多智能体强化学习中的低熵通信

Lebin Yu、Yunbo Qiu、Qiexiang Wang、Xudong Zhang、Jian Wang 电子工程系，
清华大学。中国北京 100084 邮箱：{yulb19、qyb18、wqx19}@mails.tsinghua.edu.cn，{
zhangxd、jian-wang}@tsinghua.edu.cn

arXiv:2302.05055v1 [cs.MA] 10 Feb 2023

摘要——**近年来，多智能体强化学习中的通信因其在协作中发挥的重要作用而备受关注。然而，多智能体系统可能受到通信资源限制，因此在现实场景中需要高效的通信技术。根据香农-哈特利定理，在较差信道中，要可靠传输的消息需要具有更低的熵。因此，我们旨在降低多智能体通信中的消息熵。一个根本挑战是，熵的梯度要么为 0，要么为 ∞，这使得基于梯度的方法无法使用。为了解决这一问题，我们提出了一种伪梯度下降方案，通过巧妙地调整消息分布来降低熵。我们在两个基础通信框架和六种环境设置下进行了实验，结果表明，我们的方案可将消息熵降低多达 90% ，而协作性能几乎没有损失。**

一、引言

近年来，多智能体强化学习（MARL）因其在协作任务中的广泛应用而日益受到关注，例如机器人导航[1]，、交通信号灯控制 [2]和大规模车队管理[3]。为促进智能体之间的协作，一些研究人员设计了智能体之间的通信框架，利用神经网络生成消息 [4]–[8]。

然而，许多多智能体通信框架仅以协作性能作为衡量指标，而不考虑通信效率。因此，这些框架会随意生成并发送消息，使其不适用于通信资源受限的场景 [9]–[11]。为了解决这一问题，一些研究人员设计了高效的多智能体通信协议，这些协议可分为两类。第一类是减少通信次数 [12]–[16]，，包括明智地选择通信时机和通信对象。第二类是降低生成消息的熵。其动机在于，根据香农-哈特利定理 [17]，熵较低的消息所需带宽较少，因此能够在较差的通信信道中可靠传输。大多数基于学习的多智能体通信框架使用实值连续向量进行通信，因此，当前所有降低消息熵的方法都聚焦于微分熵[1] [19]–[21]。

然而，这些方法存在三个缺陷。首先，其中一些方法依赖专门设计的架构来最小化消息熵 [21]，，这损害了它们的泛化能力。其次，在没有先验信息的情况下，微分熵难以估计；而且有些方法简单地将消息分布视为单一高斯分布，这可能与实际情况相去甚远。第三，降低微分熵不如降低量化消息的离散熵重要。这是因为在现代通信系统中，连续变量必须先量化为离散变量才能传输，因此在高效通信研究中，离散熵比微分熵重要得多。

本文提出一种称为离散熵最小化（简称 DisEM）的方案，可应用于常见的基于学习的多智能体通信框架，在几乎不降低协作性能的情况下减少量化消息的离散熵。核心问题在于，量化会截断梯度，使所有基于梯度的训练算法都无法使用。为克服这一挑战，我们提出了一种新颖的伪梯度下降法，根据精心设计的伪梯度调整消息分布，从而降低离散熵。我们还从理论上证明了该方法的有效性。直观而言，DisEM 会使消息变量从不太常见的量化区间移动到相邻的、更常见的区间，从而改变消息分布。因此，DisEM 不会过度改变消息分布，进而能够在性能几乎不受影响的情况下降低熵。为通过实证说明 DisEM 的有效性，我们在三个通信关键型多智能体任务中开展实验，共设置六种实验条件。同时，我们将该方案应用于两个常见的多智能体通信框架 IC3NET [22] 和 TARMAC [23]，，以展示其泛化能力。实验表明，在某些设置下，DisEM 在不降低协作性能的情况下，最多可将熵降低 90%。

II. 系统模型

A. 带通信的多智能体强化学习

我们考虑一个智能体之间能够通信的部分可观测 n-智能体马尔可夫博弈[24] 。这一过程可以用元组 $\langle S, A, R, T, O, \Omega\ M_S,\ n,\ \gamma\rangle$ 描述，其中 S表示环境的状态空间，A 表示可用动作集合，R 是奖励函数 $R\colon S \times A \to \mathbb{R}$，$T$是转移函数 $T\colon S \times A \to S$，O 是智能体的观测空间，Ω 是智能体的观测函数：$\Omega\colon S \to O$，M_S 表示消息空间，n 表示

