[[1]](#footnote-1)

Coordination as Inference in Multi-Agent Reinforcement Learning

First A. Author, *Fellow, IEEE*, Second B. Author, and Third C. Author, Jr., *Member, IEEE*

***Abstract*—尽管集中式训练和分散式执行范式由于其卓越的性能而备受关注，但是使用集中训练学习到的完全分散的策略可能会由于非平稳性而失效。本质上最类似于人类合作模式的协作模型是完全分散的，其中代理的策略是独立优化的，即independent learning (IL)。然而，为了使得智能体通过IL学习协作，我们仍需回答两个问题：在分散训练下，智能体如何意识到其他智能体的存在，以及如何与其他智能体协作来改进自己的策略。为此，在本文中，我们提出了一种基于推断的协调MARL 方法。具体来说，它在我们的方法中包括两个部分。第一个是基于有条件的深度生成模型的个体代理意图建模；第二个是基于因果推断的代理级协调。提出的模型在 StarCraftII和Multi-Agent MuJoC基准上进行了广泛的实验。结果表明，与包括 MAPPO 和 HAPPO 在内的强大基线相比，即使不采用CTDE范式也可以学习到智能体间的协调行为。**

***Index Terms*—多智能体、深度强化学习、部分可观察性、非平稳性、变分推断、因果推断、心智理论**

# I. INTRODUCTION

合作多智能体强化学习（MARL，Oroojlooy jadid &amp; Hajinezhad，2019）的核心挑战之一是如何学习多个代理之间的协调行为，尤其是在部分可观测环境中。这是由于在执行时，一组代理在共享环境中相互交互，每个代理将其他代理视为其环境的一部分并仅根据其本地观察采取行动，这就引入了著名的非平稳性问题。该领域的先前工作（例如，Foerster 等人（2017；2016））通常采用集中训练分散执行（MADDPG）的范式来缓解非平稳性。利用这种范式的一些有前途的方法包括价值函数分解方法(VDN [20], QMIX [14],QTRAN)，基于通讯的方法(TarMAC, I2C, TMC)和基于协作图的方法(DCG, CASEC)。

然而，价值函数分解方法假设最优联合动作代表了每个智能体的最优动作集合, 且对联合 Q 函数可分解性有限制性假设。不幸的是，这些约束导致了有限的可扩展性。所有代理之间的通信（Sukhbaatar 等，2016；Peng 等，2017）需要大量带宽且会导致延迟，而大量的冗余信息使得难以提取有价值的合作信息。尽管最近有大量的工作致力于简洁且鲁棒的通信，但这仍然需要具有足够带宽的瞬时通信通道，这在分散式控制中难以部署。基于协作图的方式通过构建有向或无向的协作图来捕捉智能体之间的相互作用，但静态和密集协作图会在动态的环境中失效，且会导致大量的无效消息传递，而如何学习动态和稀疏协调图仍是一个亟待解决的问题。

虽然CTDE范式取得了重大进展，然而为了执行这样的策略，代理需要访问全局状态，以使它们能够聚合他们的个人观察。这些要求在实践中通常不成立。并且在具有随机转移函数的部分可观察环境中，使用集中培训学习到的完全分散的策略可能会由于非平稳性而失效，从而导致不协调问题。

本质上最类似于人类合作方式的协作模型是完全分散的模型，其中代理的策略是独立优化的，即independent learning（IL）。然而，即使在简单的合作任务Foerster et al. (2017) and Omidshafiei et al. (2017) 中，分散模型也表现不佳。在分散的多智能体强化学习环境中难以进行协作的一个关键问题在于智能体没有意识到其他智能体的存在，也即不能根据其他智能体的行动来调整自己的策略。在某些情况下，人类使用他们自己的心理过程来模拟他人的行为通过采用他们的观点（Gordon, 1986; Gallese &Goldman, 1998）。这使他们能够理解他人的意图或动机，并在社会环境中采取相应的行动，进而形成协作关系。事实上，人类善于从观察到的行为中进行推断。并本能地进行这种推断，从而阅读行为背后的额外含义。让我们想象一对夫妇在舞厅里跳舞。为了正确有效地表演，每个舞者不仅要能对对方的动作做出反应，还必须预测舞伴的动作。此外，两个舞伴必须有共同的目标和意图，以便有效地完成联合行动。这即是人类经常进行的心智理论，也被称为 "心理状态推理"。（Baker等人，2017年）。这种推理在预测外部事态和内部心态之间的关系，以及这包含了哪些关于私人观察分布的信息。

