心强化学习。在 IA2C 中，每个智能体都有自己的评论家和演员神经网络，前者将个体观测映射为该智能体自身奖励函数下的联合动作价值

， $Q_0(o,\ a_0,\ \mathbf{a}_{-0})$，后者则将个体观测映射为个体动作概率，$\pi_{0,\ \boldsymbol{\theta}}(a_0|o)$，$\boldsymbol{\theta}$ 是其参数集合。IA2C 将优势估计为

$$A_0(o,a_0,\hat{\mathbf{a}}_{-0}) = avg[r+\gamma Q_0(o,a_0',\hat{\mathbf{a}}'_{-0})-Q_0(o,a_0,\hat{\mathbf{a}}_{-0})]$$

而行动者的梯度估计为

$$avg[\nabla_{\boldsymbol{\theta}} \log \pi_{0,\boldsymbol{\theta}}(a_0|o)\ A_0(o,a_0,\hat{\mathbf{a}}_{-0})]$$

其中，r，o 和 a_0' 是从轨迹中采样得到的样本，$\hat{}\mathbf{a}_{-0}$ 和$\hat{}\mathbf{a}'_{-0}$ 是其他智能体的预测动作，avg则是对采样轨迹取平均。与先前的多智能体深度强化学习算法不同，IA2C无需获取其他智能体的动作和/或梯度。相反，智能体会维护关于其他智能体可能模型的信念分布。设 $\mathbf{a}_{-0} = a_1 \ldots,\ a_N$。给定智能体的先验信念 b_0、动作 a_0以及其公共和私有观测 o_0'、 ω_0'后，智能体会更新其对智能体 j模型的信念 $m_j' = \langle \pi_j', h_j' \rangle$，如下所示：

$$b_0'(m_j'|b_0,a_0,o_0',\omega_0') \propto \sum_{\mathbf{a}_{-0}} \Big(\prod_{k=1,k\neq j}^{N} \sum_{m_k \in M_k} b_0(m_k)$$
$$Pr(a_k|m_k) \sum_{m_j \in M_j} b_0(m_j)\ Pr(a_j|m_j)\ \delta_K(\pi_j',\pi_j)$$
$$\delta_K(APPEND(h_j,\langle a_j,o'\rangle),h_j') \Big)\ W_0(a_0,\mathbf{a}_{-0},\omega_0') \quad (1)$$

其中，公共观测和私有观测是对状态及其他智能体动作的含噪观测。 m_j 表示智能体j的模型，π_j 是 j的策略， h_j 是其动作—观测历史。 W_0 是私有观测函数，用于将联合动作映射为主体智能体的私有观测。 δ_K是克罗内克δ函数，而APPEND返回一个字符串，该字符串是将第二个参数追加到第一个参数后得到的结果。信念更新由一个独立的信念滤波器执行，该滤波器集成在评论家网络中。

2.2 多智能体交互式 A2C

MARL 的另一个关键挑战是联合动作空间的指数级增长。许多多智能体环境天然具有动作匿名性 特征，这意味着环境动态和奖励取决于动作配置 （以下记作 $\mathcal{C}$），也就是群体中动作的计数分布。多智能体 IA2C（IA2C^{++}） [He 等人， 2022] 通过利用动作匿名性的这一关键洞见扩展了 IA2C，从而使用随智能体数量呈多项式增长的配置，而不是使用呈指数级增长的联合动作向量。

IA2C^{++}BU 改编公式 1，以更新智能体 0 对另一个智能体可能采用的模型的信念，具体是在动作

配置而非联合动作的情境下。

$$b_0'(m_j'|b_0,a_0,o_0',\omega_0') \propto \sum_{m_j\in M_j} b_0(m_j) \sum_{a_j} Pr(a_j|m_j)$$
$$\sum_{\mathcal{C}\in\mathcal{C}^{\mathbf{a}-\mathbf{o}}} Pr(\mathcal{C}|b_0(M_1),b_0(M_2),\ldots,b_0(M_N))\, W_0(a_0,\mathcal{C},\omega_0')$$
$$\delta_K(\pi_j,\pi_j')\delta_K(APPEND(h_j,\langle a_j,o'\rangle),h_j'). \quad (2)$$

术语 $Pr(\mathcal{C}|b_0(M_1)\ b_0(M_2)\ldots,b_0(M_N))$ 表示配置 $\mathcal{C}$ 在配置集合 $\mathcal{C}^{\mathbf{a}-\mathbf{o}}$上的分布中的概率。该分布通过 [Jiang 等人2011]提出的动态规划过程得到。与单个规模呈指数级增长的大型信念 $|M_j|^N$相比，该算法的输入仅为 N 个信念，每个信念的大小为 $|M_j|$ 。图 1 展示了 IA2C^{++}BU 的网络架构。