[1]为便于阅读，遵循 [18]，，我们用离散熵表示离散变量的熵，用微分熵表示连续变量的熵。

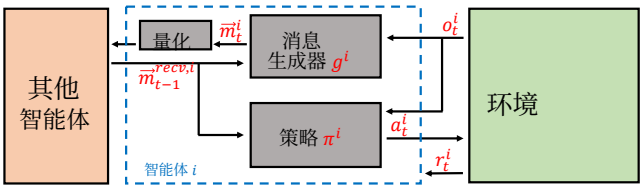


图1. 智能体 i 与其他智能体及环境的交互方式。

智能体的数量，γ 是折扣因子。在采用通信的基础多智能体强化学习框架中，智能体 i 需要一个策略 π^i 和一个消息生成器 g^i 来完成任务。在时间步t，智能体 i 首先从环境中获取观测 o_t^i，并接收来自其他智能体的消息 $\vec{m}^{recv,\ i}_{t-1}$：

$$\vec{m}_{t-1}^{recv,i} = \{\vec{m}_{t-1}^1, ..., \vec{m}_{t-1}^{i-1}, \vec{m}_{t-1}^{i+1}, ..., \vec{m}_{t-1}^n\}$$

其次，它会做出决策 $a_t^i = \pi^i(o_t^i,\ \vec{m}^{recv,\ i}_{t-1})$，并向其他智能体广播消息 $\vec{m}_t^i = g^i(o_t^i,\ \vec{m}^{recv,\ i}_{t-1})$。最后，它从环境中获得奖励 r_t^i。我们在图1中展示了智能体 i与其他智能体及环境的交互方式。

在本文中，我们关注基于学习的多智能体通信框架，其中 g^i 通过神经网络实现，并使用基于梯度的方法进行训练。在这种情况下，我们可以用 θ_i 表示其策略和消息生成器的参数，训练目标是最大化目标函数 $J(\theta_i)$：

$$J(\theta_i) = \mathbb{E}_{\theta_i}\left[\sum_{t=0}^{\infty}\gamma^t r_t^i\right] \tag{1}$$

借助策略梯度方法 [25], θ_i 可以根据目标函数 $\nabla_{\theta_i}J(\theta_i)$的梯度进行更新。

B. 通信设置

在我们的系统中，智能体生成的消息会先经过量化，然后广播给其他智能体。具体而言，我们将智能体消息生成器的输出范围设为 $[-1, 1]$，并使用量化区间长度为 Δ 的均匀量化函数 $f^Q(x)$：

$$f^Q(x) = k\Delta - 1,\ \forall x \in [(k-0.5)\Delta - 1, (k+0.5)\Delta - 1) \tag{2}$$

其中，$k \in \{0, 1, ..., K\}$、$K = 2/\Delta$和 $K+1$是量化区间的数量。

在主要实验中，我们假设信息传输无损，因为我们的目标是降低信息源（即消息生成器）的熵，而这与信道条件无关。

III. 方法

A. 量化消息的熵

为表述如何计算量化消息的熵，我们定义如下 $h_k(\cdot)$：

$$h_k(x) = \begin{cases} 1, & x \in [(k-0.5)\Delta - 1, (k+0.5)\Delta - 1) \\ 0, & x \notin [(k-0.5)\Delta - 1, (k+0.5)\Delta - 1) \end{cases} \tag{3}$$

不失一般性，我们首先关注消息长度为 1 的情形，这意味着一条消息是一个数，而不是一个向量。给定一个消息集合$M = \{m_1, m_2, ..., m_N\}$，量化消息的熵计算如下：

$$H(M) = -\sum_{k=0}^{K}\frac{\epsilon + \sum_{i=1}^N h_k(m_i)}{N}\log\frac{\epsilon + \sum_{i=1}^N h_k(m_i)}{N} \tag{4}$$

其中，ϵ 是一个较小的数，用于避免计算时出现 log0。如果消息长度大于 1，则可以通过将每一位的熵相加来计算熵。

B. 在不损害协作的情况下降低熵

我们的目标是在不损害性能的情况下，降低基于梯度的多智能体通信框架的消息熵。具体而言，对于智能体 i，首先使用基线框架训练 θ_i，直至 π^i，g^i 得到充分训练。其次，我们尝试在不损害协作性能的情况下降低消息的熵。一种常见的方法是在原始训练目标中添加一个正则项 [20], [26]：

$$\max_{\theta_i} J'(\theta_i) = J(\theta_i) - \alpha H(M_i) \tag{5}$$

其中，α 是正则项的权重，其取值取决于最大化 $J(\theta_i)$ 与最小化$H(M_i)$ 之间的权衡。于是，θ_i 相对于这一新目标的梯度变为：

$$\nabla_{\theta_i}J'(\theta_i) = \nabla_{\theta_i}J(\theta_i) - \alpha\nabla_{\theta_i}H(M_i) \tag{6}$$

请注意，$h_k(\cdot)$ 的梯度要么为 0，要么为 ∞。因此，$\nabla_{\theta_i}H(M_i)$ 要么为 0，要么为 ∞。这使得基于梯度的训练方法无法使用，因而难以降低 $H(M_i)$。我们提出一种伪梯度下降法来解决这一问题。

