比特预算信道上的多智能体系统集中式控制的任务有效观测压缩

Arsham Mostaani，*IEEE*学生会员，Thang X. Vu，*IEEE*高级会员，Symeon Chatzinotas，*IEEE*会士，以及 Bj¨orn Ottersten，*IEEE*会士

arXiv:2301.01628v1 [cs.IT] 4 Jan 2023

摘要—**本文研究由中央控制器（CC）控制多个智能体时所产生的任务有效量化问题。智能体需要将观测传送给 CC 以供决策，但智能体与 CC 之间链路的比特预算受限通信可能会限制系统的任务有效性，该有效性以系统阶段成本/奖励的平均总和衡量。因此，每个智能体都应压缩/量化其观测，使控制任务阶段成本/奖励的平均总和受到的影响最小。本文提出两种不同的基于动作的状态聚合（ABSA）算法，以间接方式和联合设计方式实现多智能体系统中的控制策略与通信策略设计，从而解决最大化阶段奖励平均总和的问题。ABSA-1 的适用范围仅限于单智能体系统，但它提供了一个分析框架，可作为设计 ABSA-2 的基础。ABSA-2 则对多智能体系统的控制与通信进行联合设计。我们通过求解多智能体几何一致性问题所进行的数值实验，采用平均回报作为性能指标，对这些算法进行了评估。最后，我们通过引入一项衡量多智能体系统通信有效性的新指标，总结数值结果。**

索引词—**语义通信、任务有效数据压缩、面向目标的通信、面向机器学习的通信、多智能体系统、强化学习。**

I. 引言

随着 5G 的推广，物联网（IoT）、工业物联网（IIoT）和自动驾驶汽车等一系列新应用不断涌现。据预测，到 2030 年，约有 300 亿台物联网设备将实现连接 [1]。随着非人类类型联网设备的不断普及，通信设计的重点正从传统性能指标（例如比特错误率和通信时延）转向语义和任务导向的性能指标，例如含义/语义错误率 [2]，[3] 以及信息及时性 [4]。

传统上，人们可以通过测量网络的总速率来评估网络资源的利用效率；而在信息物理系统时代，鉴于网络的资源约束，我们希望了解如何以期望的方式有效执行一项或多项任务 [5]，[6]。我们正见证通信系统领域的一次范式转变：传统系统所针对的性能指标已不再适用。这给如何设计通信以最终实现任务有效性带来了新的重大挑战 [6]。

作者来自卢森堡大学安全、可靠性与可信中心，卢森堡。电子邮箱：{arsham.mostaani、thang.vu、symeon.chatzinotas、bjorn.ottersten}@uni.lu。本工作由欧洲研究委员会（ERC）通过 AGNOSTIC 项目（资助协议编号：742648）支持。

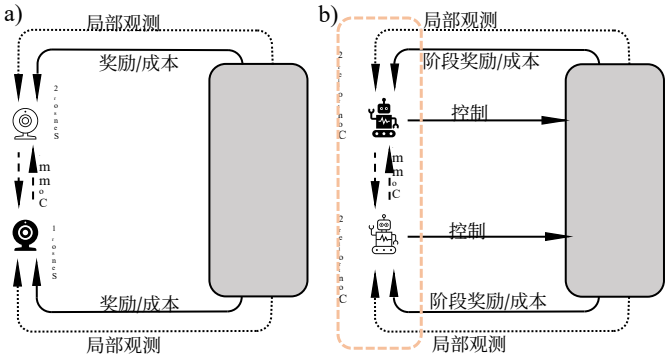


图 1。任务有效通信：a）估计任务与 b）控制任务——橙色虚线框的详细内容见图 2 和图 3。

这项研究也部分受到多智能体系统中以“涌现式通信”为名的新型机器学习技术和算法取得成功的推动 [7]。预计将这些新技术和理念引入通信工程，将对通信系统设计的多个领域产生颠覆性影响。

根据香农和 Weaver 的观点，通信问题可分为三个层次 [8]：（i）技术问题：在给定信道和网络约束的情况下，通信符号/比特能够以多高的准确度传输？（ii）语义问题：在给定信道和网络约束的情况下，通信符号能够以多高的准确度传达所需含义？（iii）有效性问题：在给定信道和网络约束的情况下，通信符号能够以多高的准确度帮助完成所需任务？传统的通信设计主要解决技术问题，而近年来，语义问题 [2], [3], [5], [9], [10] 以及有效性问题 [6], [11]–[18] 已引起广泛的研究兴趣。

与香农的技术层通信框架不同，语义通信可以利用信源与目的地之间的先验知识来提升性能 [4], [19]。然而，基于语义的设计未必具有任务有效性 [20]。人们可以设计发射机，在尽可能少地损害所传输语义的前提下压缩数据 [2], [3] ，但传输过程可能并不考虑任务 [21]。与语义层和技术层通信设计不同，任务有效通信系统的性能最终以与任务相关的平均回报/成本来衡量 [11]。在（任务）有效性问题中，我们关注的不仅是语义的通信，

还关注消息交换如何帮助接收端提升其在估计任务 [4], [13], [14], [16], [22] 或控制任务 [11], [12], [14], [17], [18], [23], [24]中的预期成本/奖励表现。

估计任务与控制任务的任务有效通信设计存在根本差异——见图 1。（i）在后者中，每个智能体都能生成控制信号，直接影响该智能体的下一次观测。因此，在控制任务中，信息源——智能体的局部观测——通常是具有记忆的随机过程，例如线性过程或马尔可夫决策过程-[11], [17], [18]。然而，在估计任务中，信息源通常被假设为独立同分布随机过程[13], [16], [22]。（ii）在控制任务中，控制信号对系统状态的影响往往会持续较长时间，而非仅限于单个阶段/时间步。例如，某个控制动作可能导致短期内期望奖励降低，但长期来看期望奖励更高。这使得控制任务本质上对设计控制策略所针对的时间范围十分敏感。估计任务则不同，尤其当观测过程为独立同分布时，可以在单个阶段/时间步内完成求解——因为一个阶段/时间步的解不会影响另一个阶段/时间步，也就是说，每个时间步都可以单独求解 [22], [25]。（iii）估计任务的成本函数通常采用差异/失真函数的形式，而控制任务的成本函数则可以有许多其他形式。

本文聚焦于控制任务的有效性问题。具体而言，我们研究多智能体系统（MAS）的分布式通信设计，最终目标是最大化逐阶段奖励总和的期望，即期望回报。多个智能体在中央控制器（CC）的协助下，在多智能体系统中选择控制动作并进行通信，以完成一项协作任务——也就是说，该系统的通信网络拓扑为星型拓扑，其中枢纽节点为中央控制器，外围节点为智能体——见图 2。所考虑的系统架构可应用于多个领域，例如物联网、新兴的网络物理系统、实时交互系统、车联网通信 [26]以及协作感知 [27]。

A. 相关工作：面向控制任务的任务有效通信

文献 [11], [12], [14], [17], [18], [23], [24] 的作者研究了不同设定下的任务有效通信设计。 [12], 针对通过理想通信信道设计应用定制协议这一特定问题，采用了任务有效通信设计；而在 [11], [14], [17],[18], [23], [24]中，通信信道被视为非理想信道。文献 [14] 的作者为单智能体系统中任务有效联合信源信道编码的设计提供了算法方面的贡献。 [11], [14], [17],旨在研究多智能体系统的任务有效联合信源信道编码，而[18], [23]则侧重于任务有效数据压缩和量化。与本文类似，文献中采用了星型拓扑

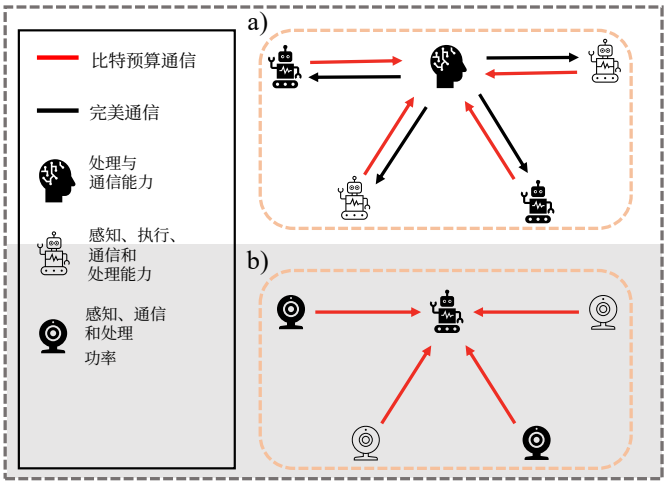


图 2. 通信拓扑及其适用场景：a）具有共置执行器和传感器的多智能体系统集中式控制；b）由单个控制器与单个执行器共置的分布式感知。橙色虚线框详示了图 1 和图 3 中的同一方框。

用于智能体间通信；而 [11], [12]采用枢纽节点与外围节点之间的完美通信， [12] 则假设枢纽节点与外围节点之间通信完美， [11] 假设外围节点的下行链路通信信道非理想。与上述所有工作不同，据我们所知，本文首次研究了上行链路（智能体到枢纽节点）信道非理想（受比特预算限制）时的星型拓扑——见图 2。相应地，每个智能体观测环境，并通过非理想（受比特预算限制）的通信信道，将其局部观测的抽象版本发送给中央控制器——图 2 中的红色链路。随后，中央控制器生成控制动作，并通过理想通信信道将其发送给智能体——图 2 中的黑色链路。中央控制器选择控制动作，以最大化协作任务的平均回报；回报是与任务完成情况相关的性能指标。

B. 贡献

在我们先前的工作 [18]，中，我们提出了一个通用框架，用于解决具有全网状连接的多智能体系统（MAS）的面向任务的通信问题。本工作可视为将该框架应用于一种新的问题设定，以设计任务有效通信，其中智能体采用星形网络拓扑进行连接。沿着这一思路，本工作将所提框架的适用范围扩展到 [18] 中所解决的特定问题之外，并进一步揭示了如何在更广泛的意义上、在范围更广的设定下使用该框架。具体而言，本工作的贡献如下所列。