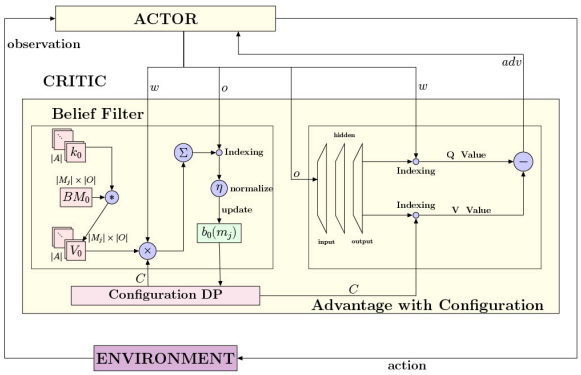


图 1：He 等人 [2022]的 IA2C^{++}BU 网络架构。信念滤波器根据观测到的配置，通过动态规划过程更新智能体对群体的信念。

IA2C^{++}BU 需要预先定义的模型集来进行信念更新。IA2C^{++}DM 则通过使用狄利克雷-多项式模型来近似整个智能体群体的动作分布，从而消除了这一假设，进而可以得到平均动作。假设同质智能体群体的动作空间为 $\{a_1,a_2,\ldots,a_{|A|}\}$，对于每个智能体 i，选择动作 a_n 的概率为 θ_n。令分布 $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_{|A|})$。$\boldsymbol{\theta}$ 具有参数为 $\boldsymbol{\alpha}$ 的狄利克雷-多项式先验分布，若

$$Pr(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_n \alpha_n)}{\Pi_n \Gamma(\alpha_n)} \Pi_n \theta_n^{\alpha_n-1} \quad (3)$$

其中，$\alpha_n > 0$ 对于所有 n，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots\ldots,\alpha_N)$，且$\sum_n \theta_n = 1$。我们可以将其写为 $\boldsymbol{\theta}\sim Dir(\boldsymbol{\alpha};\boldsymbol{\theta})$。那么，动作配置 $\mathcal{C}$ 的概率表示为：

$$Pr(\mathcal{C}|\boldsymbol{\theta}) = Pr(\#a_1,\#a_2,\ldots,\#a_{|A|}|\boldsymbol{\theta}) = \Pi_{n=1}^{|A|}\theta_n^{\#a_n} \quad (4)$$

其中 $\#a_k$ 是选择第 k 个动作的智能体数量。在执行动作 a_0 并接收（带噪声的）私有观测 $\omega_0'=(\#a_1,\ \#a_2,\ \ldots\ldots,\ \#a_{|A|})$ 后，主体智能体的

通过注意到后验分布 $Pr(\boldsymbol{\theta}|a_0,\ \omega_0')$ 与以下内容成正比，可以更新狄利克雷-多项式分布：

$$\propto \sum_{\mathcal{C}} W_0(a_0,\mathcal{C},\omega_0')Pr(\mathcal{C}|\boldsymbol{\theta})\cdot Dir(\boldsymbol{\alpha};\boldsymbol{\theta}) \quad (5)$$
$$= \sum_{\mathcal{C}} W_0(a_0,\mathcal{C},\omega_0')Dir(\boldsymbol{\alpha}+\mathcal{C};\boldsymbol{\theta}) \quad (6)$$
$$\approx Dir(\boldsymbol{\alpha}+\mathcal{C}';\boldsymbol{\theta}) \quad (7)$$

式 5 作出了一个简化的建模假设，即其他智能体的真实动作计数 $\mathcal{C}$ 在给定 $\boldsymbol{\theta}$ 的条件下与 a_0 条件独立。为便于处理，最后一步被近似为单个分量 $Dir(\boldsymbol{\alpha}+\mathcal{C}';\boldsymbol{\theta})$。$\mathcal{C}'$ 可以通过若干种替代方法计算，正如 He 等人所讨论的那样 [2022], ，而本文采用 修正方法。

在 IA2C^{++} 中，评论家网络将优势估计为：

$$A_0(o,a_0,\mathcal{C}^{\mathbf{a}-\mathbf{o}}) = avg[r+\gamma\, Q_0(o'',a_0',\mathcal{C}^{\mathbf{a}'-\mathbf{o}}) - Q_0(o',a_0,\mathcal{C}^{\mathbf{a}-\mathbf{o}})]$$

而行动者网络的梯度为：

$$avg\,[\nabla_{\boldsymbol{\theta}}\, log\ \pi_{0,\boldsymbol{\theta}}(a_0|\boldsymbol{o})\ A_0(o',a_0,\mathcal{C}^{\mathbf{a}-\mathbf{o}})] \quad (8)$$

其中，r、o'、o'' 和 a_0' 是样本，$\mathbf{a_{-0}}$ 和 $\mathbf{a'_{-0}}$ 分别是其他智能体在当前时间步和下一时间步的预测联合动作，并由其对应的配置替代；avg 由采样得到的轨迹接管。下一时间步的预测动作从更新后的狄利克雷分布中采样。

3 开放组织领域

典型的商业组织（Org）既包含员工之间为改善组织整体财务健康状况而进行的合作，也包含员工个人竞争。Org 的状态对应组织的财务健康状况。智能体（员工）只能观察到组织收到的订单数量，这些订单是反映Org潜在状态的噪声指标。智能体的奖励来自个人奖励部分和群体奖励部分：前者取决于智能体自身的动作，后者取决于所有智能体的联合行动，而这可能促使智能体与他人协调。