C. 使用伪梯度下降降低熵

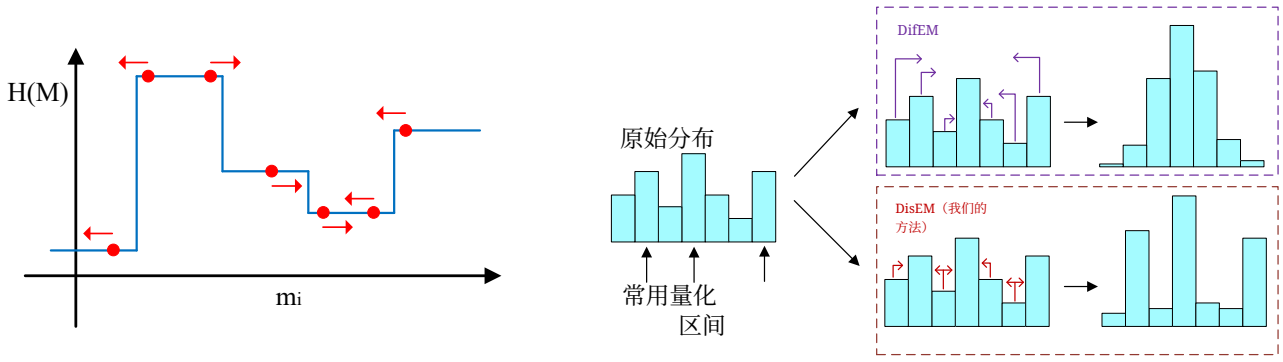
$H(M)$可视为一个具有 N个变量 $H(m_1, m_2, ..., m_N)$的多元函数，在这一部分中，我们将重点讨论如何通过调整 m_i来降低 $H(M)$。首先，我们给出梯度下降方法的核心函数：给定一个连续可微函数 $f(x_1, x_2, ..., x_n)$和 $\eta \to 0$，

$$\begin{aligned} x_i' &= x_i - \eta\nabla_{x_i}f \\ f(x_1, ..., x_i', ..., f_n) &\le f(x_1, ..., x_i, ..., f_n) \end{aligned} \tag{7}$$

由于 $\nabla_{m_i}H(M) = \sum_k \nabla_{h_k}H(M)\nabla_{m_i}h_k = 0$或 ∞，我们无法使用梯度下降来最小化 $H(M)$。因此，我们尝试设计一个伪梯度 $\nabla^p_{m_i}H(M)$，以实现与(7)类似的效果。

我们的核心思路是将 $\nabla_{m_i}h_k$ 替换为 $s_k(m_i)$：

$$s_k(x) = \begin{cases} 1, & x \in ((k-1)\Delta - 1, k\Delta - 1) \\ -1, & x \in (k\Delta - 1, (k+1)\Delta - 1) \\ 0, & x = k\Delta - 1 \end{cases} \tag{8}$$



(a) (b)

图 2。（a）我们如何通过调整 m_i，使伪梯度下降法降低 $H(M)$。红点的 x 坐标表示 m_i 的一个可能取值，其 y 坐标表示相应的 $H(M)$ 值。红色箭头表示伪梯度下降的方向。伪梯度下降会使 m_i 朝着可能降低 $H(M)$ 的方向移动。（b）我们的方法（离散熵最小化，简称 DisEM）和传统方法（微分熵最小化，简称 DifEM）如何改变消息的分布。y 轴表示概率，x 轴表示消息变量的值。DisEM 通过将消息的值从不太常用的量化区间移至相邻的、更常用的量化区间来降低熵，而 DifEM 则只是使分布更接近低方差高斯分布。

伪梯度 $\nabla^p_{m_i} H$（M）的表达式为：

$$\nabla^p_{m_i} H(M) = \sum_k \nabla_{h_k} H(M) s_k(m_i) \tag{9}$$

接下来，我们证明伪梯度下降具有与梯度下降相似的性质。给定 $H(M) = H(m_1, \ldots\ldots, m_i, \ldots\ldots, m_n)$ 和 $\eta \to 0$，

$$\begin{aligned} m_i' &= m_i - \eta \nabla^p_{m_i} H(M) \\ H(m_1, ..., m_i', ..., m_N) &\le H(m_1, ..., m_i, ..., m_n) \end{aligned} \tag{10}$$

证明见下一小节。

我们在图 2(a) 中直观展示伪梯度下降方法如何调整 m_i，并在图 2(b) 中展示我们的方法和以往方法（旨在最小化微分熵的方法）如何改变消息的分布。简而言之，我们的方法通过将消息值从较不常见的量化区间移至相邻的较常见量化区间来降低熵，同时不会过多改变消息。相比之下，传统方法只是让分布更接近低方差高斯分布，而不考虑原始分布是什么。