为了在合作多智能体环境中利用这种范式，我们首先提出了一个新颖的框架Generative Intention Network(GIN)来挖掘动作本身所包含的信息（意图），即一个智能体决定执行某个特定动作而不是另一个动作的事实，这对于发现好的策略至关重要。其次，不是每个智能体都能提供有用的信息，多余的信息甚至会损害协作。受到因果推断(CL)的启发，我们认为智能体需要对与谁协作有一种信念，即智能体更有可能与那些可能对其策略施加更多影响的人合作。具体地说，导致代理策略发生较大变化的其他代理的意图被认为是有影响力的，并被反馈给代理的策略网络和价值网络。这相当于增加他们的意图之间的互信息，并对意图进行筛选。筛选后的意图可以改善智能体的环境模型。当这些额外信息为智能体提供关于其尚未访问的状态空间部分的信息时，它们会产生巨大的影响。这些信息可以用来使智能体探索状态空间中最有希望的区域，从而降低探索成本并大大加快收敛速度。最重要的是，智能体间的迭代推理和决策改进会不断促进他们间的协作。令人震惊的是，即使不存在显式地联盟生成算法，我们的方法也可能导致一种自发的现象：动态团队组成。此外，非本地实体之间的结构化关系和迭代推理将使智能体能够捕捉到复杂问题解决所需的高阶信息。我们根据 MA​​DDPG (Lowe et al., 2017a)、IPPO (de Witt et al., 2020b) 和 MAPPO (Yu et al., 2021) 等强基线，在 StarCraftII 和 Multi-Agent MuJoCo 的基准上评估 本文提出的方法；结果清楚地证明了independent learning即使在复杂任务中不采用CTDE范式，且无法进行显式通信或共享动作/观察，也能实现协调行为。

# II. Related work

在多代理强化学习（MARL）中，实现代理之间的成功协调是一项重要而具有挑战性的任务，人们已经做出了一些尝试来解决这个问题。独立Qlearning（IQL Tan，1993）通过将每个代理建模为独立的Q-learner，分散了代理的策略。然而，由于其他代理改变了他们的策略，从单个代理的角度来看，每个代理都将其他代理视为其环境的一部分，任务变得不稳定。因此，许多工作诉诸于集中训练和分散执行（CTDE）的框架，它假设在训练阶段所有系统信息均可用和在执行阶段的分布式策略。基于该框架，现有的关于代理协调的 MARL 研究大致可以分为三类：

第一种方法确定每个代理在联合奖励中的角色，然后以某种方式将其份额从中分离出来。这类算法称为值函数分解。VDN假设联合动作价值函数可以加法分解为 N 个代理的 N 个 Q 函数。QMIX (Rashid et al., 2018) 通过超网络保证联合Q函数与单独Q函数间的关系是单调的。QTRAN [31] 进一步消除了 QMIX 中对非负权重的约束，并为联合 Q 值提供了一般分解。然而，如何根据分解很好地捕捉复杂的智能体间协调以及如何学习这些分解仍是亟待解决的问题。与该类方法不同，我们的方法不需要对联合价值函数的可分解性进行任何限制性假设。

第二类方法通过学习通信协议来增强多智能体的协作能力。“A mathematical theory of cooperative communication”: 将合作通信解释为最佳运输问题，并将部分先前提出的合作通信模型规约为它的特例。“Bi-Level Actor-Critic for Multi-Agent Coordination”: 从博弈论角度将协作问题定义为具有SE均衡的博弈，并提出了双级行为体批评学习方法，然而他们假设智能体间的决策具有优先级顺序。“COMMUNICATION IN MULTI-AGENT REINFORCEMENT LEARNING: INTENTION SHARING”: 提出了一种基于多个代理之间的共享意图的通信协议。“Learning to Ground Multi-Agent Communication with Autoencoders”： 提出了一个通过自动编码建立多智能体通信的框架，假设所有代理具有相同的模型架构。“Learning Multi-Agent Communication through Structured Attentive Reasoning”: Rangwala 和 Williams 提出SARNet，其中代理在执行操作之前提取其他代理信息的相关性并根据接收到的通信和过去的记忆进行推理。最近有大量工作考虑有限带宽的情况。“Learning Efficient Multi-agent Communication An Information Bottleneck Approach”: 证明了当通信容量受到限制时，有限的带宽约束需要低熵消息，并开发了IMAC来学习有效的通信协议和调度。“Learning Individually Inferred Communication for Multi-Agent Cooperation”: 为了减少信息冗余，提出了I2C，以使代理能够通过因果推理学习代理-代理通信的先验。 “Neurosymbolic Transformers for Multi-Agent Communication”: 提出了一种为多智能体系统的分散控制综合程序化通信策略的方法，显著减少通信量。“Succinct and Robust Multi-Agent Communication With Temporal Message Control”: 提出了TMC，提供显着降低的通信开销和更好的传输丢失鲁棒性。我们的方法与前述方法的一个区别是，我们使用推断来协作，而不是通讯。

第三类方法从协调图的角度对协调行为进行建模。多Agent环境被建模为一个图。每个代理是一个节点，代理的局部观察的编码是节点的特征，一个节点和它的每个邻居之间有一条边。Graph-Based Coordination Strategy for Multi-Agent Reinforcement Learning：引入一个图生成器和基于图的协调策略，以动态地表示基本的决策依赖结构。Deep Coordination Graphs：DCG通过将所有代理人的联合价值函数按照协调图分解为代理人对之间的报酬，在代表能力和概括性之间进行了灵活的权衡。Context-Aware Sparse Deep Coordination Graphs：然而DCG专注于预先定义的静态和密集的拓扑结构。研究如何学习动态稀疏协调图，使用成对报酬函数的方差作为指标来选择边缘。GRAPH CONVOLUTIONAL REINFORCEMENT LEARNING：DGN提出图卷积强化学习适应多代理环境的底层图的动态并捕捉代理之间的相互作用。