- 首先，我们考虑一种新颖的问题设定：多智能体系统由中央控制器控制，而中央控制器只能通过比特预算受限的分布式通信获取智能体的局部观测。这种问题设定可以

用于协作感知系统和车联网通信，而此前类似研究所探讨的问题设定无法涵盖这一情形。

- 其次，我们的解析研究建立了所考虑的联合通信与控制设计问题同传统数据量化问题之间的关系。特别是，引理 1 展示了本文所处理的问题如何成为传统数据量化问题的推广版本。这一表述十分有用，因为它有助于在更强的条件下通过 ABSA-1，以及在更宽松的条件下通过 ABSA-2，为该问题找到精确解。

- 此外，我们的解析研究帮助我们设计出一种间接的[1]任务有效数据量化算法——ABSA-2。为 ABSA-2 设计任务有效的数据量化方案，可以等价地转化为一种针对任意深度Q网络进行间接特征选择的方法。基于对 ABSA-1 所开展的分析，ABSA-2 在智能体与中央控制器之间设计了分布式且比特预算受限的通信。通过增加中央控制器的记忆，ABSA-2 被证明能够逐步接近最优性能。事实上，增加中央控制器的记忆会带来更高的计算复杂度。因此，ABSA-2 被认为在计算复杂度与任务效率之间实现了权衡。

- 我们在几何一致性任务上开展数值实验，以评估所提方案在多智能体系统于该任务中的期望回报最优性方面的性能。在具有局部可观测性和比特预算受限通信的多智能体[2] 场景中，我们将 ABSA-1 和 ABSA-2 与 [18], 介绍的若干其他基准方案进行比较。

- 最后，我们将介绍一种名为任务相关信息的新指标，用于衡量面向任务的通信策略的有效性。与正向倾听和正向信号传递等现有指标相比，该指标能更好地解释多种任务有效通信方案的行为。该指标能够衡量面向任务的通信/压缩策略的有效性，而无需测试联合设计的控制策略，也无需在目标任务中测试联合设计的策略。

C. 技术方法

我们的目标是对智能体的局部观测进行高效表示，以确保满足通信链路的比特预算，同时尽量减小其影响任务平均回报的量化。 为此，我们首先需要为所有智能体设计任务有效的数据量化策略。在任务有效的数据量化中，需要考虑平均回报函数的性质以及与该任务相关的最优控制策略 [15]。除了为所有智能体设计量化策略之外，我们还需要确保中央控制器的控制策略能够仅凭量化后的消息做出近乎最优的决策，因此这构成了一个联合控制与数据压缩问题。我们将这一联合控制与数据压缩问题表述为一种广义的数据压缩形式：面向任务的数据压缩（TODC）。基于这一新颖的问题表述，我们提出了两种间接的基于动作的状态聚合算法（ABSA）：(i) ABSA-1 为任务有效量化提供解析证明，即在期望回报方面达到最优性能。为此，ABSA-1 放宽了对底层马尔可夫决策过程可合并性的假设，根据该假设 [18][条件。 6], 所提方法的性能保证得以确立。由于 ABSA-1 仅适用于系统由一个智能体和中央控制器组成的情况，我们还提出了 ABSA-2。在 ABSA-1 解析结果的基础上，借助最大后验估计放宽 ABSA-1 的上述限制，并利用中央控制器处的深度Q网络控制器；我们将介绍 ABSA-2，将其作为一种更通用的方法。(ii) ABSA-2 求解 TODC 问题的近似版本，并为任意数量与中央控制器通信的智能体进行量化。得益于中央控制器所采用的深度Q网络控制器，ABSA-2 能够解决更复杂的问题，在这类问题中，控制器可利用更大的记忆。因此，ABSA-2 能够在复杂度与通信效率之间进行权衡。最后，我们将在特定任务中评估所提方案的性能：有限可观测性下的几何一致性问题 [28]。

D. 文章结构

本文其余部分安排如下。第二节介绍多智能体系统，并陈述联合控制与通信问题。第三节提出两种基于动作的状态聚合算法。第四节展示所提算法在几何一致性问题中的性能。最后，第五节总结全文。为方便读者，表 I 总结了本文采用的符号。粗体表示随机矩阵或标量，其实现值以普通字体表示。

II. 系统模型与问题陈述

我们在此引入的问题设定可用于分析图 2 所示的两种场景。不过，为了统一术语，本文始终聚焦于该图中的场景 (a)。具体而言，我们使用“智能体”一词时，指的是一种必然具备以下全部硬件能力的对象：感知、执行、通信和数据处理。然而，多智能体系统并不一定仅由智能体构成，也可能由智能体与其他对象组合而成，

[1]这里所说的间接算法，是指不依赖于我们对某一特定任务的了解的方法。间接方法适用于任何任务或广泛的任务。与间接方案不同，我们还有专为某一特定应用领域设计的直接方案 [16]。正如 [6] 所定义的：“直接方案旨在通过设计针对特定任务量身定制的通信策略，来保证或提升信息物理系统在该任务中的性能。”

[2]由于第四节中所述的复杂度相关问题，数值结果仅限于双智能体和三智能体场景。

这些其他对象至少具备通信和数据处理能力。这里的中央控制器应具备处理相对大量数据的硬件能力，以及通信能力。多智能体系统内部以及多智能体系统与环境之间的交互如图 3 所示。

A. 系统模型

我们考虑一个多智能体系统，其中多个智能体$i \in \mathcal{N} = \{1, 2, ..., N\}$在中央控制器的辅助下协作完成一项任务。遵循集中式动作策略，中央控制器通过理想通信信道向智能体提供其动作，同时通过不完美通信信道接收智能体的观测 [3]。所考虑的设置类似于传统的多智能体系统集中式控制 [18], [30],，不同之处在于，智能体向中央控制器发送的通信通过比特预算受限的通信信道传输。智能体与中心节点通信被视为瞬时且同步的 [18]。这与具有延迟的 [17], [31] 以及顺序式/迭代式通信模型 [32]–[34]形成对比。需要指出的是，在所考虑的系统中不存在直接的智能体间通信——通信仅发生在智能体与中央控制器之间。系统以离散时间步运行 t。每个智能体 i 在时间步 t 的观测表示为 $\mathrm{o}_i(t) \in \Omega$，系统的状态 $\mathrm{s}(t) \in \mathcal{S}$ 由联合观测定义，即 $\mathrm{s}(t) \triangleq \langle \mathrm{o}_1(t) \ldots, \mathrm{o}_N(t) \rangle$[4]。每个智能体 i 在时间 t 的控制动作表示为 $\mathrm{m}_i(t) \in \mathcal{M}$，系统的动作向量 $\mathrm{m}(t) \in \mathcal{M}^N$ 由联合动作定义，即 $\mathrm{m}(t) \triangleq \langle \mathrm{m}_1(t) ..., \mathrm{m}_N(t) \rangle$。观测空间 Ω、状态空间 $\mathcal{S}$和动作空间 $\mathcal{M}$ 均为离散集合。环境由一个潜在的[5] 马尔可夫决策过程支配

[3]在本研究中，我们遵循网络化控制文献中的一项常见假设 [29]：比特预算仅限制智能体的上行通信，不限制其下行链路。因此，智能体按照中央控制器的指示选择控制动作。

[4]根据此定义，在任意给定时刻，t 任意两个智能体的观测 i，$j \in \mathcal{N}$在欧几里得空间中线性无关。任意智能体的控制动作也满足相同条件。

[5]根据文献中的定义，[10], 底层的 MDP′ 是由一个假想的单个智能体定义的时域为 T' 的 MDP，该智能体采取联合动作 $\mathbf{m}(t) \in \mathcal{M}^N$并观测标称状态 $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t) \ldots, \mathbf{o}_N(t) \rangle$；其状态转移模型 $T(\cdot)$ 和奖励模型 $r(\cdot)$ 与我们的多智能体系统所处环境的相同。

符号	含义
$\mathbf{x}(t)$	时刻 t生成的通用随机变量
$\mathrm{x}(t)$	$\mathbf{x}(t)$的实现值
$\mathcal{X}$	$\mathbf{x}(t)$的字母表
$\lvert\mathcal{X}\rvert$	$\mathcal{X}$的基数
$p_{\mathbf{x}}(\mathbf{x}(t))$	$\Pr(\mathbf{x}(t) = \mathrm{x}(t))$ 的简写
$H(\mathbf{x}(t))$	$\mathbf{x}(t)$ 的信息熵（比特）
$\mathcal{X}_{-\mathbf{x}}$	$\mathcal{X} - \{\mathbf{x}\}$
$\mathbb{E}_{p(\mathrm{x})}\{\mathbf{x}\}$	随机变量 X 在概率分布 $p(\mathrm{x})$上的期望
$\mathrm{tr}(t)$	系统轨迹在时间 t的实现值

表 I 符号表

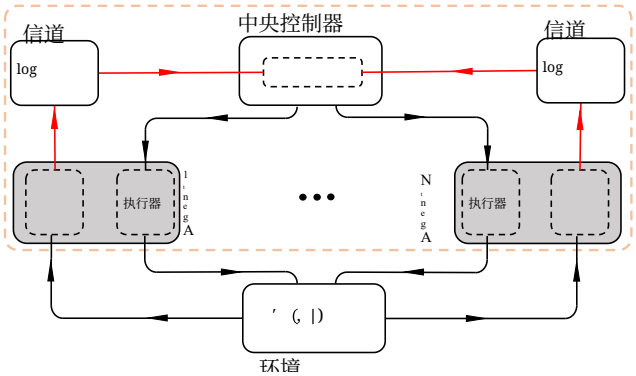


图 3. 中央控制器与智能体交互以控制环境的示意图。红色连线表示比特预算受限的通信信道，这意味着中央控制器仅具备局部（而非全局）可观测性。橙色虚线框详细展示了图 1 和图 2 中的同一框。