D. 式 (10) 的证明

为简洁起见，我们设 $N_k = \epsilon + \sum_{i=1}^{N} h_k(m_i)$。注意，$N_k$ 表示落入量化区间 $[(k-0.5)\Delta - 1, (k+0.5)\Delta - 1)$ 的消息变量数量。为证明 (10)，我们给出三个引理：

引理 III.1。如果 $m_i \in (u\Delta - 1, (u+1)\Delta - 1)$，则

$$\begin{aligned} \nabla^p_{m_i} H(M) &> 0, \ if \ N_u > N_{u+1} \\ \nabla^p_{m_i} H(M) &< 0, \ if \ N_u < N_{u+1} \\ \nabla^p_{m_i} H(M) &= 0, \ if \ N_u = N_{u+1} \end{aligned} \tag{11}$$

注释 该引理表明，对于两个相邻的量化区间，伪梯度下降法会使消息变量从较不常用的区间移动到较常用的区间。

证明。

$$\begin{aligned} \nabla^p_{m_i} H(M) &= \sum_{k=0}^{K} \nabla_{h_k} H(M) s_k(m_i) \\ &= -\frac{1}{N} \sum_{k=0}^{K} (1 + \log \frac{N_k}{N}) s_k(m_i) \end{aligned} \tag{12}$$

注意到 $\forall m_i \in (u\Delta - 1, (u+1)\Delta - 1)$、$s_u(m_i) = -1$、$s_{u+1}(m_i) = 1$ 以及 $s_k(m_i) = 0 \ \forall k \in \{0, 1, ..., K\} \setminus \{u, u+1\}$。因此，我们得到

$$\nabla^p_{m_i} H(M) = -\frac{1}{N} log \frac{N_{u+1}}{N_u} \tag{13}$$

于是我们可以得到引理 III.1 中的方程 □

引理 III.2。如果将 $N_u > N_{u+1}$ 和 $m_i \in (u\Delta - 1, (u+1)\Delta - 1)$ 更新为 m_i'，其中 $m_i' = m_i - \eta \nabla^p_{m_i} H(M)$ $\eta \to 0$，则此次更新只会导致以下两种可能的结果：

(a) $N_k' = N_k, \ \forall k \in \{0, 1, ..., K\}$

(b) $N_u' = N_u + 1, \ N_{u+1}' = N_{u+1} - 1$ *and* $N_k' = N_k, \ \forall k \in \{0, 1, ..., K\} \setminus \{u, u+1\}$

If $N_u < N_{u+1}$，类似地，*It* 的更新也只会导致两种可能的结果：

(a) $N_k' = N_k, \ \forall k \in \{0, 1, ..., K\}$

(b) $N_u' = N_u - 1, \ N_{u+1}' = N_{u+1} + 1$ *and* $N_k' = N_k, \ \forall k \in \{0, 1, ..., K\} \setminus \{u, u+1\}$

证明。我们首先关注第一种情况（即 $N_u > N_{u+1}$）。由于 $\nabla^p_{m_i} H(M) > 0$，

$$m_i' = m_i - \eta \nabla^p_{m_i} H(M) \in (m_i - \eta \max(\nabla^p_{m_i} H(M)), \ m_i) \tag{14}$$

其中

$$\max(\nabla^p_{m_i} H(M)) = \frac{1}{N} \log \frac{N_s + \epsilon}{\epsilon} \tag{15}$$

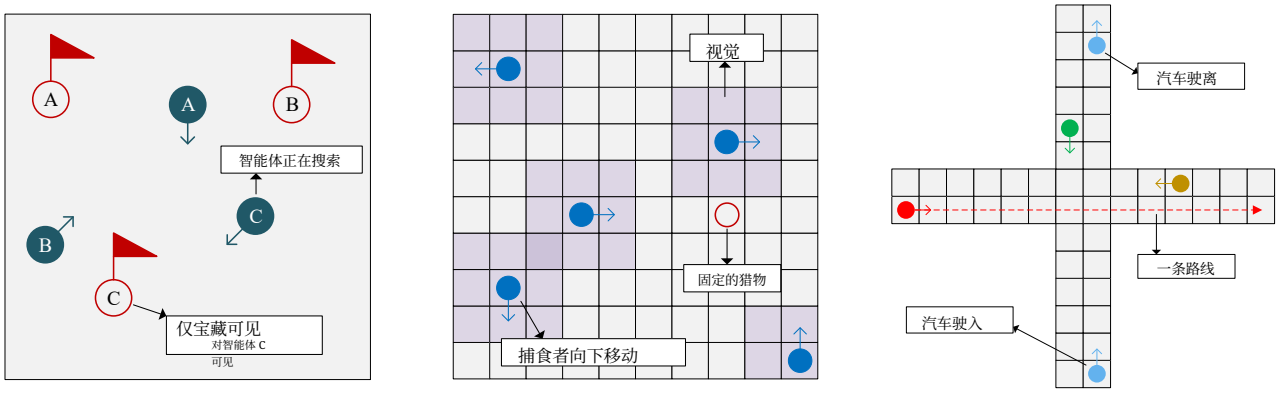


图 3. 我们的三个实验环境的可视化：寻宝、捕食者-猎物和交通路口。

由于 $\eta \to 0$, $\eta \max(\nabla^p_{m_i} H(M)) < 0.5\Delta$。注意， $m_i \in (u\Delta-1,(u+1)\Delta-1)$，因此 $m'_i \in ((u-0.5)\Delta-1,(u+1)\Delta-1)$。因此， m_i 和 m'_i的值只有两种可能：