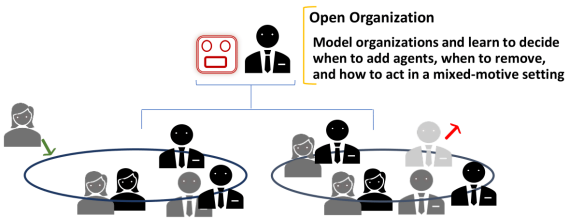


图 2：在开放组织领域中，经理可以从群体中移除员工或将员工加入群体，而任何员工都可以辞职。

Org 提供的挑战比以往的多智能体强化学习评估领域更加贴近现实。然而，原始Org 假设这是一个封闭系统 [He 等人，2021]。在现实生活中的组织中，员工可能会因退休或跳槽等各种原因离开组织。组织会招聘员工来填补职位空缺，或因业务萎缩而裁员。为了模拟这种情况，我们引入开放式 *Org* 领域，将Org推广到开放式多智能体系统（可视化示例见图 2）。开放式Org 包含两类智能体：员工 和经理。员工可以选择离开组织。经理可以根据组织当前的财务健康状况招聘新员工或解雇现有员工。与现有的开放系统领域不同，在这些领域中，外部智能体会定期和/或随机进入系统 [Radulescu 等人，2018]，；在开放式Org中，系统开放性完全由内部智能体控制。内部智能体的目标是在系统开放性的影响下优化其收益。

3.1 状态与观测

五种状态表示组织的财务健康状况水平，如图 3 所示：极低（s_{vl}）、低（s_l）、中等（s_m）、高（s_h）和极高（s_{vh}）。智能体无法观测状态。相反，智能体会接收组织收到的订单数量的观测，这些数量与状态相关，并且该信息是公开的。当组织处于 s_{vl} 或 s_l 时，会观测到少量（o_e）。当组织处于 s_m 或s_h 时，会观测到若干（o_s）。当组织处于 s_{vh} 时，会观测到大量（o_m）。智能体还会接收表示其他智能体动作的私有观测。当 $0 \leq \delta < 1$时，私有观测以 $1-\delta$ 的概率揭示其他智能体的真实动作。

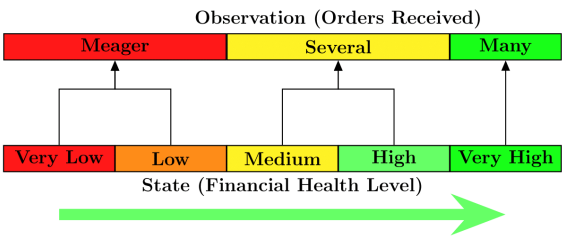


图 3：开放式Org中的状态、观测及其相互关系。

3.2 动作与状态转移

员工有四种可能的动作：自我、平衡、群体和辞职。自我 动作可获得较高的个体奖励，但没有群体奖励；群体 动作可获得中等的群体奖 励，但没有个体奖励；平衡 动作可获得较低的个体奖励和较低的群体奖励。辞职 动作会将员工移出组织。在本文中，我们让Org 仅包含一名经理。经理有六种可能的动作：self、balance、group、fire和hire。Self、balance和group 与员工的动作类似。fire 动作会移除最近雇用的员工，让人想起近期科技行业为纠正疫情时期过度招聘而进行的裁员。hire 动作会为组织添加一名新员工。

联合动作由选择自我 的智能体人数与选择 群体的智能体人数决定，这可能导致状态转变。选择自我 的智能体越多，组织的财务健康状况就越差。选择群体的智能体越多，组织的财务健康状况就越好。如果选择两者的人数相同，组织的财务健康状况则保持不变。员工的动作辞职 以及经理的动作 解雇 和 雇用 不会影响组织的财务健康状况。此外，智能体只能获得他人动作的含噪观测。为了最大化总体回报，智能体需要权衡个人利益与群体福利，以优化长期收益。

3.3 奖励

员工的奖励函数与原始Org中的相同。经理的奖励函数由三部分组成：个体、群体和成本：

$$R_m^t = \sigma(\sum_i R_i^t) + R_G^t - C^t$$

其中，σ 是一个Sigmoid函数，用于模拟员工过多时个体奖励的边际递减。$\sum_i R_i$ 是所有智能体个体奖励的总和；R_G 是由联合行动决定的群体奖励；$C = c\times$ 表示此时间步雇用的员工数量，其中 c 是雇用一名员工的成本。

3.4 组织中的智能体开放性

与封闭式多智能体系统相比，开放式多智能体系统面临更严重的非平稳性问题。这是因为任何员工智能体都可以随时离开系统或被移除，也可能有新员工加入系统。新智能体可能不了解该领域，从探索性动作开始；也可能是具备领域知识的预训练智能体。

在目前研究的若干开放智能体领域中，内部智能体并不能完全控制系统开放性。例如，Radulescu 等人 [2018]提出了一个高速公路驾驶领域的开放系统模拟。高速车辆会随机进入系统，而内部车辆若距离主观车辆过远，则会离开系统。另一个例子是 Chandrasekaran 等人 [2016]提出的消防领域。内部消防员在灭火剂耗尽后会离开系统。只有此前离开系统的消防员