该系统由元组 $M = \{\mathcal{S}, \mathcal{M}^N, r(\cdot)\ \gamma, T(\cdot)\}$ 描述，其中 $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \to \mathbb{R}$ 是阶段奖励函数，标量 $0 \leq \gamma \leq 1$ 是折扣因子。函数 $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \to [0, 1]$ 是一个条件概率质量函数（pmf），用于表示状态转移，使得 $T(\mathrm{s}(t+1), \mathrm{s}(t), \mathrm{m}(t)) = \Pr(\mathrm{s}(t+1) | \mathrm{s}(t), \mathrm{m}(t))$。根据阶段奖励信号，时间范围 T' 内的系统回报记为

$$\mathbf{g}(t') = \sum_{t=t'}^{T'} \gamma^{t-1} r(\mathbf{o}_1(t), ..., \mathbf{o}_N(t), \mathbf{m}_1(t), ..., \mathbf{m}_N(t)). \quad (1)$$

虽然智能体可以共同观测系统状态[35],，但每个智能体 i的观测 $\mathrm{o}_i(t)$ 都是局部的 [6]。每个时间步，智能体 $i \in \mathcal{N}$可以通过通信消息 $\mathrm{c}_i(t)$ 向中央控制器发送其局部观测。智能体与中央控制器之间的通信以同步（非顺序）且同时（无延迟）的方式进行 [17]。每个智能体 i 按照其通信策略$\pi_i^c(\cdot) : \Omega \to \mathcal{C}$生成通信消息 $\mathrm{c}_i(t)$。与其他所有智能体并行，智能体 i遵循通信策略 $\pi_i^c(\cdot)$，将当前观测 $\mathrm{o}_i(t)$映射为通信消息 $\mathrm{c}_i(t)$，中央控制器将在同一时间步收到该消息 t。码本 $\mathcal{C}$ 是由有限个通信码字 $\mathrm{s}\ \mathrm{c}, \mathrm{c}', \mathrm{c}'', ..., \mathrm{c}^{(|\mathcal{C}|-1)}$ 组成的集合 —— 我们也用相同的记号表示动作、观测和状态空间中的不同元素。智能体的通信消息通过无差错有限速率比特管道发送，其速率约束为 $R \in \mathbb{R}$（每次信道使用的比特数），等价地也可表示为（每个时间步的比特数）。因此，量化码本的大小应满足不等式 $|\mathcal{C}| \leq 2^R$。中央控制器利用最近 d个时间步内收到的通信消息 $\mathrm{c}(t) \triangleq \langle \mathrm{c}_1(t) ..., \mathrm{c}_N(t) \rangle$，按照控制策略 $\pi^m(\cdot) : \mathcal{C}^{Nd} \to \mathcal{M}^N$生成动作信号 $\mathrm{m}(t)$。根据上述描述，从中央控制器的角度来看，环境

[6]在我们的问题设定中，由于局部可观测性，每个智能体都无法将环境视为马尔可夫决策过程（MDP）。我们仅假设环境存在一个底层的 MDP，这一假设在强化学习算法文献中被广泛采用，例如 [36] [37]。我们作出这一假设，是因为我们的性能保证依赖于控制任务解的最优性； [7], [10] 中也作出了同样的假设。需要指出的是，在我们所有的数值研究中，即使中央控制器（CC）获得了所有智能体的联合观测，也无法观测到环境的真实/名义状态。

而且，从智能体的角度来看，环境也未必是一个 MDP——因为没有任何智能体能够观察到环境的名义状态。

B. 问题陈述：联合控制与通信设计（*JCCD*）问题

现在我们定义 JCCD 问题。令 M 为支配环境的马尔可夫决策过程，标量 $R \in \mathbb{R}$ 为所有智能体上行链路的比特预算。在任意时间步 t'，我们的目标是选择元组 $\pi = \langle \pi^m(\cdot)\ \pi^c \rangle$，其中 $\pi^c \triangleq \langle \pi_1^c(\cdot).....,\ \pi_N^c(\cdot) \rangle$ 以求解以下变分动态规划

$$\underset{\pi}{\text{argmax}}\ \ \mathbb{E}_{\pi}\Big\{\mathbf{g}(t')\Big\};\ \ \text{s.t.}\ \ |\mathcal{C}| \leq 2^R, \tag{2}$$

其中，期望是对系统轨迹的联合概率质量函数 $\{\mathbf{tr}\}_{t'}^{T'}$ $=\mathbf{o}_1(t').....,\ \mathbf{o}_N(t')\ m(t').....,\ \mathbf{o}_1(T').....,\ \mathbf{o}_N(T')$、$\mathbf{m}(T')$ 求取的，智能体遵循策略元组 π。在下一节中，与 [18] 类似，我们将通过对观测进行基于动作的量化，分离动作策略与通信策略的设计。与 [18]，不同，这里假设多智能体系统的通信网络采用星型拓扑。这种分离的思路旨在提取控制设计问题中可能影响通信设计的特征，并在设计通信方案时将其纳入考量。因此，我们的通信设计将考虑控制任务的关键特征。我们利用解析技术以及强化学习 [17], [18]来提取控制任务的关键特征。事实上，这种称为 TODC 的新通信问题将不再类似于传统通信问题，因为它是受 JCCD 问题启发而提出的。

在 [18], [23],中，作者将智能体观测对于给定任务的价值，作为通信设计中所考虑控制任务的关键特征。因此，其思路是将价值相似的观测点聚为一簇。与 [18], [23],不同，后者将观测价值视为控制任务的显式关键特征；本文则将分配给每个观测的最优控制/动作值作为关键特征。因此，只要观测点被分配了相似的最优控制/动作值，ABSA 就会将相应的观测值聚合在一起。强化学习领域已将基于动作的状态聚合作为一种方法引入，用以降低强化学习算法的复杂度，同时保持平均回报性能 [38], [39]。

III. 基于动作的观测无损压缩

本节将以另一个示例（除 [18] 外）说明如何使用通用框架解决 JCCD问题。在 [18], 中，作者针对分布式控制和量化解决了一个类似问题，并借助一个假设的泛函 Π^{m^*}，将面向任务的通信策略与动作策略的设计分离开来。具体而言，泛函 Π^{m^*} 是从所有可能的通信策略所构成的向量空间映射到 $\mathcal{K}^c$ 最优对应控制策略所构成的向量空间 $\mathcal{K}^m$ 的映射 $\pi^{m^*}(\cdot)$。有了泛函 Π^{m^*}后，只要 JCCD问题中出现函数 π^m，就可以用 $\Pi^{m^*}(\pi^c)$ 替代，从而得到一个只需设计通信策略 π^c 的新问题。[18], 中，作者使用 $\Pi^{m^*}(\pi^c)$ 的近似来构建面向任务的量化器设计问题；而在本研究中，我们针对式（3）的简化版本推导出精确解——其中，与中央控制器通信的智能体数量限制为一个。为使 ABSA 适用于问题（3）的通用设定，在 ABSA-2 中，我们将借助一种近似技术解除这一限制。

JCCD问题已可表述为一种数据量化问题。引理 1 指出了本文旨在优化的量化度量，并将JCCD问题重新表述为一种新型的广义数据量化问题。

引理 1。*JCCD*问题 *(2)* 也可以表示为如下广义数据量化问题

$$\underset{\pi}{\textit{argmin}}\ \ \mathbb{E}_{p(\mathbf{s}(t))}\Big|V^{\pi^*}\big(\mathbf{s}(t)\big) - V^{\pi^m}\big(\mathbf{c}(t)\big)\Big|,\ \ \textit{s.t.}\ \ |\mathcal{C}| \leq 2^R, \tag{3}$$

其中，由 π^c 生成的通信向量 $\mathbf{c}(t)$ 是系统状态$\mathbf{s}(t)$的量化版本。

证明。 附录 A。 ■

与经典的数据量化问题不同，此处的失真度量衡量的是原始信号及其量化版本的两个不同函数之间的差异，即 $V^{\pi^*}{}_{(\cdot)}$ 和 $V^{\pi^m}{}_{(\cdot)}$；因此，我们通过求解 (3) 来优化的失真度量并非常规度量。事实上，该变分最小化问题是在联合量化策略 π^c 函数与动作策略 π^m函数构成的向量空间上求解的。

*A. ABSA-1*算法

所提出的 ABSA-1 仅适用于两个在数学上等价的场景：(i) 只有一个智能体与中央控制器通信——如图 2-a 所示，此时仅有一个智能体连接到中央控制器；或 (ii) 智能体通过中继与中央控制器通信。在后一种场景中，中继可以完全访问智能体的通信观测，即$\mathbf{o}_i, \forall i \in \mathcal{N}$，而中继与中央控制器之间的信道则受到比特预算限制。这一受限场景有助于我们对问题 (3) 开展解析研究，使我们能够从理论上证明 ABSA-1 的压缩无损性及其最优平均回报性能。引理 2 将验证这些结论，其结果也有助于设计 ABSA-2。ABSA-1 的核心思想是：当且仅当 $\pi^*\big(\mathbf{s}^{(i)}\big) = \pi^*\big(\mathbf{s}^{(j)}\big)$时，使用相同的通信消息$\mathbf{c}$ 表示任意两个状态$\mathbf{s}^{(i)}$,$\mathbf{s}^{(j)}$，其中 $\pi^*(\cdot)$: $\mathcal{S} \rightarrow \mathcal{M}^N$ 是智能体在能够访问所有智能体观测的情况下的最优控制策略。因此，ABSA-1 和 ABSA-2 分三个不同阶段求解 JCCD问题：(i) 通过强化学习在完美通信条件下求解集中式控制问题，即

使用 Q学习 求得 $\pi^*(\cdot)$[7]，；(ii) 通过某种数据聚类方法求解面向任务的数据量化问题，以求得 π^c；(iii) 找出与 π^c 相对应的 π^m。