(a) 若 $m_i \in ((u+0.5)\Delta-1,\ (u+1)\Delta-1)$ 且 $m'_i \in ((u-0.5)\Delta-1,\ (u+0.5)\Delta-1)$，则 $N'_u = N_u+1$, $N'_{u+1} = N_{u+1}-1$ 且 $N'_k = N_k$, $\forall k \in \{0,1,...,K\} \setminus \{u, u+1\}$(b) 否则， $N'_k = N_k$, $\forall k \in \{0,1,...,K\}$

第二种情况下的证明类似（即 $N_u < N_{u+1}$） □

用 $H(M)$ 表示 $H(m_1, ... m_i, ... m_N)$，用 $H(M')$ 表示 $(m_1, ... m'_i, ... m_N)$，我们提出第三个引理：

引理 III.3. $H(M') < H(M)$ 若 $N_u > N_{u+1}$, $N'_u = N_u+1$, $N'_{u+1} = N_{u+1}-1$ 且 $N'_k = N_k$, $\forall k \in \{0,1,...,K\} \setminus \{u, u+1\}$。

$H(M') < H(M)$ 若 $N_u < N_{u+1}$, $N'_u = N_u - 1$, $N'_{u+1} = N_{u+1}+1$ 且 $N'_k = N_k$, $\forall k \in \{0,1,...,K\} \setminus \{u, u+1\}$。

证明。 我们首先关注第一种情况（即 $N_u > N_{u+1}$）。

$$H(M') - H(M) = \frac{N_u}{N}\log\frac{N_u}{N} + \frac{N_{u+1}}{N}\log\frac{N_{u+1}}{N} - \frac{N_u+1}{N}\log\frac{N_u+1}{N} - \frac{N_{u+1}-1}{N}\log\frac{N_{u+1}-1}{N} \tag{16}$$

为简洁起见，设 $x_1 = \frac{N_u}{N}$, $x_2 = \frac{N_{u+1}}{N}$, $\delta = \frac{1}{N}$和 $f(x) = x\log x$。于是得到：

$$H(M') - H(M) = f(x_2) - f(x_2-\delta) - (f(x_1+\delta) - f(x_1)) \tag{17}$$

根据中值定理，存在 $x_{1m} \in (x_1, x_1+\delta)$ 和 $x_{2m} \in (x_2-\delta, x_2)$，使得 $f(x_2) - f(x_2-\delta) = \delta f'(x_{2m})$ 且 $f(x_1+\delta) - f(x_1) = \delta f'(x_{1m})$。由于$f''(x) > 0$且 $x_{2m} < x_2 < x_1 < x_{1m}$, $H(M') - H(M) < 0$。第二种情况的证明类似。 □

由引理 III.2 和引理 III.3 可得，若$H(M') \neq H(M)$，则 $H(M') < H(M)$。因此，我们得到$H(M') \leq H(M)$。

E. 实现细节

遵循策略梯度方法和我们提出的伪梯度下降方案，智能体 i 的参数 θ_i 可根据以下梯度表达式进行训练：

$$\nabla_{\theta_i} J'(\theta_i) = \nabla_{\theta_i} J(\theta_i) - \alpha \nabla^p_{\theta_i} H(M_i) \tag{18}$$

此外，我们首先在不使用熵正则项的情况下训练智能体（即设置 $\alpha = 0$），使其优化策略和消息生成器。经过 T_N 个训练轮次后，我们加入正则项（即设置 $\alpha = \alpha_p$），并继续训练 $T_{max} - T_N$ 个训练轮次。 T_N, α_p和 T_{max}是预定义的超参数，其取值取决于具体任务。至于量化函数，我们设置为 $\Delta = 0.25$。

IV. 实验

A. 环境

为了演示，我们考虑三种环境，每种环境有两种设置：寻宝（TH）、捕食者-猎物（PP）和交通路口（TJ） [22]，如图 3 所示。这些环境都对通信至关重要，因此降低消息熵具有实际意义。