除了以上三类方法，“Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning”、 “Multi-Agent Common Knowledge Reinforcement Learning”、“Coordination Between Individual Agents in Multi-Agent Reinforcement Learning”、 “Learning to Simulate Self-Driven Particles System with Coordinated Policy Optimization”也研究了多智能体协调算法，然而它们均属于CTDE范式，且部分需要修改奖励函数或仅考虑邻域间的协调。不同于CTDE范式，本文提出的方法延续IL范式，引入隐式合作(“Real-World Human-Robot Collaborative Reinforcement Learning”)的方法来稳定IL并学习协作行为。我们的方法主要包含两个部分：意图生成网络和因果推断。“Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning”、 “Learning Individually Inferred Communication for Multi-Agent Cooperation”和我们相似，都使用因果推断，但他们要么需要修改奖励函数，要么仍需要显式的通讯协议。 除此之外，“Learning to Share and Hide Intentions using Information Regularization”同样也对智能体意图进行建模，然而它仍属于CTDE范式且需要进行奖励塑造。同样没有显式通道进行通信的工作还有“Learning to Communicate Implicitly by Actions”，我们均没有预定义的显式通信协议，但我们无需有监督训练和任何其他修改。相反的，我们通过意图建模来指导状态估计和策略更新，且仅根据观察到的历史行为轨迹进行意图更新，这样产生的协作机制在训练和执行阶段是一致的。

# III. Preliminaries

在本节中，我们首先介绍了协作 MARL 问题。然后我们回顾条件生成模型，最后介绍一个目前最先进的 MARL 算法。

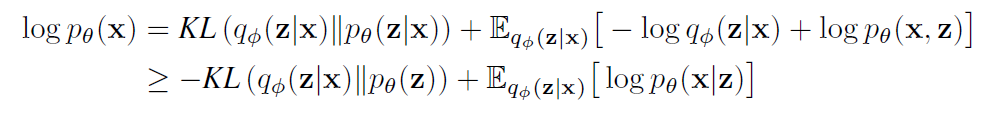
## A. 问题描述

我们考虑分散的部分可观测马尔科夫决策过程（Dec-POMDPs），通过元组hN;O;A;R;P;i来定义。其中，N=f1;::;ng是代理人的集合，O=Qn i=1 Oi是代理人的局部观测空间的乘积，即联合观测空间，A=Qn i=1 Ai是代理人的行动空间的乘积，即联合行动空间，R：O A ! [Rmax;Rmax]是联合奖励函数，P : O AO ! R是转移概率函数，2 [0; 1]是折扣系数。在时间步骤t 2 N，每个代理人i 2 N观察到个人观察值oi t 2 Oi (2)，并根据其策略i采取行动ai t，这是代理人联合策略的第i个组成部分，即。环境执行联合动作，并根据转移函数和奖励函数分别生成下一时刻的状态和即时奖励。所有智能体共享一个联合奖励R(ot; at)，并观察到ot+1，其概率分布为P(jot; at)。所有代理协调以最大化累积折扣收益 E𝝉∼𝜋 ��∞ 𝑡=0 𝛾𝑡𝑟 (𝑠𝑡, 𝒖𝑡)

## B. (条件)生成模型

变分自动编码器 (VAE) [16, 24] 是一个有向图模型，具有某些类型的潜在变量，例如高斯潜在变量。 VAE的生成过程如下：一组潜在变量z由先验分布p(z)生成，数据x由以z为条件的生成分布p(xjz)生成：z p (z)； x p (xjz)。

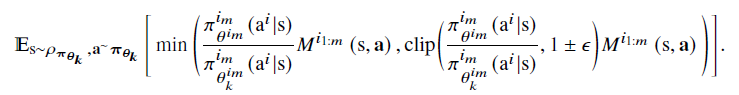
.通常，由于难以处理的后验推断，有向图模型的参数估计通常具有挑战性。然而，可以在随机梯度变分贝叶斯 (SGVB) [16] 框架中有效地估计 VAE 的参数，其中对数似然的变分下限用作代理目标函数。变分下界写为：



conditional variational auto-encoder1 (CVAE) 通过使用辅助协变量调节生成模型来扩展标准 VAE 框架。CVAE由多个MLP组成，如识别网络q(zjx;y)、(条件)先验网络p(zjx)和生成网络p(yjx;z)。该模型的变分证据下限目标可以通过调节包含协变量 x的等式1的概率来获得。

## C. Heterogeneous-Agent Proximal Policy Optimisation (HAPPO)

Heterogeneous-Agent Proximal Policy Optimization (HAPPO) [14] 目前是 SOTA 算法之一，它充分利用多智能体优势分解定理 来实现具有单调改进保证的多智能体信任区域学习。在更新期间，代理随机选择一个排列 i1:n，然后按照排列中的顺序，每个代理 im 选择 im new = im 最大化目标.