为了解释 ABSA-1，我们引入具有集中式控制的面向任务的数据压缩问题。TBIC 采用 [18] 中的类似技术推导而来，但针对的是不同的设置，即 MAS 的通信网络具有星型拓扑。TBIC 问题不再是联合控制与通信问题，而是一个考虑了控制问题特征的量化设计问题。为了从 JCCD 问题得到 TODC 问题，我们使用泛函 Π^{m^*} 来替换 $\pi^m(\cdot)$ 与 $\Pi^{m^*}(\pi^c)$。在 Π^{m^*}可用的情况下，将其代入 JCCD 问题（2），我们将得到一个新问题

$$\underset{\pi^c}{\text{argmin}} \quad \mathbb{E}_{p(\mathbf{s}(t))}\left|V^{\pi^*}(\mathbf{s}(t)) - V^{\Pi^{m^*}(\pi^c)}(\mathbf{c}(t))\right|, \quad \text{s.t.} \ |\mathcal{C}| \le 2^R, \tag{4}$$

其中，我们仅针对本地中继的通信策略 $\pi^c(\cdot)$ 最大化系统回报。中央控制器的最优控制策略 $\pi^{m^*}(\cdot)$ 通过映射 $\Pi^{m^*}(\pi^c(\cdot))$自动计算得到。这里将该问题称为 TODC 问题。在 Π^{m^*}可用的情况下，JCCD 问题（2）可化简为（4）。为通过获得中继的通信策略 $\pi^c(\cdot)$ 以及相应的 Π^{m^*}来求解（2），定义1 给出了形式化求解（4）的精确方法。

定义1。 ***ABSA-1*** **中的量化和控制策略：**

ABSA-1 设计的通信策略 $\pi^{c,\ ABSA-1}(\cdot)$ 将通过求解以下 *k*-中位数聚类问题获得

$$\min_{\mathcal{P}} \quad \sum_{i=1}^{|\mathcal{C}|}\sum_{\mathbf{s}(t)\in\mathcal{P}_i}\left|\pi^*(\mathbf{s}(t)) - \mu_i\right|, \tag{5}$$

其中，$\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_B\}$ 是 $\mathcal{S}$ 的一个划分，μ_i 是每个簇 i 的质心。*ABSA-1* 的通信策略 - $\pi^{c,\ ABSA-1}(\cdot)$ - 是一个任意的非单射映射，使得 $\forall k \in \{1, ..., B\}$：$\pi^{c,\ ABSA-1}(\mathbf{s}) = \mathbf{c}^{(k)}$ 当且仅当$\mathbf{s} \in \mathcal{P}_k$。现在，令 C_g 为一个函数复合算子，使得 $C_g f = g \circ f$。我们定义算子 $\Pi^{m^*} \triangleq C_g$，其中 $g = \pi^*(\pi^{c,\ ABSA-1^{-1}}(\cdot))$[8]。

所提出的 ABSA-1 算法的最优性随后将在定理 2 中给出。

引理 2. 根据定义 *1* 所述的通信策略 $\pi^{c,\ ABSA-1}$ -若 $|\mathcal{C}| \ge |\mathcal{M}|^N$，则将针对平均回报对观测数据进行无损压缩。

证明。 附录 B。 ■

备注： ABSA-1 还将针对失真度量对观测数据进行无损压缩

[7]随着智能体数量增加，Q学习的复杂度也随之上升，这正是 ABSA 的瓶颈所在 N。任何需要集中式训练阶段的其他算法也面临类似的局限 [7], [30]8

注意，当 $\pi^{c,}$ 其逆函数对于任何输入都不会产生唯一的输出。因此，我们用 $\pi^*(\pi^{c,ABSA-1^{-1}}(\mathbf{c}'))$ 表示$\pi^*(\mathbf{s}')$，其中$\mathbf{s}'$ 可以是 $\pi^{c,}$ $^{ABSA-1^{-1}}(\mathbf{c}')$的任意输出。

在问题 (3) 中引入。鉴于引理 2 和引理 1 的证明，该备注的证明显而易见，因此省略。

ABSA-1 中量化的无损性意味着，与在完美通信条件下使用最优策略 $\pi^*(\cdot)$ 控制多智能体系统的情况相比，π^{ABSA-1} 不会造成系统平均回报的损失。因此，控制策略 $\pi^{m,\ ABSA-1}(\cdot)$ 是最优的。再次说明，这里我们不使用传统的量化失真度量，而是选择一种局部观测的表示方式，使传递的消息能够最大化平均任务回报。

请注意，在 [7],中，作者并未找到能够将通信与控制联合问题简化为面向任务的通信设计的高阶函数 Π^{m^*}，而是求解了面向任务的通信设计问题的近似版本。然而，在本文中，我们通过 ABSA-1 引入了一个闭式 Π^{m^*}，它可以将 ABSA-1 引入的每一种通信策略 $\pi^{c,\ ABSA-1}$ 映射为精确的最优控制策略。这意味着 ABSA-1 给出的解也是通信与控制联合设计（JCCD）问题的最优解。

B. ABSA-2 算法

我们在前面的引理 2 中看到，通过求解问题 5 得到的通信策略是最优的，并且当 $|\mathcal{C}| \ge |\mathcal{M}|^N$ 时能够实现无损的平均回报性能。然而，要解决问题 5，我们需要知道 $\pi^*(\mathbf{s}(t))$。这是一个限制性假设，在 ABSA-1 中可转化为两种不同的系统模型，它们都不如图 3 所示的系统一般：（i）在智能体与中央控制器之间设置一个额外的中继，该中继与智能体之间具有理想的下行信道，并通过一个比特预算受限的信道与中央控制器通信。（ii）多智能体系统仅由一个智能体和一个中央控制器组成，其中智能体的上行链路受比特预算限制，而下行链路为理想信道。

我们提出的第二种算法 ABSA-2 无需知道 $\pi^*(\mathbf{s}(t))$，并且能够在图 3 所示的更一般设置下运行。具体做法是：根据智能体 i 的局部观测 $\mathrm{o}_i(t)$，近似该智能体处 $\pi^*(\mathbf{s}(t)) = \langle \mathrm{m}_1*(t) ..., \mathrm{m}_N*(t)\rangle$ 中的局部元素 $\mathrm{m}_i^*(t)$。也就是说，在集中式训练阶段，我们可以获得 $p(\mathrm{o}_i, \mathrm{m}_i^*)$ 的经验联合分布，并据此得到 $\hat{\mathrm{m}}^*{}_i$ 的数值最大后验估计量。因此，ABSA-2 支持完全分布式的通信策略。具体而言，每个智能体在与中央控制器或其他智能体通信之前，分别对自己的通信消息进行编码。这种编码方式通常称为分布式编码。此外，ABSA-2 在每个智能体处执行的编码过程复杂度低、功耗低，在此之前无需智能体间通信。在这种情况下，每个智能体都通过比特预算受限的通信信道，将其编码观测直接发送给中央控制器。为了提高中央控制器的学习效率，中央控制器可以考虑时间范围 $[t-d, t]$ 内收到的所有通信消息，再作出控制决策 $m(t)$。因此，ABSA-2

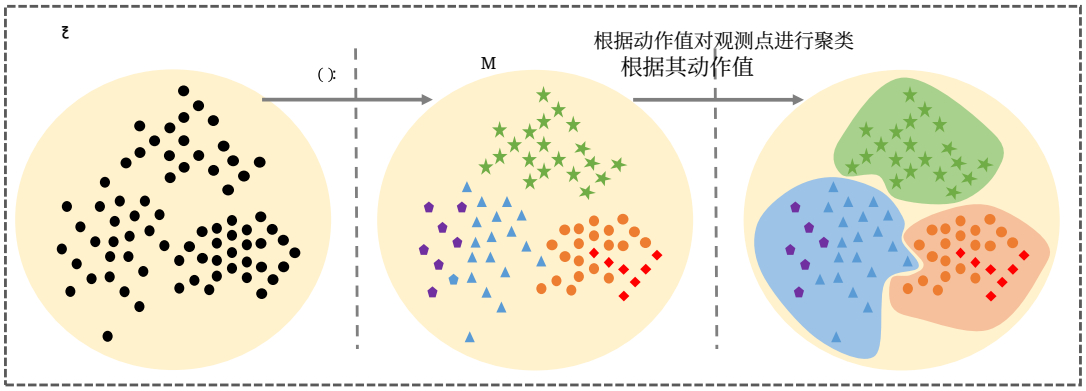


图4。ABSA-2 中状态的抽象表示，其中 $|\mathcal{C}| = 3$ 和 $|\mathcal{M}| = 5$ - $|\mathcal{M}|$ 由用于展示观测点的形状数量表示，而 $|\mathcal{C}|$ 由右侧子图中显示的簇数量表示。左侧子图展示了聚合前的观测点。在集中式训练阶段，我们首先根据 $\pi_i^*(\cdot)$计算 $\pi^*(\cdot)$：由此可获得 $\Omega \to \mathcal{M}$。我们使用满射 $\pi_i^*(\cdot)$将高维/高精度观测空间映射到低维/低精度空间。中间子图展示了观测点及其对应的动作值——每种独特的形状代表一个独特的动作值。**这种对观测点的新表示，将控制问题的特征嵌入数据量化问题中**。最后，我们根据观测点的动作值对其进行聚类——分配给某个（或某组）动作值的所有观测点将被聚集在一起。右侧子图展示了聚合观测空间，其中每个簇中的所有观测点都将使用相同的通信消息表示。使用深度Q网络运行的集中式控制器在每个时间步通过从所有智能体接收到的全部这些聚合观测/通信来观测环境。

该算法可以在中央控制器处计算复杂度与智能体通信有效性之间实现权衡：前者直接受 d 的取值影响，后者则受 $|\mathcal{C}|$ 的取值反向影响。此外，ABSA-2 可以轻松扩展，以便为每个智能体 i采用不同的 $|\mathcal{C}|$取值，而不是让所有智能体共用一个固定的比特预算$R = \log_2 |\mathcal{C}|$。