之后才能重新进入系统。当内部智能体对系统开放性只有最低限度的控制权时，理性决策就需要对开放性进行建模，并预测其造成的影响。相比之下，在开放式Org中，内部智能体可以主动管理系统开放性，以实现最优状态。

4 潜在交互式A2C

本节介绍潜在交互式A2C（LIA2C），该方法利用编码器-解码器网络，将智能体群体的底层隐藏状态和预测动作建模为潜在变量，以用于部分可观测多智能体环境中的强化学习。

4.1 编码器-解码器网络

我们的目标是用编码器-解码器组合替换 IA2C^{++} 评论家的信念过滤模块。图 4 展示了该编码器-解码器的神经网络架构。

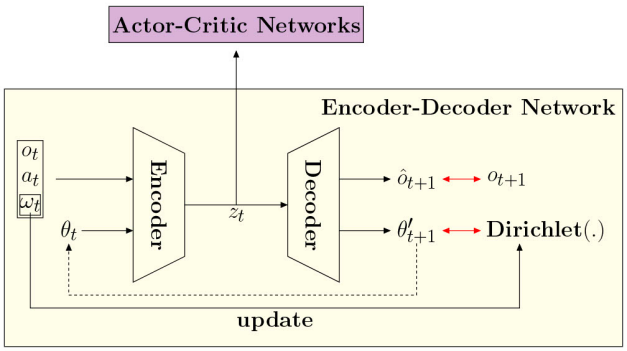


图 4：用于建模底层状态和智能体群体的编码器-解码器网络。其中，ω_t 和 o_t 分别是当前的私有观测和公共观测，a_t 是自身动作，θ_t 是解码器根据时间步 t预测的动作分布。$\hat{}o_{t+1}$ 是预测的下一时刻公共观测，θ'_{t+1} 是智能体群体更新后的动作分布。请注意，原始的狄利克雷分布在更新时也依赖于 ω_t。

在每个时间步 t，编码器网络的输入包括公共观测 o_t、私有观测 ω_t、自我动作a_t以及智能体群体的动作分布 θ_t。编码器网络根据输入和标签生成潜在嵌入 z_t。潜在嵌入 z_t 会发送至演员-评论家网络以进行强化学习，并发送至解码器网络以进行重构。给定 z_t，解码器网络通过观测重构头 f_d^o生成下一时间步的公共观测 o_{t+1}，并通过模型重构头 f_d^θ生成更新后的智能体群体动作分布 θ_{t+1}。编码器-解码器网络的损失函数定义如下：

$$\begin{aligned}\mathcal{L} =& \frac{1}{H}\sum_{t=1}^{H}[(f_d^o(z_t) - o_{t+1})^2 - \log Pr(f_d^\theta(z_t)|\omega_t, a_t) \\ &+ \mathcal{D}_{KL}(Dir(\mathcal{C}; f_d^\theta(z_t))||Dir(\alpha + \mathcal{C}'; \theta))] \\ =& \frac{1}{H}\sum_{t=1}^{H}[(f_d^o(z_t) - o_{t+1})^2 - \log Dir(\alpha + \mathcal{C}'; f_d^\theta(z_t)) \\ &+ \mathcal{D}_{KL}(Dir(\mathcal{C}; f_d^\theta(z_t))||Dir(\alpha + \mathcal{C}'; \theta))] \end{aligned} \tag{9}$$

损失函数的第一项是重构的下一时刻观测$\hat{}o_{t+1}$ 与真实下一时刻观测 o_{t+1}之间的均方误差。第二项是狄利克雷-多项式模型的对数后验（见式 7），我们希望最大化该项。$\mathcal{C}'$ 通过第 2.2 节所述的整流方法计算得出。最后一项损失是重构模型与通过维持狄利克雷分布、利用私有观测历史得到的模型之间的 KL 散度。这里，KL 散度第二项中的 θ 是根据式 3 从分布 $Pr(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{\mathcal{C}'})$中采样得到的。KL 散度第一项中的 $\mathcal{C}$ 则根据式 4 从 $Pr(\mathcal{C}|f_d^\theta(z_t))$中采样得到。通过最后两项损失，可以将该网络视为既学习匹配由狄利克雷-多项式模型驱动的外部信念更新算法（式 3–7），又学习最大化由此得到的后验。直观而言，该网络学习更新隐变量上的信念，从而在信念空间中创建潜在表示 z_t。这一潜在表示为我们的去中心化评论家提供了恰当的增广。

如果网络没有学会匹配外部信念更新算法，而是学会了该算法的内部版本，会产生什么后果？相应的损失函数将不包含最后一个损失项，而后验最大化（第二项）将成为这一学习方面的唯一驱动力。我们推测，这一版本的收敛速度会更慢。此外，它可能会学到与现有基线方法相似的策略，而这些基线方法同样没有显式地对信念更新进行建模。我们将此版本作为实验中的一个基线方法，称为 LIA2C-w/oKLD，以评估这些推测。