# IV. MULTI-AGENT Coordination LEARNING Via inference

在一个未知的合作环境中，每个代理意识不到其他实体的存在，从而将其他代理视为其环境的一部分。此外，由于对其他代理的意图一无所知，智能体并不知道与哪些智能体协作，这使得代理很难在非平稳的环境中学习协调策略。因此，本文的模型包含一个生成式意图网络，它学习智能体意图的表示，以及因果推断，它考虑强关联意图的智能体。

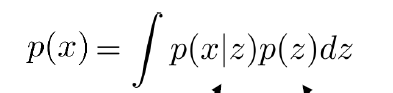
## A. Generative Intention Network

在部分可观测环境中，每个智能体只能通过自己的观察进行行动，无法感知全局状态。为了缓解部分可观察性，我们将其他智能体的意图与当前个人观测结合，以重建或部分重建全局状态，因为意图中包含了关于他们观察分布的信息。

那么要问的关键研究问题是：

我们如何合理地建模智能体的意图？

人类合作者具有社会感知方面的能力，被定义为根据基本行为信号编码另一个人的心理状态的能力。此外，当人们理解他人的行为时，还有一种机制在很大程度上依赖于感知者的自我认知：为了理解他人的行动，感知者必须首先观察该行动，然后形成对该行动的心理模仿，通过镜像神经元网络在观察和模仿的行动之间进行比较。

根据以上理论，我们提出了一个有条件的深度生成模型根据历史动作轨迹进行意图建模，从而表示人类的社会感知能力。并为每个代理人编码潜变量来代表代理的意图。即可表示为：

在每个时长为T的幕中，我们有N个代理。用A 属于 R NxT表示这个场景中的所有历史动作数据，ati表示第i个代理人在第t步的动作。我们将A分为过去历史动作和未来历史动作，即根据给定Tsep <T将A分为两部分。我们有过去历史动作轨迹AP = A1:Tsep，任务是根据过去历史动作轨迹，生成智能体意图后还原未来历史动作轨迹AF = A(Tsep+1):T，即心理模仿。GIN可以通过最大化证据下限（ELBO，[35]）来训练，ELBO通常被用作难以解决的对数似然的替代物。在训练过程中，当收集到一整幕的动作轨迹后，我们将数据分为过去和未来部分，我们使用给定AP和AF的编码器推断潜变量，并用Adam优化ELBO。在执行阶段，我们不需要编码器，而是使用给定AP的先验模型来建模意图。我们的模型GIN，如图1所示。我们在下面的小节中描述细节。

* 1. **Encoder**

人类具有时间整合的能力，即构建和整合一段时间内信息的过程和能力，使之成为一个连贯的整体，从而能够理解和预测随着时间推移发生的事件。因此，我们沿时间维度串联XP和XF，并使用递归神经网络（RNN）来计算每个代理的全长轨迹的特征，并使用层归一化进行特征正则化。具体地，编码器以AP和AF为输入，计算出一个近似的后验q(z|XP;XF)。我们假设z遵循多变量高斯分布N(; I)，利用重参数化技巧[31, 16]，我们对N(0; I)进行采样，得到z = + ，其中和是由多层感知器（MLP）预测的。

综上，编码器将基本行为信号编码另一个人的心理状态。

* 1. **Prior**  
     在执行阶段，智能体只能从记录的历史动作轨迹进行意图建模，因此，先验p(zjXP)表征了仅给定AP的z的条件分布。考虑到先验与编码器仅在模型输入数据上有区别，即RNN具有不同序列长度的输入，因此我们将先验与编码器共享模型权重。从另一个角度看待公式1，我们发现优化ELBO可以最小化分别从先验和编码器提取的信息间的KL散度。因此，模型被鼓励学习从AP和[AP;AF]中编码相同的信息。
  2. **Decoder**  
     解码器形成对行动的心理模仿，并在观察和模仿的行动之间进行比较。对互动伙伴的预测和实际行动进行比较，可以帮助个人调整自己的行动，以适应互动伙伴的行动。

解码器p(AF|z;AP )产生以随机变量z和过去轨迹AP为条件的未来轨迹。按照[18]，我们假设潜伏的z是对动作历史的充分总结，并将似然定义为p(At+1|z;A1:t) = p(At+1|z;At)。因此，给定z和At，解码器可以预测^At+1。在这里，我们采用预测下一个动作值与当前动作值之间差异的函数，即ai t+1 ai t，而不是下一个动作值ai t+1，以减少学习早期阶段的模型偏差。