如图4所示，在 ABSA-2 中，每个智能体 i 通过在其局部观测空间而非全局状态空间上求解聚类问题，获得通信策略函数 $\pi_i^c(\cdot)$，其形式如下：

$$\min_{\mathcal{P}_i} \quad \sum_{j=1}^{|\mathcal{C}|} \sum_{\mathbf{o}_i(t)\in\mathcal{P}_{i,j}} \left| \tilde{\pi}_i^*(\mathbf{o}_i(t)) - \mu_{i,j} \right|, \qquad (6)$$

其中，$\mathcal{P}_i = \{\mathcal{P}_{i,1}, \ldots, \mathcal{P}_{i,|\mathcal{C}|}\}$ 是 Ω 的一个划分，且

$$\tilde{\pi}_i^*(\mathbf{o}_i(t)) = \text{argmax}_{\mathbf{m}_i^*} \; p_{\pi^*}(\mathbf{m}_i^* | \mathbf{o}_i(t)), \qquad (7)$$

而 $\mathbf{m}_i^*$ 是智能体 i的最优动作，它是$\mathbf{m}^* \triangleq \pi^*(\mathbf{o}_1(t) \ldots, \mathbf{o}_N(t))$中的第 i个元素。因此，$\tilde{\pi}_i^*(\mathbf{o}_i(t))$ 是在给定局部观测 $\mathbf{o}_i(t)$ 时，对$\mathbf{m}_i^* = \pi^*(\mathbf{s}(t))$ 的最大后验估计器。

完成式 (6) 中的聚类后，每个智能体 i 将训练其局部通信策略 $\pi^c_{,i}{}^{ABSA-2}(\cdot)$，该策略是满足以下条件的任意非单射映射：$\forall k \in \{1, \ldots, |\mathcal{C}|\} : \pi^c_{,i}{}^{ABSA-2}(\mathbf{o}_i) = \mathbf{c}^{(k)}$ 当且仅当 $\mathbf{o}_i \in \mathcal{P}_{i,k}$。获得通信策略 $\langle \pi_i^{c,ABSA-2}(\cdot) \rangle_{i=1}^N$后，为了在中央控制器（CC）处获得与这些通信策略相对应的适当控制策略 $\pi^m(\cdot)$，我们采用单智能体强化学习。为此，并为了控制算法在 d 较大时的复杂度，我们提议在中央控制器处使用深度Q网络架构 [41]。

IV. 性能评估

本节通过对广受关注的多智能体几何一致性问题进行数值分析，评估所提方案。

算法1. 基于动作的状态聚合（ABSA-2）

1：将回放记忆库 D 初始化为容量 10’000。 2：使用随机权重 θ初始化状态-动作价值函数 $Q(\cdot)$。 3：使用权重 $\theta^t = \theta$ 初始化目标状态-动作价值函数 $Q^t(\cdot)$。 4：使用 Q学习[40]* 求解式 (2)，得到 $\pi^*(\cdot)$ 和 $Q^*(\cdot)$，其中 $R >> H(\mathbf{o}_i(t))$ $\forall i \in \mathcal{N}$。 5：计算 $\pi_i^*(\mathbf{o}_i(t)) =$ 众数$\left[\mathbf{m}_i^* | \mathbf{o}_i(t)\right]$，其中 $\forall \mathbf{o}_i(t) \in \Omega$，适用于 $i \in \mathcal{N}$。 6：通过对问题 (5) 应用 k- 中位数聚类，求得 $\mathcal{P}_i$ 和 $\pi_i^c(\cdot)$，适用于 $i \in \mathcal{N}$。 7：对于每个回合 $k = 1$：200’000，执行 8：随机初始化观测 $\mathbf{o}_i(t = 0)$，适用于 $i \in \mathcal{N}$；9：随机初始化消息 $\mathbf{c}(t = 0)$；10：对于 $t = 1$： T'，执行 11：智能体 i按照 $\pi_i^c(\cdot)$选择 $\mathbf{c}_i(t)$，适用于 $i \in \mathcal{N}$；12：在中央控制器处获取消息 $\langle \mathbf{c}_1(t) \ldots, \mathbf{c}_N(t) \rangle$；13：中央控制器遵循 ϵ-贪心策略生成动作 $\mathbf{m}_i(t)$，适用于 $i \in \mathcal{N}$；14：在中央控制器处获取奖励 $r(t) = R(\mathbf{s}(t), \mathbf{m}(t))$；16：将转移 $\{\mathbf{c}(\,)$ $\mathbf{m}(\,)\,(\,)\,\mathbf{c}(\,)\}$ 存储到：17：结束；18：从 中抽取 $\mathbf{c}(\,)\,\mathbf{m}(\,)\,(\,)$ $\mathbf{c}(\,)$；对于小批量中的每个转移：，执行 20：计算深度Q网络的平均损失 $L_{t'}(\theta) = \frac{1}{2}\left(r(t') + \mathbf{max}_{\mathbf{m}^*} \; Q^t(\mathbf{c}(t'+1), \mathbf{m}^*, \theta^t) - \mathbf{max}_{\mathbf{m}^*} \; Q(\mathbf{c}(t'), \mathbf{m}^*, \theta)\right)^2$，21 $L_{t'}\ \theta$ ；21：对 $(\,)$ 关于 θ 执行一次梯度下降；22：结束；23：每 1000 步更新一次目标网络 $Q^t(\cdot)$；24：结束

.....问题[9]。通过间接设计，ABSA-1 和 ABSA- 2 从不依赖于任何特定任务的明确领域知识，例如几何一致性。因此，我们推测，它们的间接设计使其能够应用于几何一致性问题之外的领域，并适用于范围广得多的任务。为了使几何一致性任务适用于评估我们提出的算法，类似于 [18],，我们在智能体与中央控制器之间的通信信道中加入了比特约束。在会合问题的背景下评估所提出的算法后，我们尝试通过现有指标——正向倾听——来解释所有算法的行为，该指标用于衡量通信的任务有效性。由于正向倾听不足以解释所研究算法行为的所有方面，我们还将引入一个新指标。这个新指标称为任务相对信息，有助于以更高的准确性和可靠性进一步解释不同算法的行为。

A. 几何一致性问题

本节通过会合问题 [42]，[43],的数值结果评估所提方案。会合问题是在有限可观测性 [28]条件下的一种特殊几何一致性问题。遵循第二节 A 和图 2 中分别说明的瞬时同步通信模型与星形网络拓扑，会合问题可描述如下。在每个时间步 t，依次发生若干事件。首先，智能体 i 获取局部观测 $\mathrm{o}_i(t)$——这相当于其在网格世界中的位置。随后，智能体 i遵循其量化/通信策略，将观测压缩为压缩版本 $\mathrm{c}_i(t)$，并通过比特预算受限通信链路发送给中央控制器。中央控制器收到所有智能体的量化观测后，遵循其控制策略，确定并选择联合动作向量 $\mathrm{m}(t)$，然后相应地将每个智能体 i的局部动作 $\mathrm{m}_i(t)$发送给它。通过理想通信信道传回智能体 i 的局部动作 $\mathrm{m}_i(t) \in \mathcal{M}$ 是在网格世界中的单向移动，即$\mathcal{M} = \{$左、右、上、下或暂停$\}$。给定每个智能体 i的动作 $\mathrm{m}_i(t)$后，环境随之演化并转移至下一个时间步 $t+1$，届时每个智能体 i 都会获取新的局部观测 $\mathrm{o}_i(t+1)$。所有智能体均获得单一团队奖励

$$\mathrm{r}_t = \begin{cases} C_1, & \text{if } \exists i,j \in N : \mathrm{o}_i(t) \in \Omega^T \ \ \& \ \ \mathrm{o}_j(t) \notin \Omega^T \\ C_2, & \text{if } \nexists i \in N : \mathrm{o}_i(t) \in \Omega - \Omega^T, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

其中，$C_1 < C_2$ 和 Ω^T 是终止观测的集合，即当 $\exists i \in \mathcal{N}$: $\mathrm{o}_i(t) \in \Omega^T$时，回合结束。因此，当并非所有智能体都到达目标点时，获得的奖励$C_1 = 1$ 较低；而当所有智能体同时到达目标点时，则获得较高的奖励 $C_2 = 10$ 。

[9]在数值实验中，假设折扣因子为 $\gamma = 0.9$。所有实验均在 8×8 大小为 的网格世界中进行，其中会合的目标点位于网格编号 $\Omega^T = \{22\}$处。

我们将提出的 ABSA算法与 [18] 中提出的启发式非通信（HNC）、启发式最优通信（HOC）和 SAIC算法进行比较。这些算法是针对本文所解决的特定几何一致性问题，联合设计通信和控制策略的直接方案。与采用间接设计的 ABSA-1 和 ABSA-2 不同，HOC 和 HNC 的直接设计使其无法应用于其他问题，只能用于本文所述具有有限可观测性的特定几何一致性问题，即会合问题。

B. 数值实验

使用精确 Q 学习时，采用恒定学习率 $\alpha = 0.07$ 来获取 ABSA-2 的 $\pi^*(\cdot)$；使用深度Q网络学习 ABSA-2 的 $\pi^m(\cdot)$时，采用 $\alpha = 0.0007$ 。对于精确 Q 学习，考虑使用 UCB[10] 探索率 $c = 1.25$ 。用于近似 Q 值的深度神经网络采用具有 10 层深度的全连接前馈网络，并使用 Adam优化器进行优化。经验回放缓冲区的大小设为 10’000，迷你批次大小为 62。目标Q网络每 1000 步更新一次；探索时使用衰减的 ϵ-贪婪策略，其初始$\epsilon = 0.05$ 和最终 $\epsilon = 0.005$ 分别为 [41]。在以平均折扣累积奖励报告各方案性能的图中，使用记忆长度为 20,000 次迭代的移动平均滤波器对训练迭代次数期间获得的奖励进行平滑处理。如第三节-A所述，ABSA-1 和 ABSA-2 均需先经过集中式训练阶段，才能以分布式方式执行。