请注意，狄利克雷分布仅在训练期间才是必需的。在执行期间，无需维护狄利克雷分布，因为我们的行动者网络不使用 z_t 或该网络生成的任何其他内容。因此，在执行时，评论家和该编码器-解码器网络都可以被丢弃。

4.2 用于强化学习的演员-评论家网络

我们扩展了评论家网络的输入，将潜在嵌入 z_t 纳入其中，该嵌入来自编码器-解码器网络。具体而言，智能体 0 的 Q-函数现在为 $Q_0(z, o, a_0, \omega_0)$。

评论家的优势函数定义如下，

$$A_0(z, o, a_0, \omega_0) = avg\,[r + \gamma\, Q_0(z', o'', a_0', \omega_0'') - Q_0(z, o', a_0, \omega_0')] \quad (10)$$

其中， r、 o'、 o''、 ω_0'、 ω_0'' 和 a_0' 是样本， z 和 z' 是编码器-解码器网络的潜在嵌入， avg 接管了采样轨迹。智能体群体在下一时间步的预测动作通过从更新后的解码器模型中采样获得。请注意，评论家更新产生的梯度不会反向传播至编码器-解码器网络。

将潜在嵌入也纳入行动者网络的输入同样是可取的，因为它能为行动者的计算提供有关隐藏状态的有用信息。然而，我们出于两个原因避免这样做。第一个原因是技术上的困难：为了运行行动者并输出 a_t，我们不能将 z_t 用作其输入之一，因为编码器本身需要将 a_t 作为其输入之一。改用 z_{t-1} 作为行动者的输入则会引入一步延迟。第二个原因是历史先例。正如 Lee 等人 [2020]在论文中所指出的：“.....策略并不以潜在状态为条件，因为这可能导致过于乐观的行为：算法会为能够完美获取潜在状态的策略学习 Q 值。相反，我们算法中学习到的策略直接以过去的观测和动作为条件。这样做还有一个好处，即学习到的策略可以在运行时执行，而无需推断潜在状态。”这一观察同样适用于 LIA2C。因此，我们让行动者网络仅根据公共观测 o来推荐动作。请注意，潜在状态仍会通过优势项影响行动者的更新，因为其梯度为

$$avg\,[\nabla_\theta \log \pi_\theta(a_0|o)\; A_0(z, o, a_0, \omega_0)]. \quad (11)$$

图 5 展示了 LIA2C 的整体网络架构。行动者接收来自环境的观测，并将自我动作发送到编码器-解码器网络。编码器-解码器网络根据行动者的动作和来自环境的观测更新近似模型θ' 。编码器-解码器网络的潜在嵌入被发送到评论家以计算优势。评论家网络根据潜在嵌入和环境奖励更新其参数，并将优势值发送给行动者以更新其梯度。

5 实验

我们使用 Python [1] 实现了 LIA2C，并在两个领域的多个实例上评估其学习表现，与重要的基线方法进行比较。

[1]我们的代码将在论文被接收后发布。

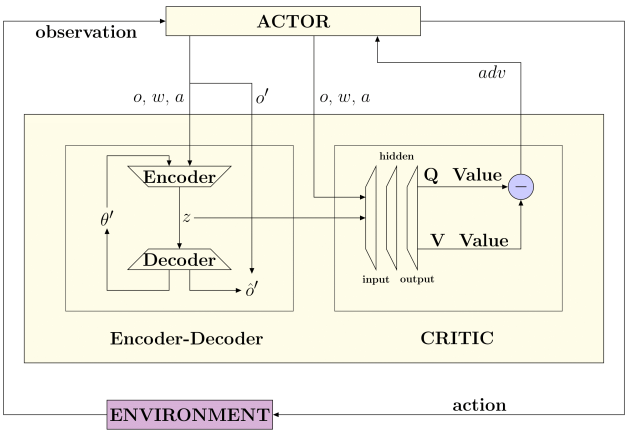
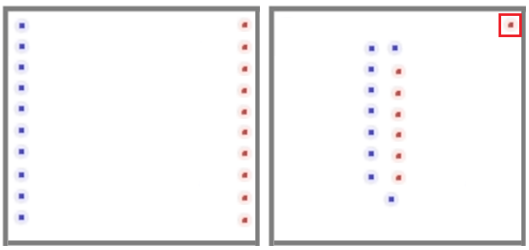


图 5：LIA2C 的总体架构。编码器-解码器网络更新智能体群体的模型。评论家的优势函数现在取决于潜在嵌入、自我动作和观测，而不是配置。

我们的第一个实验领域是Org；我们使用原始（封闭式）Org [He 等人， 2022]以及我们新提出的开放式 Org（见第 3 节）。我们的Org 是完全连通的，所有智能体共享同一个邻域。对于开放式Org 领域， c 雇用一名员工的成本设为 1，而辞职、雇用和解雇的个体奖励均设为 0。环境初始时有一名经理和一名员工。

我们的第二个实验领域是 Battlefield [He 等人，2021]，位于 MAgent 环境 [Zheng 等人， 2018] 中，每支队伍有 10 个智能体，如图 6 所示。我们通过添加一个新的动作撤退，将系统开放性引入 Battlefield。所有蓝色智能体都在战场左侧生成，所有红色智能体都在战场右侧生成。如果智能体在当前时间步未被淘汰，撤退 动作会将该智能体从当前位置移除，并根据其颜色将其重新生成在左侧或右侧。