## B. Casual Inference

一方面，正如DMS所表明的那样，社会互动需要对他人的心理状态进行神经表征，并对合作伙伴的意图具有相应的信念。此外，交互的每个代理都有自己的DMS并且会及时调整。然而，只有当智能体的神经过程同步时，才能调整DMS和行为策略。大脑间的同步化是人际行动协调和社会互动行为的重要和必然的机制。因此，我们认为感知智能体除了理解他人的行为外，还应通过理解他们的意图对感知者的策略的因果效应来将自己的策略与伙伴策略成功匹配。直观地说，一个代理人更有可能与那些可能对其策略施加更多影响的实体进行协作，希望获得关于他们倾向于如何行动以及如何做出合作反应的线索。因此，其他代理人的因果效应可以被看作是以其他代理人的意图为条件进行决策的必要性。

另一方面，假设我们已得到所有合作伙伴的意图，那么在此有理由提出这样的问题：如果感知者已经跟踪和预测所有互动伙伴的意图，为什么还需要推断意图间的因果关系，而不是直接将所有意图整合？在我们看来，简单的意图整合会导致信息冗余，甚至可能影响学习过程，而因果推断则将使代理能够捕获高阶关系，甚至形成有共同的目标和意图的动态即时团队。

这里，我们通过以不同意图为条件生成的策略分布间的KL散度来量化代理人之间的因果效应：

\documentclass{article}
\usepackage{amsmath}
\pagestyle{empty}
\begin{document}

$ KL\left [ \pi \left ( a\mid o,[z_{intention}^m, z_{intention}^n, \cdots z_{intention}^k] \right )\parallel \pi \left ( a\mid o,[z_{intention}^n, \cdots z_{intention}^k] \right ) \right ]  $


\end{document}

Kullback-Leibler（KL）散度被用来衡量这两个条件概率分布之间的差异。Ij i的大小表明，如果不考虑代理人j的意图，代理人i会对其策略做出多大的调整，也表明代理人j的战略与代理人i的策略有多大的关联。如果Ij i>阈值δ，则智能体间为强关联，并将该智能体加入到协作集中，否则为弱关联，其中delta是一个超参数。阈值G设定为0.5，实验中分析了不同阈值G的性能。

每个智能体根据部分其他智能体的意图进行决策，产生的新的决策被智能体感知后，意图也会随之更新，从而进一步更新策略，这样便形成了无限嵌套的信念。

## C. Long Short-Term Intention Clip

我们工作中的其中一个主要技巧是使用裁剪的意图重要性比率--这试图限制意图在迭代之间的急剧变化。剪切的强度由超参数控制：大的允许更大的意图变化。

我们认为对意图进行剪裁，在一定程度上控制了由变化的多代理政策引起的非平稳性。如第5.4节所示，我们观察到，虽然较低的数值减缓了学习速度，但它们对应的是更一致的政策改进。另一方面，较高的数值会导致较大的方差和较大的性能波动。



我们工作中的其中一个主要技巧是使用长短周期的意图裁剪。智能体的探索可能会产生不符合宏观意图的动作从而干扰意图建模。此外，在竞争情形下，对手可能在短时间内做出假动作（与长期意图相反）。因此，为了学习更稳定的意图，我们提出了一个短期意图来消除干扰性动作。

短期意图网络与GIN中**Prior**网络共享参数，和先验只是在公式1中的RNN的输入数据方面有所不同。具体地，先验使用整个历史动作轨迹作为输入，而短期意图网络使用长度为n的动作轨迹作为输入，长度n设定为3，实验中分析了不同n的性能。根据先验和短期意图网络输出的意图，我们比较了他们的KL散度，如果KL散度大于delta，则将n个动作进行masking，否则保留。

## D. Action Mask

在多智能体环境中，每个单元由一个独立的代理控制，该代理只对以该单元为中心的有限视域内的局部观测进行控制(如，smac),智能体不可能在每个时刻都可以观察到其他所有智能体的动作，因此在智能体观察的动作轨迹中会存在缺省值。对这些缺省值进行意图建模，无疑是毫无意义的，甚至会对策略训练造成巨大影响，我们的建议是简单地使用一个特定于代理人的常数向量，即一个带有代理人ID的零向量，作为代理人动作缺省值的输入。我们称这种技术为 "动作屏蔽"。

## E. HAPPO with Intention Ratio

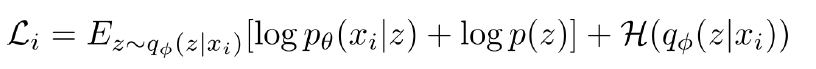
TRUST REGION POLICY OPTIMISATION IN MULTI-AGENT REINFORCEMENT LEARNING开发了异质代理近端政策优化（HAPPO）算法。与许多现有的MARL算法不同，HATRPO/HAPPO不需要代理共享参数，也不需要对联合价值函数的可分解性进行任何限制性假设。最重要的是，他们在理论上证明了HAPPO的单调性改进特性，建立了新的技术水平。