寻宝 在此任务中， N 个智能体协同在场地中寻找宝藏。每个智能体都能获得其目标宝藏的坐标，而其他智能体无法看到该宝藏。需要注意的是，智能体无法独自收集自己的宝藏，而应通过通信帮助其他智能体找到它。场地大小为 1×1, 智能体可以以速度 v向八个方向移动。如果所有宝藏都被找到，或时间步达到上限 t_{max}，则回合结束。因此，时间步越少，性能越好。在设置 A（TH-A）中， $N = 3$, $v = 0.15$ 和 $t_{max} = 20$。在设置 B（TH-B）中， $N = 6$, $v = 0.09$ 和$t_{max} = 60$。至于训练超参数， $T_N = 100$、$\alpha_p = 0.2$ 和 $T_{max} = 200$。

捕食者-猎物 [22]在此任务中， N 具有有限视野的智能体需要在大小为 $D \times D$ 的网格世界中到达固定的猎物。如果所有智能体都到达猎物，或时间步达到上限 t_{max}，则一个回合结束。因此，时间步越少表示性能越好。在设置 A（PP-A）中，

表 I 六种设置下八种通信框架的协作性能和消息熵。

		基于 IC3NET				基于 TARMAC			
		ORI	ZC	DifEM	DisEM	ORI	ZC	DifEM	DisEM
TH-A	时间步↓	8.3 ±0.3	19.7 ±0.1	9.1 ±0.4	9.0 ±0.3	8.2 ±0.2	19.7 ±0.2	10.0 ±0.5	9.2 ±0.2
	熵↓	139 ±7	0	70 ±2.8	15 ±1	70 ±1	0	29 ±1	9 ±1
TH-B	时间步↓	24.4 ±1.0	59.9 ±0.0	42.9 ±6.5	24.1 ±0.7	21.0 ±0.6	59.9 ±0.0	36.3 ±6.0	22.8 ±0.4
	熵↓	127 ±4	0	39 ±26	30 ±1	72 ±4	0	25 ±6	15 ±1
PP-A	时间步↓	10.1 ±0.1	17.4 ±0.5	14.4 ±2.5	10.7 ±0.4	10.0 ±0.2	17.5 ±0.2	11.3 ±0.4	10.4 ±0.1
	熵↓	94 ±3	0	20 ±14	9 ±1	60 ±3	0	13 ±1	6 ±1
PP-B	时间步↓	16.5 ±1.4	33.1 ±0.7	17.5 ±0.5	16.3 ±0.7	16.0 ±1.6	33.6 ±0.5	16.8 ±0.5	15.4 ±0.5
	熵↓	122 ±11	0	67 ±6	27 ±3	58 ±5	0	25 ±3	14 ±1
TJ-A	成功率↑	0.87 ±0.13	0.29 ±0.02	0.66 ±0.05	0.85 ±0.11	0.77 ±0.18	0.28 ±0.02	0.71 ±0.04	0.74 ±0.21
	熵↓	141 ±10	0	96 ±26	83 ±10	27 ±4	0	27 ±13	27 ±4
TJ-B	成功率↑	0.95 ±0.02	0.74 ±0.02	0.73 ±0.02	0.93 ±0.01	0.95 ±0.02	0.72 ±0.02	0.90 ±0.09	0.95 ±0.01
	熵↓	75 ±30	0	77 ±32	38 ±29	44 ±2	0	22 ±10	14 ±6

$N = 3$、 $D = 5$、 $t_{max} = 20$ ，视觉设置为 0。在设置 B（PP-B）中， $N = 5$、 $D = 10$、 $t_{max} = 40$ ，视觉设置为 1。至于训练超参数， $T_N = 100$、 $\alpha_p = 0.05$和 $T_{max} = 150$。

交通路口 [27]在此任务中，汽车以概率 p_{arr} 从所有入口进入路口。路口中汽车的最大数量设为 N。每辆汽车都分配了一条固定路线，并且每一步都有两个选择：沿路线前进或停留。汽车需要快速完成路线，同时避免发生碰撞。如果经过 t_{max} 后没有发生碰撞，则本回合计为成功，否则计为失败。值得注意的是，汽车只能观察自身位置，因此必须通过通信来了解其他汽车的位置，从而避免碰撞。在设置 A（TJ-A）中， $N = 5$, $p_{arr} = 0.3$ 和 $t_{max} = 20$。在设置 B（TJ-B）中， $N = 10$, $p_{arr} = 0.05$ 和 $t_{max} = 40$。至于训练超参数， $T_N = 1000$, $\alpha_p = 0.05$ 和$T_{max} = 1250$。

B. 测试算法

为测试我们方案的有效性，所测试的算法基于两种多智能体通信框架：IC3NET [22] 和 TARMAC [23]。具体而言，我们按照下文所述，分别测试每种框架的四个变体。