对于所有黑色曲线，都需要先经过一个集中式训练阶段，以获得 $\pi^*(\cdot)$。如第三节所述，所提出的算法 ABSA-1 和 ABSA-2 利用 $\pi^*(\cdot)$来设计 π^c ，然后再 π^m 。虚线曲线 HOC 和 HNC 由 [18] 提出，它们提供了启发式方案，利用其设计者关于会合任务的领域知识，因此不适用于会合问题以外的任何任务。HOC 采用联合控制与通信设计，而 HNC 不进行通信。请注意，HNC 和 HOC 要求智能体在任务起始点之前进行通信/协调，而其他方案均无此要求。这些由 [18], 引入的方案详述如下。

- 在会合问题中，**利用领域知识** 设计联合通信与控制策略。HNC 智能体接近目标点，并在附近等待若干时间步，以确保另一个智能体也已到达。只有在此之后，它们才会前往目标点。请注意，该方案要求智能体在任务起始点之前进行通信/协调，因为它们必须事先就这种协调方案达成一致。

- 在会合问题中，设计了一种**利用领域知识的**联合通信与控制策略。

[10]UCB 是精确强化学习中用于平衡探索与利用 [40]的标准方案。

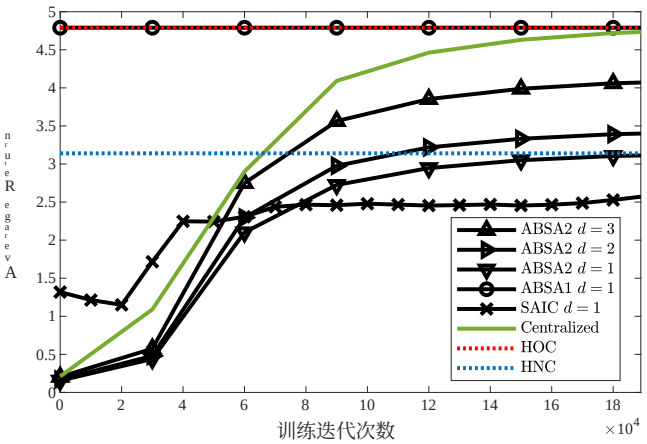


图5. 在 [18] 中引入的所提方案与一些基准方法的平均回报对比——恒定比特预算下的三智能体场景。

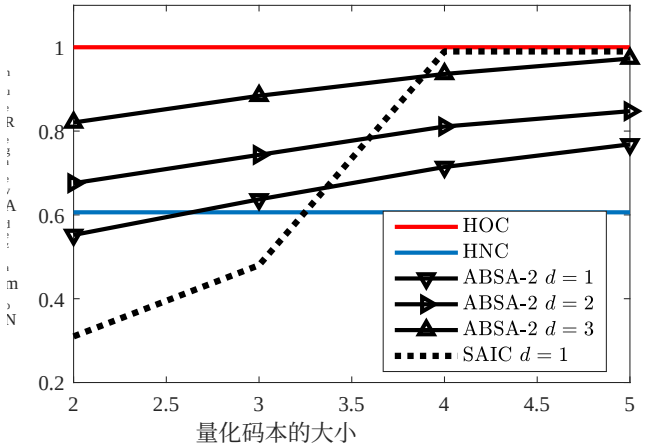


图 6. 比较所提方案与 [18] 双智能体场景中引入的一些基准方法在不同码本大小 $|\mathcal{C}|$ 下获得的归一化平均回报。

HOC 智能体在目标点旁等待，直到另一个智能体通知它们自己也已到达那里。只有在此之后，它们才会前往目标点。请注意，这种方案要求智能体在任务起点之前进行通信和协调，因为它们必须事先就这种协调与通信方案以及每条通信消息所传达的含义达成一致。

为了获得图 5 所示的结果，我们对三智能体系统的会合问题进行了仿真。黑色曲线展示了中央控制器（CC）处正在进行的训练阶段，以获得 π^m ，此后 π^c 已通过公式 (5) 和 (6) 计算得出。 我们观察到，ABSA-1 无损地实现了最优平均回报，无需进行任何 (第 2 轮)训练。为实现观测过程的完全分散式量化，提出了 ABSA-2；可以看到，随着d 增大，它逐渐接近最优解。所有 ABSA-2 曲线均绘制为 $|\mathcal{C}| = 3$，而 ABSA-1 曲线在三智能体场景（图 5）中绘制为 $|\mathcal{C}| = |\mathcal{M}|^N = 125$ ，在双智能体场景（图 6）中则绘制为 $|\mathcal{C}| = |\mathcal{M}|^N = 25$ 。

图 5 展示了 ABSA-2 在不同量化率下与 HNC、HOC 和 SAIC 的性能对比。正如预期，随着量化码本大小的增加，ABSA-2 的平均回报性能逐渐提升，并在 $d = 3$ 时接近最优性能。我们还观察到，在比特预算极为紧张、SAIC 的性能大幅下降时，ABSA-2 的性能优于 SAIC。可以看到，随着 d 增大，即使在较高量化率下，ABSA-2 也能达到接近最优的回报性能；然而，较大的 d 会增加 ABSA-2 的计算复杂度。

C. 学习所得通信策略的可解释性

文献中用于评估通信有效性的一项常见指标 [37] 是正向倾听$I\left(\mathbf{c}_i(t);\mathbf{m}_j(t)\right)$ $j \in \mathcal{N}-\{i\}$，即互信息智能体产生的通信$\mathbf{c}_i(t)$与另一个智能体接收到来自智能体 i 的通信$\mathbf{c}_i(t)$后所选动作$\mathbf{m}_j(t)$之间的互信息。正向信号传递 $I\left(\mathbf{o}_i(t);\mathbf{c}_i(t)\right)$ 是 [37],提出的另一项指标，用于衡量智能体 i 在同一时间步的观测$\mathbf{o}_i(t)$与其自身产生的通信消息$\mathbf{c}_i(t)$之间的互信息。然而，正如以下将要展示的，这些指标无法充分反映所有方案的潜在性能趋势。因此，我们首次引入一项名为任务相关信息(RI)的新指标，以解释所学习通信策略的任务有效性。

衡量正向倾听是量化智能体 i 传递的消息对智能体 j 动作选择所作贡献的一种方式。另一方面，正向信号传递用于衡量传递的消息$\mathbf{c}_i(t)$与智能体观测$\mathbf{o}_i(t)$之间的一致性及相关性。由于 SAIC 和 ABSA 使用确定性映射将观测 $\mathbf{o}_i$ 转换为通信消息 $\mathbf{c}_i$，因此它们始终能够保证正向信号传递 [37] ——不过，其程度受上行信道的比特预算 $R = \log_2|\mathcal{C}|$限制。因此，在现有的通信效果衡量指标中，我们将数值研究限定于正向倾听的测量。已知正向倾听越高，我们预期智能体之间的协调程度就越强（但不一定越好）。也就是说，正向倾听越高，智能体之间（其动作与观测）的依赖程度就越高，而这并不一定足以使团队智能体完成任务。

图7说明，智能体与中央控制器之间更强的协调通常会提升多智能体系统获得更高平均回报的表现。例如，如图 7 所示，SAIC 的正向倾听性能从 $|\mathcal{C}| = 3$ 量化器提升至 $|\mathcal{C}| = 4$ 量化器，这带来了平均回报性能的提升，如图 6 所示。这一指标也合理地解释了 ABSA-2 性能的提升：通过增加 d ——中央控制器的记忆——以及扩大量化码本的大小，ABSA-2 获得了更高的回报

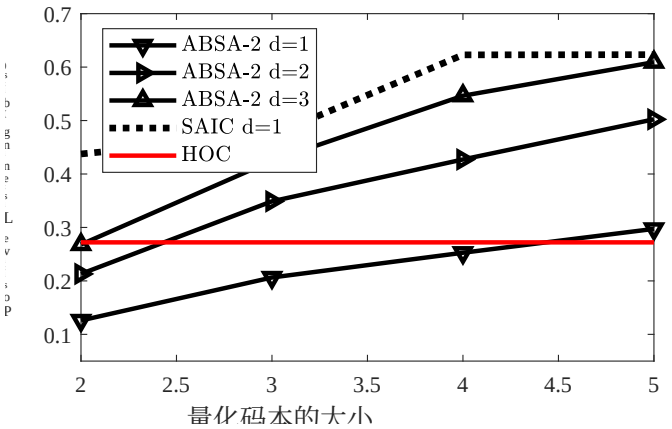


图7. 不同方案在一系列参数范围内的正向倾听 $I(\mathbf{c}_i(t);\mathbf{m}_j(t))$ 性能比较。

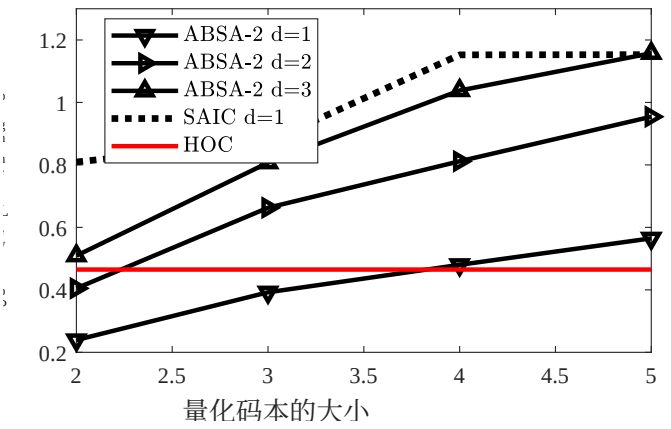


图 8. 比较不同方案下正向倾听 $I(\mathbf{c}_i(t);\mathbf{m}(t))$ 的性能。

$|\mathcal{C}|$。此外，与 HOC 相比，ABSA-2 中智能体与中央控制器之间的协调更强。因此，我们预期 ABSA-2 的平均回报性能会更好，这与图 5 的结果相反。这一现象表明，通过正向倾听衡量的更强协调未必会带来平均回报性能的提升，因为协调可能无法与任务需求完全匹配。