我们在HAPPO的基础上，引入了意图重要性比率，这使得每个代理人的目标都不仅考虑到所有先前代理人的更新，还考虑到意图的更新。

## F. Training

* 1. GIN

对于训练，我们从编码器中取样z = [zI]，q(zjXP ;XF)，然后优化ELBO：



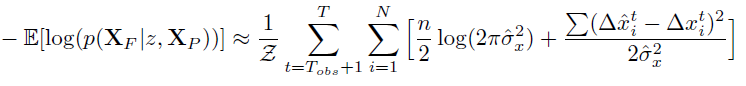
可以从另外一个角度看待ELBO：



最大化对数似然的同时，尽量缩小编码模型和先验分布的KL divergence



由于我们的解码器还原代理人i在步骤t时的动作，并且它遵循高斯分布。我们用样本平均数对期望值进行近似。



其中Z是归一化项，n是xti的维度。由于先验和编码器的输出是高斯的，所以第二个项可以用分析法计算。



优化的目标为最大化ELBO：

由于正在更新的GIN 也用于计算策略梯度和td误差，因此策略和价值网络的更新容易发散。我们的方案是使用软更新，即通过 polyak 平均更新GIN网络：

此外，我们使用几个训练技巧：

延迟更新：不准确的意图估计可能会导致糟糕的Critic和策略更新，这特别有问题，因为会产生一个反馈回路，错误的意图导致Critic产生有偏差的状态估计，从而造成策略朝着错误地方向更新，而有偏差的意图也会直接对策略改进造成影响，进而改变后续的意图更新。因此，我们引入TD3中的延迟策略更新，策略网络的更新频率低于价值网络和意图网络，使得策略在价值误差和意图误差尽可能小时更新，即价值网络和意图网络更新d次后再更新策略网络。

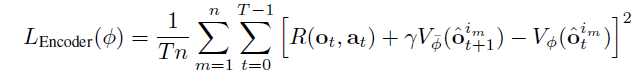
首次更新：在ppo-epoch的内循环中对意图进行优化，计算量很大。相反，我们仅在首个ppo-epoch更新意图，而在其他ppo-epoch对意图进行微调。这样做的结果是，只要意图的变化足够慢，策略和Critic就会被维持在其最优解附近。这种策略类似于GAN的训练方式。

预训练：在训练早期，智能体策略十分随机，导致意图更新很不稳定。因此我们在前100episodes进行预训练，即不考虑意图，直到预训练结束，我们开始意图更新。

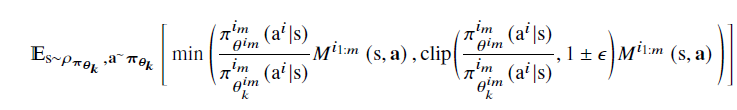
详细伪代码在附录中列出。

* 1. HAPPO：

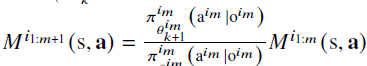
价值网络通过最小化经验贝尔曼误差更新：



通过最大化PPO-Clip的目标更新策略网络：







A^t 是智能体i以个人观察及其他智能体意图为条件下的优势函数的估计。我们使用广义优势估计 (GAE) [24]作为价值函数的稳健估计器

# V. EXPERIMENTS AND RESULTS

## 我们考虑两个最常见的基准--StarCraftII多Agent挑战赛（SMAC）（Samvelyan等人，2019）和多Agent Mujoco（de Witt等人，2020b）--用于评估MARL算法。所有的超参数设置和实现细节都可以在附录中找到。