(a) 初始位置　　(b) 重新生成的红方智能体

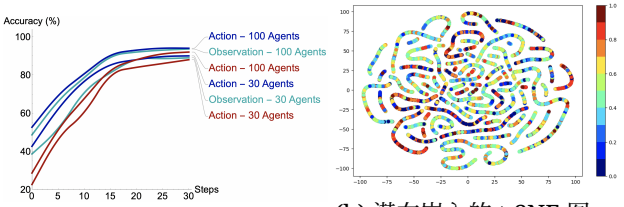
图 6：(*a*) 蓝队智能体部署在左侧，红队智能体部署在右侧。(*b*) 一名红方智能体在撤退后重新生成。

基线方法 基线方法是指在损失函数中去掉 KL 散度项的 LIA2C，标记为 LIA2C-

w/oKLD、IA2C^{++}DM [He 等人， 2022]，以及一种近期用于建模其他智能体、同样使用编码器-解码器的方法，标记为 LIAM [Papoudakis 等人， 2021]。实现细节和超参数详见附录 A。

5.1 模型性能

预测准确率 我们通过预测准确率评估编码器-解码器模型的性能。图 7a 展示了 LIA2C 中编码器-解码器模型对动作（蓝色）和观测（青色）的预测准确率，以及 IA2C^{++}DM 中狄利克雷模型对动作（红色）的预测准确率。



(a) 预测准确率 (b) 潜在嵌入的 t-SNE 图嵌入表示

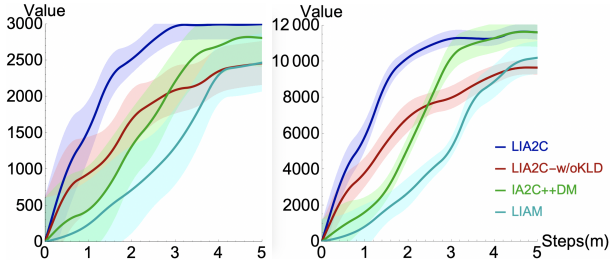
图 7：(a) 动作和观测的预测准确率随时间步的变化。蓝色和青色曲线表示 LIA2C 的预测准确率，红色曲线表示 IA2C^{++}DM 的预测准确率。(b) 编码器-解码器潜在嵌入的 t-SNE 图。不同颜色表示 θ 的不同分布，不同的点表示不同的时间步。

可视化 我们使用 t-SNE [Maaten and Hinton, 2008] 对编码器-解码器模型学到的潜在嵌入进行可视化。图 7b 展示了潜在嵌入的二维投影。我们观察到，颜色相近的点彼此相连，表明对他人动作分布的信念平滑过渡。

5.2 性能比较

Org 和开放Org 图 8 展示了来自原始Org 的实验结果，其中包含 30 个和 100 个智能体。在原始Org中，LIA2C 经过约 2M 步收敛到最优策略。IA2C^{++}DM 在经过 3.5M 步后收敛到最优策略。LIA2C-w/oKLD 在 5M 步内未收敛到最优策略。由于私有观测存在噪声，LIAM 只能收敛到次优策略。此外，LIA2C 的平均方差比 IA2C^{++}DM 低 32%，比 LIAM 低 44%。

开放式 Org 实验从 1 名经理和 1 名员工开始。经理的个人奖励部分按 0.9^{E-1} 进行折扣，其中 E 是员工数量



(a) 30 个智能体组织 (b) 100 个智能体组织

图 8：(a) LIA2C 的方差低于 IA2C^{++}DM 和 LIAM。(b) 与基线方法相比，LIA2C 的样本效率更高，因为它只需少得多的样本即可收敛。

。经理为雇用一名员工所承担的成本 C 是每个时间步 1。图9a展示了开放式Org中不同方法的策略值。

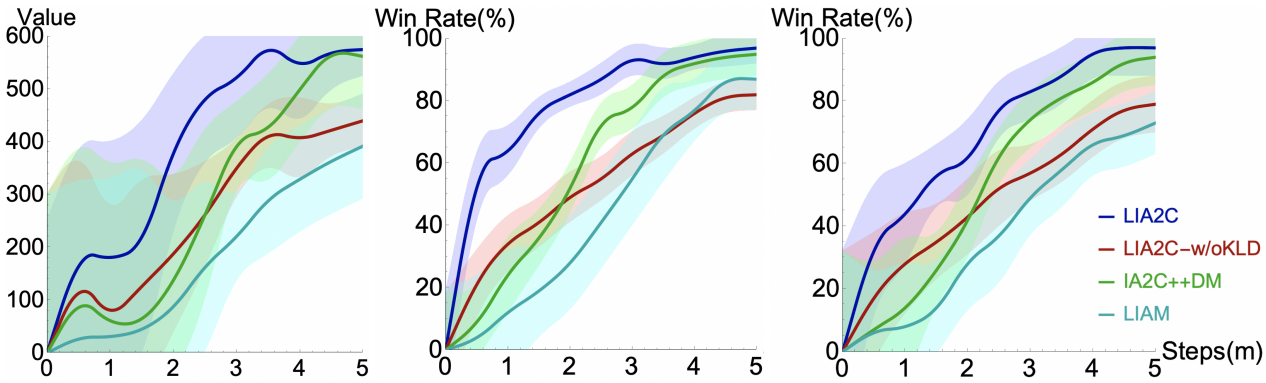
我们观察到，在开放式组织中，所有方法的方差都更高。系统开放性大大加剧了环境的非平稳性。LIA2C 在 350 万步后收敛，而 IA2C^{++}DM 在 450 万步后收敛。LIA2C 和 IA2C^{++}DM 都学会了雇用 8 名员工；根据奖励函数，这是应雇用的最佳员工人数。员工的策略与原组织保持一致。LIA2C-w/oKLD 和 LIAM 未能收敛到高质量策略。