关于 HOC 方案的曲线让我们回想起，在上述几何一致性任务中，0.3 比特的正向倾听足以维持实现最优性能所需的协调。因此，在 ABSA-2 和 SAIC 方案中，通信消息仍会对接收端所选的动作产生不必要的影响。事实上，接收端收到的信息并非全部都有助于提高系统的平均回报。因此，ABSA 设计的通信消息中仍存在一些不必要的数据，其中不包含特定于任务或有用的信息。

因此，我们认为，正向倾听无法明确量化面向任务的通信算法的有效性；故而也不足以解释这些算法的行为。即使将正向倾听计算为 $I(\mathbf{c}_i(t);\mathbf{m}(t))$，以捕捉智能体 i 的通信与所有智能体的控制信号之间的互信息，我们得到的模式也几乎相同——见图8。

图 9 通过一种新颖的性能指标——任务相关信息（TRI），考察了多种方案的性能。我们将任务相关信息指标定义为

$$I\Big(\pi^c\big(\mathbf{o}_i(t)\big);\pi^*\big(\mathbf{s}(t)\big)\Big) = I\big(\mathbf{c}_i(t);\mathbf{m}^*(t)\big), \qquad (9)$$

它衡量智能体 i 传输的消息与中央控制器处联合最优动作向量 $\mathbf{m}^*(t)$ 之间的互信息（以比特为单位）；该向量由最优集中式控制策略 $\pi^*(\cdot)$ 选出。如图 9 所示，TRI 是衡量通信有效性的间接指标，可以解释不同通信设计的行为。还可以观察到，随着不同方案越来越接近最优性能，TRI 指标会放大它们之间的性能差距。尽管如此，TRI 可作为独立指标，用于量化通信设计的有效性，因为它几乎能完美预测通信策略的平均回报表现，而无需在解决实际任务时对通信进行测试。

请注意，我们根据采用量化算法时可获得的平均回报来衡量其任务有效性。此外，为了衡量在通信策略 $\langle\pi_1^c(\cdot)\ldots,\pi_N^c(\cdot)\rangle$ 下可获得的平均回报，我们必须在中央控制器处设计控制策略 $\pi^m(\cdot)$，使其仅能访问智能体的量化观测 $\mathbf{c}(t)$，并据此选择控制向量 $\mathbf{m}(t)$。因此，若没有针对多智能体系统控制策略的具体设计，我们就无法衡量其通信策略的有效性。即便完成了多智能体系统控制策略的设计，要判断算法的次优性能究竟是由控制策略设计不当还是通信策略设计不当造成的，也颇具挑战。事实上，很难将控制策略和通信策略对多智能体系统平均回报的影响区分开来。我们提出的 TRI 指标能够单独衡量任意通信策略的性能，并避免控制策略的影响体现在 TRI 的数值中。

因此，引入这一指标具有多方面的重要意义：（i）通过将 TRI 用作间接指标，我们可以衡量通信策略对任何特定任务的有效性；（ii）它使我们能够在设计任何控制策略之前，衡量通信方案的有效性；（iii）它有助于设计任务有效的通信策略，并使其与控制策略设计完全分离。

五、结论

本文研究了集中式控制和分布式通信策略下多智能体系统中控制与通信的联合设计。我们首先提出了一种基于动作的状态聚合算法（ABSA-1），用于

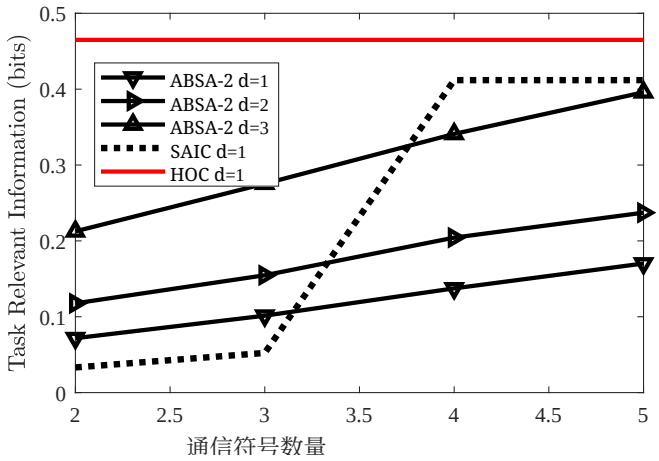


图 9. 比较一系列方案的任务相关信息（TRI）性能。可以看出，TRI 能够全面解释某一任务中所有任务有效量化方案的表现，而无需通过这些方案在任务中产生的平均回报来衡量其有效性——请将此图与图 6 进行比较。

无损压缩，并通过解析证明其最优性。随后，我们提出了 ABSA-2，该算法提供了一种完全分布式的通信策略，并可在计算复杂度与通信效率之间进行权衡。最后，我们通过在几何一致性问题上进行数值实验，并采用若干具有代表性的指标，展示了所提算法的任务有效性。 此外，我们的数值研究表明，亟需进一步研究，以寻找能够更准确地衡量或解释通信任务有效性的指标。面向任务的设计中的可扩展性也是未来研究需要解决的另一项核心挑战。

附录 A 引理 1 的证明

证明。将亚当定律应用于公式 (2)，得到

$$\underset{\pi}{\text{argmax}} \quad \mathbb{E}_{p(\mathbf{c}(t))}\Big\{\mathbb{E}_{p_{\pi^c,\pi^m}(\{\mathrm{tr}\}_{t'}^{T'}|\mathbf{c}(t))}\{\mathbf{g}(t')|\mathbf{c}(t)\}\Big\}, \text{ s.t. } |\mathcal{C}| \le 2^R \tag{10}$$

其中，$\mathbf{c}(t)$ 由通信策略 π^c 生成，系统轨迹 $\{tr\}_{t'}^{T'}$ 的联合概率质量函数直接受动作策略 π^m 的影响。条件概率质量函数 $p_{\pi^c,\pi^m}(\{\mathrm{tr}\}_{t'}^{T'}|\mathbf{c}(t))$ 表示在遵循策略 $\pi^c(\cdot)$ 和 $\pi^m(\cdot)$ 时，给定接收到的通信 $\mathbf{c}(t)$ 的系统轨迹的联合概率。接下来，我们对公式 (10) 取负，并在目标函数中添加第二项；该项相对于问题的决策变量为常数，从而得到

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{\mathbb{E}_{p_{\pi^*}(\{\mathrm{tr}\}_{t'}^{T'}|\mathbf{s}(t))}\{\mathbf{g}(t')|\mathbf{s}(t)\}\Big\} - \\ & \mathbb{E}_{p(\mathbf{c}(t))}\Big\{\mathbb{E}_{p_{\pi^c,\pi^m}(\{\mathrm{tr}\}_{t'}^{T'}|\mathbf{c}(t))}\{\mathbf{g}(t')|\mathbf{c}(t)\}\Big\}, \text{ s.t. } |\mathcal{C}| \le 2^R.\end{aligned} \tag{11}$$

我们用系统回报的条件期望替换为价值函数 $V(\cdot)$，[40]（第 3.5 章），于是可得

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^*}(\mathbf{s}(t))\Big\} - \mathbb{E}_{p(\mathbf{c}(t))}\Big\{V^{\pi^m}(\mathbf{c}(t))\Big\}, \\ \text{s.t.} \quad & |\mathcal{C}| \le 2^R.\end{aligned} \tag{12}$$

注意，$\mathbf{c}(t)$ 的经验联合分布可以通过在$\mathbf{s}(t)$ 的经验分布上遵循通信策略 π^c 得到。

$$\begin{aligned}\underset{\pi^c}{\text{argmin}} \quad & \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^*}(\mathbf{s}(t))\Big\} - \mathbb{E}_{p(\mathbf{s}(t))}\Big\{V^{\pi^m}(\mathbf{c}(t))\Big\}, \\ \text{s.t.} \quad & |\mathcal{C}| \le 2^R.\end{aligned} \tag{13}$$

由于对于任意$\mathbf{s}(t) \in \mathcal{S}$，$V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t)) \ge 0$ 都成立，合并两个期望可得

$$\underset{\pi^c}{\text{argmin}} \quad \mathbb{E}_{p(\mathbf{s}(t))}\Big|V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t))\Big|, \quad \text{s.t.} \quad |\mathcal{C}| \le 2^R, \tag{14}$$

由此完成该引理的证明。 ■

附录 B 引理 2 的证明

证明。我们从引理 1 的结果——问题 (3) 出发。对$\mathbf{s}(t)$的经验分布取期望，并应用贝尔曼最优性方程，可得

$$\begin{aligned}\underset{\pi}{\text{argmin}} \quad & \frac{1}{n}\sum_{t=1}^{n}\Big|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t))) - Q^{\pi^m}\Big(\mathbf{c}(t),\pi^m(\pi^c(\mathbf{s}(t)))\Big)\Big|, \\ \text{s.t.} \quad & |\mathcal{C}| \le 2^R,\end{aligned} \tag{15}$$

其中，向量 $\pi^c(\mathbf{s}(t))$ 的维度为 N，其第 i 个元素为 $\mathbf{c}_i(t)$。根据定义1，将 $\pi^{c,ABSA-1}(\cdot)$ 和 Π^{m^*} 代入方程 (15)，可得

$$\frac{1}{n}\sum_{t=1}^{n}\Big|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t))) - Q^{\pi^*}\Big(\mathbf{c}(t),\pi^*(\mathbf{s}')\Big)\Big|, \tag{16}$$