星际争霸II多Agent挑战（SMAC）。SMAC包含一组《星际争霸》的地图，其中

一组盟友单位旨在击败对手的团队。IPPO（de Witt等人，2020a）和MAPPO

(Yu et al., 2021)被认为在这个基准上取得了最高的成绩。通过采用参数

共享，这些方法在大多数地图上实现了100%的获胜率，甚至包括有异质代理的地图。

有异质性代理的地图。因此，我们假设，不一定需要非参数共享，共享政策就足以解决

需要，共享政策足以解决SMAC任务。我们在两张硬地图和一张超硬地图上测试了我们的方法。

图2的结果证实，SMAC的难度并不足以展示HATRPO的能力。

与参数共享方法相比，SMAC的难度不足以显示HATRPO/HAPPO的能力。

多Agent MuJoCo。与SMAC相比，我们认为Mujoco环境为我们的方法提供了一个更为

为我们的方法提供了更合适的测试案例。MuJoCo任务要求机器人学习一种最佳的运动方式。

多Agent MuJoCo将机器人的每个部分都建模为一个独立的代理，例如，蜘蛛的腿或手臂。

例如，蜘蛛的腿或游泳者的手臂。随着身体各部分的种类越来越多，建立模型

异质性的政策变得很有必要。图3表明，在所有情况下，HATRPO和

HAPPO享有比参数共享方法更优越的性能。IPPO和MAPPO。

并且在奖励值和方差方面都优于非参数共享的MADDPG（Lowe等人，2017b）。

奖励值和方差。还值得注意的是，HATRPO和它的竞争对手之间的性能差距，随

其竞争对手之间的性能差距随着代理数量的增加而扩大。同时，我们可以看到，HATRPO

在几乎所有的任务中都优于HAPPO；我们认为这是因为HATRPO的硬KL约束。

我们认为这是因为HATRPO中的硬KL约束与HAPPO中的剪切版本相比，更接近于算法1，而算法1可以达到

单调性改进的保证。

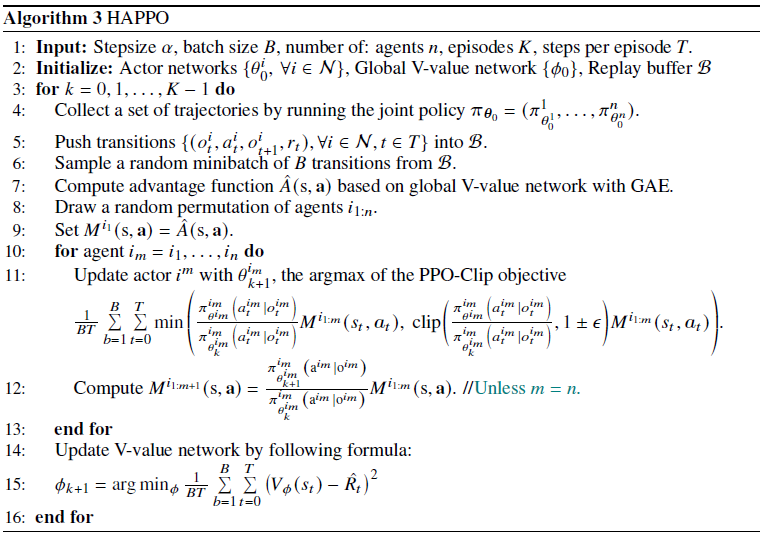
# VI. Conclusion

## 在本文中，我们提出了一种基于推断的独立学习MARL 方法来解决合作部分可观察多代理环境中的协调问题。这与采用CTDE范式的工作不同。为了缓解非平稳性问题，我们首先引入了DMS来回答如何让智能体意识到他人存在的问题，它允许代理仅根据个人的运动系统推断他人意图而无需任何其他修改。如何选择合作伙伴形成协作团队的问题通过代理之间的因果关系来回答，它准确地捕捉到了协作的必要性，并形成动态团队。离散和连续控制任务（即 SMAC 和 Multi-Agent Mujoco）的实验结果表明，在复杂任务中，即使不采用CTDE范式也可以学习到智能体间的协调行为。我们希望我们的工作能够将 MARL 扩展到具有复杂协调动态的更现实的任务。

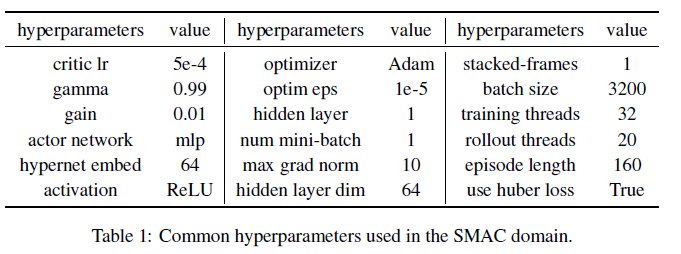
但是，我们强调了我们工作的两个限制。首先，我们的方法假设在协调稳定后，代理的个人观察结合他人意图可以部分还原全局状态。其次，我们的方法相比最先进的MARL方法需要更多的训练步。我们相信，测试这些约束的极限并放宽它们，将是未来工作的重要步骤。

Appendix

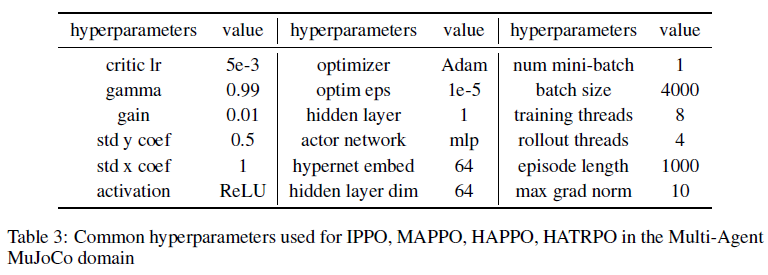
伪代码

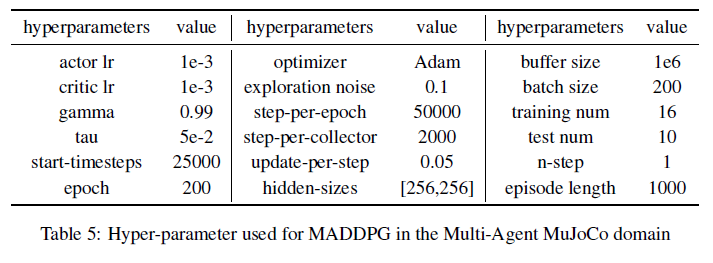


用于实验的超参数设置



MADDPG 的实现采用天寿框架（Weng et al., 2021），所有超参数保持原点最佳性能状态不变





References

*Basic format for periodicals:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, pp. xxx-xxx, Abbrev. Month, year, doi: 10.1109.XXX.1234567.

*Periodicals using article numbers:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, Abbrev. Month, year, Art. no. xxxxx, doi: 10.1109.XXX.1234567.