MAgent战场与开放式MAgent 战场在MAgent 战场领域中，所有方法均针对由独立 A2C 预训练的智能体进行测试。图9b 展示了相应方法的平均胜率。

LIA2C 在 2M 步后学得高质量策略。IA2C^{++}DM 在 4M 步后学得相似的策略。LIAM 收敛到与 LIA2C 和 IA2C^{++}DM 相近的策略，但方差更高。平均而言，LIA2C 的方差比 IA2C^{++}DM 低 29%，比 LIAM 低 41%。

在开放的 MAgentBattlefield 领域中，所有方法均针对由独立 A2C 预训练的智能体进行测试。图 9c 展示了相应方法的平均胜率。

由于系统开放性，LIA2C 需要近 4M 步才能收敛到高质量策略。IA2C^{++}DM 学到的策略与 LIA2C 非常接近。我们注意到，在开放的 MAgent Battlefield 中，LIAM 的表现有所下降。在私有观测存在噪声的情况下，LIAM 似乎难以在开放系统中学习准确的潜在表示。



(a) 开放Org (b) MAgent战场 (c) 开放式 MAgent 战场

图 9：(a) 开放Org 从一名经理和一名员工开始。LIA2C 和 IA2C^{++}DM 策略共雇用 8 名员工，从而使经理获得最优收益。(b) 经过 2M 步后，LIA2C 学会了一个胜率约为 80% 的策略，而基线方法策略的胜率均低于 50%。(c) 由于系统开放性带来了额外的不确定性，所有方法都花费了更多步数来学习高质量策略。

6 相关工作

多智能体强化学习 集中训练分散执行一直是近期多智能体强化学习研究中的常见范式。先前的工作 [Lowe 等人，2017年，Foerster 等人，2018年b，Yu 等人，2022]采用集中式评论家和分散式行动者架构。集中训练分散执行中的另一个研究方向是价值分解，以缓解基于联合行动的价值表示问题。Sunehag 等人 [2018],Rashid 等人 [2018] 将联合 Q 函数分解为个体智能体局部 Q 函数的函数。使用无监督学习技术学习环境的低维表示，已经在策略质量和样本效率方面带来了有希望的改进。近期工作 [Xu 等人，2022，Papoudakis 等人， 2021] 专注于学习可变模型，这些模型由状态表示和环境动态的联合构成。与这些工作不同，LIA2C 专注于学习交互状态的表示以及潜在状态动态。

多智能体开放系统中的规划与学习在大多数现实问题中，环境中的智能体数量通常并不固定。智能体可能因各种原因离开或进入环境。系统的开放性与多智能体设置大大加剧了环境的非平稳性。Chandrasekaran 等人 [2016]，Eck 等人 [2020]从决策理论的角度对多智能体开放系统进行建模。通过对邻居建模并预测其未来是否存在及其行为，可以应对多智能体开放系统中的不确定性。迁移学习方法也可用于多智能体开放系统中的强化学习。Radulescu 等人 [2018]提出了一种集中训练分散执行方法：先学习单智能体策略，再将知识从单智能体设置迁移到多智能体设置。该策略通过直接共享参数的方式共享，因此仅适用于同质群体，

体。近期一种名为基于图的策略学习（GPL） [Rahman 等人 2021] 的方法，基于图神经网络（GNN），在完全可观测的多智能体开放系统中学习智能体模型和联合动作价值模型。GPL 学习一个联合动作价值函数，用于建模其他智能体的动作对主体智能体回报的影响；同时，它还使用基于 GNN 的模型来预测其他队友的动作。

7 结语

我们提出了一种名为 LIA2C 的新方法，用于在分散训练分散执行（DTDE）设置下进行多智能体系统中的强化学习。该方法延续了多智能体强化学习（MARL）中潜在空间建模的当前趋势，但与现有研究有一个重要区别。LIA2C 的潜在空间并非仅对隐变量进行建模，而是同时涵盖这些变量以及针对这些变量的信念更新算法。由此带来的模型预测能力提升，使其在两个多智能体领域——Org 和 MAgent——以及本文同时引入的开放变体中，相较于近期基线方法具有更高的样本效率和相对更低的方差。此外，鉴于 LIAM 学习到的策略与 LIA2C-w/oKLD 学习到的策略质量相差不大，我们得出结论：显式建模信念更新——LIAM 也没有采用这一做法——是 LIA2C 成功的关键因素。