原始版本（ORI）：这是未经任何修改的框架原始版本，体现了基线协作性能和熵。

无通信（ZC）：该变体通过强制将所有消息设为零来禁用通信，体现了无通信时的性能。

DifEM：[20] 一种通过互信息正则项最小化微分熵的既有方法。

DisEM：这就是我们提出的方案，通过在原始损失函数中添加伪熵正则项来降低熵。

C. 主要实验结果

我们在这六种设置下，使用总计八种方法开展主要实验。为提高结果的可靠性，我们对每种方法在每种设置下重复训练五次，并对每个模型测试 500 回合。随后，我们在表 I 中展示平均结果。需要强调的是，我们的 DisEM 方案旨在实现良好的协作性能，同时保持较低的熵[2]。请注意，在 TH 和 PP 场景中，时间步越少表示协作越好；而在 TJ 场景中，成功率越高表示协作越好。

我们从结果中得出以下几点。(a) ZC（零通信）在几乎所有设置下表现不佳，尤其是在 TH-A 和 TH-B 中，智能体完全无法在没有通信的情况下完成任务。这反映了通信在这些设置中的重要性。(b) DifEM 确实以不同程度的性能下降为代价降低了通信熵。(c) DisEM 的表现优于 DifEM，能够在几乎不降低或完全不降低性能的情况下减少熵。更准确地说，它在 TH 和 PP 设置中表现相当出色，可将熵降低 80% $-$ 90%。然而，在 TJ-A 设置中，基于 TARMAC 的 DifEM 和 DisEM 均未能降低熵。这是因为该设置下的消息分布难以优化。

D. 探索性实验：缩短消息长度

缩小层大小是一种高效且常见的信息压缩方式，自动编码器中便采用了这种方法 [28]。由于消息向量由神经网络生成，且其长度在训练前手动设定，因此我们可以缩短消息向量的长度来降低熵。在这项探索性实验中，我们测试了缩短消息长度在降低熵方面是否优于我们的方法。具体而言，我们在捕食者-猎物环境中测试了不同消息长度下的 IC3NET-Original 和 IC3NET-DisEM，并在图 4 中展示结果。每个点代表一个模型的性能，其横坐标表示熵，纵坐标表示时间步。因此，位于左下方的点意味着性能良好且熵较低。在 PP-A 中，DisEM 显著降低了通信熵，同时仅造成轻微的性能下降。在 PP-B 中，DisEM 大幅降低了

[2]基于 TARMAC 的方法生成的消息熵低于基于 IC3NET 的方法，这是由原始论文中不同的设置所致。IC3NET 的默认消息大小为 128，而 TARMAC 的默认消息大小为 48。

通信熵，同时保持了基线性能。总之，尽管缩短 IC3NET-Original 的消息长度可以降低熵，但与我们的方法结合使用效果更好。

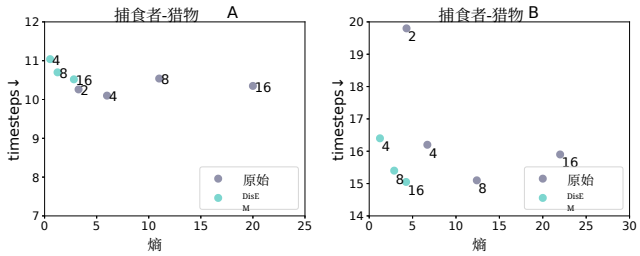


图4. IC3NET-Original 和 IC3NET-DisEM 在捕食者-猎物环境中的性能。数据点旁的数字表示其消息长度，数据点的颜色表示该模型是 IC3NET-Original 还是 IC3NET-DisEM。

V. 结论

本文提出了一种简单而有效的方案 DisEM，用于降低常见的基于学习的多智能体通信框架中的通信熵。首先，我们指出，在多智能体通信中，最小化量化消息的熵是必要的。其次，针对熵无法通过梯度下降进行优化这一问题，我们设计了伪梯度下降，通过将消息变量从较不常用的量化区间移至相邻且更常用的量化区间来降低熵。第三，我们从理论上证明了伪梯度下降的有效性。第四，我们开展了大量实验来测试该方案。具体而言，我们在六种通信关键环境设置中，对两种基础多智能体通信框架 IC3NET 和 TARMAC 的八种变体进行了测试。结果证实，我们的方案 DisEM 优于现有方案：DisEM 可将消息熵降低多达 90%，且几乎不会损失协作性能。我们还开展了几项探索性实验，发现将 DisEM 与缩小消息大小相结合能够进一步降低消息熵。我们希望这项工作能为更先进的高效多智能体通信研究奠定基础。

参考文献

[1] R. Han、S. Chen 和 Q. Hao，《利用深度强化学习实现动态环境中的协作式多机器人导航》，载于*2020*年*IEEE*国际机器人与自动化会议*(ICRA)*。IEEE，2020年，第448–454页。[2] J. A. Calvo 和 I. Dusparic，《用于交通信号灯控制的异构多智能体深度强化学习》。载于*AICS*，2018年，第2–13页。[3] K. Lin、R. Zhao、Z. Xu 和 J. Zhou，《通过多智能体深度强化学习实现高效的大规模车队管理》，载于《第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，2018年，第1774–1783页。[4] J. Foerster、I. A. Assael、N. de Freitas 和 S. Whiteson，《利用深度多智能体强化学习学习通信》，《神经信息处理系统进展》，第29卷，第2137–2145页，2016年。[5] S. Ahilan 和 P. Dayan，《修正多智能体通信中的经验回放》，载于《国际学习表征会议》，2020年。