其中，$\mathbf{s}' = \pi^{c,ABSA-1^{-1}}\Big(\pi^{c,ABSA-1}(\mathbf{s}(t))\Big)$，其任意可能取值都与 $\mathbf{s}(t)$ 一样属于同一子集 $\mathcal{P}_{k'}$；而根据 $\mathcal{P}_{k'}$ 的定义，我们知道 $\pi^*(\mathbf{s}(t)) = \pi^*(\mathbf{s}')$，若 $|\mathcal{C}| \ge |\mathcal{M}|^N$。因此，在方程 (17) 中用 $\pi^*(\mathbf{s}(t))$ 替换 $\pi^*(\mathbf{s}')$，可得

$$\frac{1}{n}\sum_{t=1}^{n}\Big|Q^{\pi^*}(\mathbf{s}(t),\pi^*(\mathbf{s}(t))) - Q^{\pi^*}\Big(\mathbf{s}(t),\pi^*(\mathbf{s}(t))\Big)\Big| = 0. \tag{17}$$

定理2的证明至此结束。

■

参考文献

[1] L. S. Vailshery，《2019 年至 2021 年全球物联网连接设备数量及 2022 年至 2030 年预测》，2022 年 8 月。 [在线]。网址：https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/[2] B. G¨uler、A. Yener 和 A. Swami，《语义通信博弈》，*IEEE*认知通信与网络汇刊，第 4 卷，第 4 期，第 787–802 页，2018 年。[3] H. Tong、Z. Yang、S. Wang、Y. Hu、W. Saad 和 C. Yin，《基于联邦学习的无线网络音频语义通信》，载于*2021 IEEE* 全球通信大会（*GLOBECOM*），2021 年，第 1–6 页。[4] N. Pappas 和 M. Kountouris，《自主系统中用于实时跟踪的目标导向通信》，载于*2021 IEEE* 自主系统国际会议（*ICAS*），2021 年，第 1–5 页。[5] E. Calvanese Strinati 和 S. Barbarossa，《6G 网络：超越香农，迈向语义与目标导向通信》，计算机网络，第 190 卷，第 107930 页，2021 年。[6] A. Mostaani、T. X. Vu、S. K. Sharma、Q. Liao 和 S. Chatzinotas，《信息物理系统中的面向任务的通信系统设计：理论与应用综述》，*arXiv* 预印本 *arXiv:2102 07166*，2021 年。

[7] J. Foerster、Y. Assael、N. de Freitas 和 S. Whiteson，《通过深度多智能体强化学习学习通信》，载于《神经信息处理系统进展论文集》，巴塞罗那，2016 年。[8] C. E. Shannon 和 W. Weaver，《通信的数学理论 [1949]。伊利诺伊州厄巴纳》，1959 年。[9] L. Hu、G. Wu、Y. Xing 和 F. Wang，《Things2vec：利用图表示学习对物联网进行语义建模》，IEEE物联网期刊，第 7 卷，第 3 期，第 1939–1948 页，2020 年。[10] J. Cai、W. Zhong 和 J. Luo，《Seminer：基于侧信息的专有工业控制协议语义挖掘器》，IEEE物联网期刊，第 9 卷，第 22 期，第 22 796–22 810 页，2022 年。[11] T.-Y. Tung、S. Kobus、J. P. Roig 和 D. G¨und¨uz，《有效通信：面向噪声信道上多智能体强化学习的联合学习与通信框架》，IEEE通信领域精选期刊，第 39 卷，第 8 期，第 2590–2603 页，2021 年。[12] M. P. Mota、A. Valcarce、J.-M. Gorce 和 J. Hoydis，《多智能体强化学习中无线 MAC 协议的涌现》，arXiv 预印本 arXiv:2108.07144，2021 年。[13] N. Shlezinger 和 Y. C. Eldar，《深度任务型量化》，熵，第 23 卷，第 1 期，第 104 页，2021 年。[14] M. A. Gutierrez-Estevez、Y. Wu 和 C. Zhou，《学习有意图地通信：导论》，arXiv 预印本 arXiv:2211 09613，2022 年。. [15] C. Zhang、H. Zou、S. Lasaulce、W. Saad、M. Kountouris 和 M. Bennis，《面向物联网的目标导向通信及其在数据压缩中的应用》，arXiv 预印本 arXiv:2211.05378，2022 年。[16] N. Shlezinger 和 Y. C. Eldar，《面向任务的量化及其在 MIMO 接收机中的应用》，arXiv 预印本 arXiv:2002.04290，2020 年。[17] A. Mostaani、O. Simeone、S. Chatzinotas 和 B. Ottersten，《面向多智能体协作的基于学习的物理层通信》，载于2019 IEEE个人、室内和移动无线电通信国际研讨会，2019 年 9 月。[18] A. Mostaani、T. X. Vu、S. Chatzinotas 和 B. Ottersten，《比特预算信道上多智能体通信的面向任务的数据压缩》，IEEE通信学会开放期刊，第 3 卷，第 1867–1886 页，2022 年。[19] M. Kountouris 和 N. Pappas，《面向联网智能系统的语义赋能通信》，IEEE通信杂志，第 59 卷，第 6 期，第 96–102 页，2021 年。[20] R. Carnap、Y. Bar-Hillel 等，《语义信息理论纲要》，1952年。[21] H. Zhang、S. Shao、M. Tao、X. Bi 和 K. B. Letaief，《采用任务无关发射机和动态数据的深度学习赋能语义通信系统》，arXiv 预印本 arXiv:2205.00271，2022 年。[22] P. A. Stavrou 和 M. Kountouris，《目标导向通信的率失真方法》，载于《2022 IEEE国际信息论研讨会（ISIT）》。IEEE，2022 年，第 590–595 页。[23] A. Mostaani、T. X. Vu、S. Chatzinotas 和 B. Ottersten，《速率受限信道上多智能体通信的状态聚合》，载于《GLOBECOM 2020—2020 IEEE 全球通信大会》。IEEE，2020 年，第 1–7 页。[24] D. Kim、S. Moon、D. Hostallero、W. J. Kang、T. Lee、K. Son 和 Y. Yi，《在多智能体强化学习中学习通信调度》，载于《国际学习表征会议》，2019 年。[25] J. Liu、S. S hao、W. Zhang 和 H. V. Poor，《语义源的间接率失真刻画：一般模型及高斯观测情形》，arXiv 预印本 arXiv:2201.12477，2022 年。[26] C.-M. Chou、C.-Y. Li、W.-M. Chien 和 K.-c. Lan，《车联网基础设施通信的可行性研究：Wi-Fi 与 WiMAX》，载于《2009年第十届移动数据管理：系统、服务与中间件国际会议》。IEEE，2009 年，第 397–398 页。[27] Y.-C. Liu、J. Tian、C.-Y. Ma、N. Glaser、C.-W. Kuo 和 Z. Kira，《Who2com：通过可学习的握手通信实现协同感知》，载于《2020 IEEE国际机器人与自动化会议（ICRA）》。IEEE，2020 年，第 6876–6883 页。[28] A. Barel、R. Manor 和 A. M. Bruckstein，《Come together：多智能体几何一致性》，arXiv 预印本 arXiv:1902.01455，2017 年。[29] S. Tatikonda 和 S. Mitter，《通信约束下的控制》，IEEE自动控制汇刊，第 49 卷，1068，2004 7，1056 期，第 – 页。[30] J. N. Foerster、G. Farquhar、T. Afouras、N. Nardelli 和 S. Whiteson，《反事实多智能体策略梯度》，载于《第三十二届 AAAI人工智能大会》，2018 年。[31] F. A. Oliehoek、C. Amato 等，《去中心化部分可观测马尔可夫决策过程简明导论》。Springer，2016 年，第 1 卷。[32] Z. Ding、W. Hong、L. Zhu、T. Huang 和 Z. Lu，《多智能体强化学习中的序贯通信》，2021 年。[33] J. Albowicz、A. Chen 和 L. Zhang，《传感器网络中的递归位置估计》，载于第九届国际网络协议会议论文集（*ICNP 2001*）。IEEE，2001年，第 35–41 页。[34] S. Dorvash 和 S. Pakzad，《随机迭代模态识别算法及其在无线传感器网络中的应用》，《结构控制与健康监测》，第 20 卷，第 8 期，第 1121–1137 页，2013年。[35] D. V. Pynadath 和 M. Tambe，《通信多智能体团队决策问题：团队合作理论与模型分析》，《人工智能研究杂志》，第 16 卷，第 389–423 页，2002年6月。[36] F. A. Oliehoek、M. T. Spaan、N. Vlassis等，《具有延迟通信的分散式部分可观测马尔可夫决策过程》，载于不确定领域中的多智能体序贯决策会议论文集，檀香山，夏威夷，2007年5月。[37] R. Lowe、J. Foerster、Y.-L. Boureau、J. Pineau 和 Y. Dauphin，《涌现式通信测量中的陷阱》，载于自主智能体与多智能体系统国际会议，2019年。[38] L. Li、T. J. Walsh 和 M. L. Littman，《迈向马尔可夫决策过程状态抽象的统一理论》，载于*AI&M*，2006年。[39] A. K. McCallum，具有选择性感知和隐藏状态的强化学习。罗切斯特大学，1996年。[40] R. S. Sutton 和 A. G. Barto，《强化学习导论》，第 2 版。麻省理工学院出版社，2017年11月，第 135 卷。[41] V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski等，《通过深度强化学习实现人类水平控制》，《自然》，第 518 卷，第 7540 期，第 529–533 页，2015年。[42] P. Xuan、V. Lesser 和 S. Zilberstein，《多智能体协作中的通信决策：模型与实验》，载于第五届自主智能体国际会议论文集，丛书 AGENTS '01。美国纽约州纽约市：计算机协会，2001年，第 616 623 [ ] 10 1145 375735 376469 – 页。[43] 在线。获取地址：https://doi.org/ 。/ 。C. Amato、J. S. Dibangoye 和 S. Zilberstein，《有限时域分散式部分可观测马尔可夫决策过程的增量式策略生成》，载于第十九届自动规划与调度国际会议，2009年。