本研究为未来的研究开辟了令人振奋的方向。在开放领域的实验中观察到的高方差表明，有必要重新关注长期以来因聚焦于集中训练分散执行（CTDE）而被忽视的非平稳性。事实上，开放性引入的波动性很可能适合采用潜在空间建模。尽管本文采用了现有的多智能体强化学习（MARL）建模概念，例如狄利克雷多项式模型和平均场动作，但其他

能够借鉴统计力学丰富文献的概念（例如，参见 [Hu等人， 2019]）或许尤其适用于开放多智能体系统，并可能从未来的研究中受益。

参考文献

Muthukumaran Chandrasekaran、Adam Eck、Prashant Doshi 和 Leenkiat Soh。《开放且具有类型的智能体系统中的个体规划》。载于人工智能中的不确定性，2016 年。

Adam Eck、Maulik Shah、Prashant Doshi 和 Leen-Kiat Soh。《开放式类型化多智能体系统中的可扩展的决策理论规划》。载于 人工智能促进协会（*AAAI*），2020 年。

Jakob Foerster、Richard Y. Chen、Maruan Al-Shedivat、Shi-mon Whiteson、Pieter Abbeel 和 Igor Mordatch。《具备对手学习感知的学习》。第122–130页。自主智能体与多智能体系统国际联合会议（AAMAS），2018a年。

Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。载于“人工智能促进协会（*AAAI*）”，2018年b。

Keyang He、Bikramjit Banerjee 和 Pranshant Doshi。依赖历史的奖励下的合作竞争强化学习。载于 自主智能体与多智能体系统（*AAMAS*），2021年。

Keyang He、Pranshant Doshi 和 Bikramjit Banerjee。部分可观测性下多智能体环境中的强化学习。载于 人工智能中的不确定性（*UAI*），2022。

Shuyue Hu、Chin-wing Leung 和 Ho-fung Leung。重复对称博弈中的多智能体Q学习动力学建模：平均场理论方法。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett（编），神经信息处理系统进展，第32卷。Curran Associates, Inc.，2019年。

Albert Xin Jiang、Kevin Leyton-Brown 和 Navin A.R. Bhat。动作图博弈。博弈与经济行为，71(1)：141–173，2011年。

Leslie Pack Kaelbling、Michael L. Littman 和 Andrew W. Moore。《强化学习：综述》。人工智能研究杂志（*JAIR*），4(1):237–285，1996 年。

Alex X.Lee、Anusha Nagabandi、Pieter Abbeel 和 Sergey Levine。《随机潜变量演员-评论家：使用潜变量模型的深度强化学习》。发表于神经信息处理系统（*NeurIPS*），2020 年。

Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。用于混合合作-竞争环境的多智能体演员-评论家。载于神经信息处理系统（*NIPS*），2017年。Laurens V. D. Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究杂志》，第 2579–2605 页，2008年 11 月。Georgios Papoudakis、Filippos Christianos 和 Stefano V. Albrecht。深度强化学习中部分可观测性下的智能体建模。载于神经信息处理系统（*NeurIPS*），2021年。

Roxana Radulescu、ManonLegrand、KyriakosEfthymiadis 和 Diedeik Roijers。《同质开放群体中的深度多智能体强化学习》。《人工智能》，第 90–105 页，2018年11月。

Arrasy Rahman、Niklas Hopner、Filippos Christianos 和 Stefano V. Albrecht。迈向基于图的策略学习的开放式临时协作。发表于国际机器学习大会（*ICML*），2021年。

Tabish Rashid、Mikayel Samvelyan、Christian Schroeder de Witt、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于机器学习国际会议（*ICML*），2018年。

Yoav Shoham、Rob Powers 和 Trond Grenager。多智能体学习如果是答案，那么问题是什么？人工智能，171(7):365–377，2007年。doi: https://doi.org/10.1016/j.artint.2006.02.006。

Peter Sunehag、Guy Lever、Audrunas Gruslys、Woj- ciech Marian Czarnecki、Vinicius Zambaldi、Max Jader- berg、Marc Lanctot、Nicolas Sonnerat、Joel Z. Leibo、Karl Tuyls 和ThoreGraepel。基于团队奖励的合作多智能体学习价值分解网络。载于自主智能体与多智能体系统（*AAMAS*），第2085–2087页，2018年。

Zhiwei Xu、Yunpeng Bai、Dapeng Li、Bin Zhang 和 Guo- liang Fan。SIDE：面向部分可观测合作多智能体强化学习的状态推断。载于自主智能体与多智能体系统（*AAMAS*），2022。

Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。《PPO 在合作多智能体博弈中的惊人效果》。载于神经信息处理系统（*NeurIPS*），2022。

Lianmin Zheng、Jiacheng Yang、Han Cai、Ming Zhou、Weinan Zhang、Jun Wang 和 Yong Yu。《MAgent：面向人工集体智能的多智能体强化学习平台》。载于人工智能促进会（*AAAI*），2018年。