*Examples:*

# J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, doi: 10.1109/TED.2016.2628402.

1. E. P. Wigner, “Theory of traveling-wave optical laser,” *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.
2. P. Kopyt *et al., “*Electric properties of graphene-based conductive layers from DC up to terahertz range,” *IEEE THz Sci. Technol.,* to be published, doi: 10.1109/TTHZ.2016.2544142. *(Note: If a paper is still to be published, but is available in early access, please follow ref [5]).)*
3. R. Fardel, M. Nagel, F. Nuesch, T. Lippert, and A. Wokaun, “Fabrication of organic light emitting diode pixels by laser-assisted forward transfer,” *Appl. Phys. Lett.*, vol. 91, no. 6, Aug. 2007, Art. no. 061103.
4. D. Comite and N. Pierdicca, "Decorrelation of the near-specular land scattering in bistatic radar systems," *IEEE Trans. Geosci. Remote Sens.*, early access, doi: 10.1109/TGRS.2021.3072864. (*Note: This format is used for articles in early access. The doi must be included.)*
5. H. V. Habi and H. Messer, "Recurrent neural network for rain estimation using commercial microwave links," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3672-3681, May 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9153027

*Basic format for books:*

J. K. Author, “Title of chapter in the book,” in *Title of Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. x, sec. *x*, pp. xxx–xxx*.*

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.
3. Philip B. Kurland and Ralph Lerner, eds., *The Founders’ Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010, [Online]. Available: http://press-pubs.uchicago.edu/founders/

*Basic format for handbooks:*

*Name of Manual/Handbook, x* ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. xxx-xxx.

*Examples:*

1. *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.
2. *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.
3. R. J. Hijmans and J. van Etten, “Raster: Geographic analysis and modeling with raster data,” R Package Version 2.0-12, Jan. 12, 2012. [Online]. Available: http://CRAN.R-project.org/package=raster

*Basic format for reports:*

J. K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. xxx, year.

*Example:*

1. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

*Basic format for conference proceedings:*

J. K. Author, “Title of paper,” in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. xxxxxx*.*

*Examples:*

1. D. B. Payne and J. R. Stern, “Wavelength-switched passively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA,1985,   
   pp. 585–590.
2. D. Ebehard and E. Voges, “Digital single sideband detection for interferometric sensors,” presented at the 2nd Int. Conf. Optical Fiber Sensors*,* Stuttgart, Germany, Jan. 2-5, 1984.
3. PROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

*Basic format for electronic documents (when available online):*

Issuing Organization. (year, month day). *Title*. [Type of medium]. Available: site/path/file

*Example:*

1. U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

*Basic format for patents:*

J. K. Author, “Title of patent,” U.S. Patent *x xxx xxx*, Abbrev. Month, day, year.

*Example:*

1. G. Brandli and M. Dick, “Alternating current fed power supply,” U.S. Patent 4 084 217, Nov. 4, 1978.

*Basic format**for theses (M.S.) and dissertations (Ph.D.):*

J. K. Author, “Title of thesis,” M.S. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

J. K. Author, “Title of dissertation,” Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

*Examples:*

1. J. O. Williams, “Narrow-band analyzer,” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.
2. N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

*Basic format for the most common types of unpublished references:*

J. K. Author, private communication, Abbrev. Month, year.

J. K. Author, “Title of paper,” unpublished.

J. K. Author, “Title of paper,” to be published.

*Examples:*

1. A. Harrison, private communication, May 1995.
2. B. Smith, “An approach to graphs of linear forms,” 2014, *arXiv:2105.02824*.
3. A. Brahms, “Representation error for real numbers in binary computer arithmetic,” IEEE Computer Group Repository, Paper R-67-85.

*Basic formats for standards:*

a) *Title of Standard*, Standard number, date.

b) *Title of Standard*, Standard number, Corporate author, location, date.

*Examples:*

1. IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969.
2. Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

**First A. Author** (Fellow, IEEE) and all authors may include biographies if the publication allows. Biographies are often not included in conference-related papers. Please check the Information for Authors to confirm. Author photos should be current, professional images of the head and shoulders. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with the type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author’s major field of study should be lowercase. 

The second paragraph uses the preferred third person pronoun (he, she, they, etc.) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. The format for listing publishers of a book within the biography is: *Title of Book* (publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author’s preferred title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter, Mx. Riley). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications.

**Second B. Author**, photograph and biography not available at the time of publication.

**Third C. Author, Jr.** (Member, IEEE), photograph and biography not available at the time of publication.

1. This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE. It is IEEE style to display support information, including sponsor and financial support acknowledgment, here and not in an acknowledgment section at the end of the article. For example, “This work was supported in part by the U.S. Department of Commerce under Grant BS123456.” The name of the corresponding author appears after the financial information, e.g. *(Corresponding author: M. Smith).* Here you may also indicate if authors contributed equally or if there are co-first authors.

   The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, First A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

   Second B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

   Third C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba 305-0047, Japan (e-mail: author@nrim.go.jp).

   Mentions of supplemental materials and animal/human rights statements can be included here.

   Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org [↑](#footnote-ref-1)