[6] T. Chu、S. Chinchali 和 S. Katti，《用于网络化系统控制的多智能体强化学习》，载于《国际学习表征会议》，2019年。[7] W. Kim、J. Park 和 Y. Sung，《多智能体强化学习中的通信：意图共享》，载于《国际学习表征会议》，2020年。[8] Y. Wang、J. Xu、Y. Wang 等，《Tom2c：目标导向的多智能体通信与心智理论协作》，载于《国际学习表征会议》，2021年。[9] M. Rangwala 和 R. Williams，《通过结构化注意推理学习多智能体通信》，载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin 编，第33卷。Curran Associates, Inc.，2020年，第10 088–10 098页。[10] A. Serra-G´omez、B. Brito、H. Zhu、J. J. Chung 和 J. Alonso-Mora，《与谁通信：学习高效通信以实现多机器人避碰》，载于《2020年IEEE/RSJ智能机器人与系统国际会议 (IROS)》。IEEE，2020年，第11 770–11 776页。[11] C. Sun、M. Shen 和 J. P. How，《面向机器人系统扩展多智能体强化学习：学习自适应稀疏通信图》，载于《2020年IEEE/RSJ智能机器人与系统国际会议 (IROS)》。IEEE，2020年，第11 755–11 762页。[12] Z. Ma、Y. Luo 和 J. Pan，《为多智能体路径规划学习选择性通信》，arXiv预印本 arXiv:2109.05413，2021年。[13] D. Kim、S. Moon、D. Hostallero、W. J. Kang、T. Lee、K. Son 和 Y. Yi，《学习在多智能体强化学习中调度通信》，载于《国际学习表征会议》，2018年。[14] V. K. Vijay、H. Sheikh、S. Majumdar 和 M. Phielipp，《在提升多智能体强化学习性能的同时最大限度减少通信》，arXiv预印本 arXiv:2106.08482，2021年。[15] H. Xiufeng 和 Z. Sheng，《面向多智能体强化学习的重要性感知消息交换与预测》，载于《2022年IEEE全球通信大会》。IEEE，2022年。[16] G. Hu、Y. Zhu、D. Zhao、M. Zhao 和 J. Hao，《具有带宽限制的多智能体强化学习事件触发通信网络》，《IEEE神经网络与学习系统汇刊》，2021年。[17] C. E. Shannon，《通信的数学理论》，《贝尔系统技术期刊》，第27卷，第3期，第379–423页，1948年。[18] T. M. Cover，《信息论基础》。John Wiley & Sons，1999年。[19] T. Wang、J. Wang、C. Zheng 和 C. Zhang，《通过通信最小化学习近似可分解价值函数》，载于《国际学习表征会议》，2019年。[20] R. Wang、X. He、R. Yu、W. Qiu、B. An 和 Z. Rabinovich，《学习高效的多智能体通信：一种信息瓶颈方法》，载于《国际机器学习会议》。PMLR，2020年，第9908–9918页。[21] S. Q. Zhang、Q. Zhang 和 J. Lin，《通过时间消息控制实现简洁且稳健的多智能体通信》，《神经信息处理系统进展》，第33卷，2020年。[22] A. Singh、T. Jain 和 S. Sukhbaatar，《学习在大规模合作与竞争任务中何时进行通信》，载于《国际学习表征会议》，2018年。[23] A. Das、T. Gervet、J. Romoff、D. Batra、D. Parikh、M. Rabbat 和 J. Pineau，《Tarmac：目标导向的多智能体通信》，载于《国际机器学习会议》。PMLR，2019年，第1538–1546页。[24] M. L. Littman，《马尔可夫博弈作为多智能体强化学习的框架》，载于《机器学习论文集 1994》。Elsevier，1994年，第157–163页。[25] R. J. Williams，《用于联结主义强化学习的简单统计梯度跟踪算法》，《机器学习》，第8卷，第3期，第229–256页，1992年。[26] N. Tishby、F. C. Pereira 和 W. Bialek，《信息瓶颈方法》，arXiv预印本 physics/0004057，2000年。[27] S. Sukhbaatar、R. Fergus 等，《通过反向传播学习多智能体通信》，《神经信息处理系统进展》，第29卷，第2244–2252页，2016年。[28] M. A. Kramer，《利用自联想神经网络进行非线性主成分分析》，《AIChE期刊》，第37卷，243，1991年2月，第233–页。