arXiv:2302.03429v1 [cs.AI] 7 Feb 2023

面向多智能体强化学习的熟练群体课程

Rundong Wang [*1] **Longtao Zheng** [*1 2] **Wei Qiu** [1] **Bowei He** [3] **Bo An** [1] **Zinovi Rabinovich** [1] **Yujing Hu** [2] **Yingfeng Chen** [2] **Tangjie Lv** [2] **Changjie Fan** [2]

摘要

近年来，多智能体强化学习（MARL）的进展使智能体能够在复杂环境中协调行为。然而，常见的多智能体强化学习算法仍面临可扩展性和稀疏奖励问题。一种有望解决这些问题的方法是自动课程学习（ACL）。ACL 通过学生（课程学习者）在教师（课程生成器）的控制下，逐步学习难度不断增加的任务。尽管取得了成功，ACL 的适用性仍受到以下因素限制：（1）缺少能够应对不同任务中智能体数量变化及稀疏奖励问题的通用学生框架；（2）由于学生策略不断变化，教师所面对的任务具有非平稳性。为解决 ACL 的这些问题，我们提出了一种新颖的自动课程学习框架——技能群体课程（SPC），将课程学习应用于多智能体协作。具体而言，我们赋予学生群体规模不变通信能力和分层技能集，使其能够从智能体数量各异的不同任务中学习合作与行为技能。此外，我们将教师建模为以学生策略为条件的上下文老虎机，使智能体团队能够改变规模，同时保留先前习得的技能。我们还分析了这一多智能体自动课程教学问题固有的非平稳性，并给出相应的遗憾界。实验结果表明，我们的方法在多个多智能体强化学习环境中提升了性能、可扩展性和样本效率。

[*]共同贡献 [1]南洋理工大学计算机科学与工程学院，新加坡 [2]网易伏羲人工智能实验室，中国 [3]香港城市大学计算机科学与工程系，中国。通信作者：RundongWang <rundong001@e.ntu.edu.sg>、Longtao Zheng<longtao001@e.ntu.edu.sg>。

预印本。

1. 引言

多智能体强化学习（多智能体强化学习）长期以来一直是复杂机器人和战略领域中的常用工具（RoboCup、2019年；OpenAI，2019年）。然而，对于大规模多智能体系统而言，如何从零开始利用稀疏奖励学习有效策略仍然充满挑战。其中一个挑战是，随着智能体数量增加，联合观测—动作空间呈指数增长。此外，稀疏奖励信号需要大量训练轨迹，这使得现有多智能体强化学习算法难以直接应用于复杂环境。因此，即使智能体之间进行协作会带来显著益处（Zhang 等人，2021年；Yang 和 Wang，2020年）。

针对稀疏奖励的大规模多智能体强化学习问题，已有多条相关研究路线，包括奖励塑形（Hu etal., 2020）、课程学习（Chen etal., 2021）和从示范中学习（Huang et al., 2021）。在这些方法中，课程学习范式尤其展现出良好的前景：在该范式下，所经历任务的难度和训练智能体群体的规模会逐步增长。在自动课程学习（自动课程学习，ACL）中，教师（课程生成器）会学习调整学生（课程学习器）所面对任务的复杂度和顺序。一些研究甚至提出了多智能体自动课程学习算法，基于近似或启发式教学方法，例如 DyMA-CL（Wang 等，2020c）、EPC（Long 等，2020）和 VACL（Chen 等人，2021）。然而，这些方法依赖于一种带有回放缓冲区的离策略学生框架，该框架忽略了智能体群体规模增大时出现的遗忘问题；或者，它们强烈假设智能体切换到不同任务时，已学习策略的价值不会发生变化。此外，由于学生策略不断变化，这些方法中的教师仍面临非平稳性问题。另一类大规模多智能体强化学习解决方案是分层学习，它利用时间抽象将任务分解为由多个子任务构成的层级。这类方法包括技能发现（Yang 等，2019）、将选项作为响应（Vezhnevets 等，2019）、基于角色的多智能体强化学习（Wang 等，2020b）以及两个抽象层次（Pang 等，2019）。然而，这些

方法大多聚焦于智能体数量固定的单一特定任务，也未考虑已学习技能的可迁移性。本文将就这一问题阐述我们的见解：

结合自动课程学习与分层学习的原则，能否在多智能体强化学习中实现带有稀疏奖励的复杂合作？

具体而言，我们提出了一种新颖的自动课程学习算法——技能群体课程（SPC），旨在应对稀疏奖励条件下为大规模多智能体系统学习有效策略所面临的挑战。SPC 的核心思想是鼓励学生从智能体数量不同的任务中学习技能，这类似于团队运动中的球员通过逐步提高任务难度并增加协同配合的球员数量来进行训练。为此，SPC 由三个关键组件构成：

上下文老虎机教师采用基于循环神经网络的（Hochreiter & Schmidhuber，1997年）模仿模型来表示学生策略，并生成老虎机的上下文。

学生模块中的种群不变通信用于处理不同任务中的智能体数量变化。SPC 将每个智能体的消息视为一个词，并使用自注意力通信通道（Vaswani 等人，2017年），从而支持任意数量的智能体共享消息。

分层技能框架用于学生模块，在稀疏奖励设置下学习可迁移技能；智能体围绕一组共享的低层策略进行高层通信。

实证结果表明，我们的方法在多智能体粒子环境（MPE）中的若干任务（Lowe 等人，2017年）以及 Google Research Football（GRF）中具有挑战性的 5对5比赛（Kurach 等人，2019年）中取得了最先进的表现。[1]

2. 预备知识

去中心化部分可观测马尔可夫决策过程。合作式多智能体强化学习问题可以表述为一个去中心化部分可观测马尔可夫决策过程（去中心化部分可观测马尔可夫决策过程）（Bernstein 等人，2002年），其定义为一个元组 $\langle n, \boldsymbol{S}, \boldsymbol{A}, P, R, \boldsymbol{O}, \boldsymbol{\Omega}\ \gamma\rangle$，其中 n表示智能体数量。 $\boldsymbol{S}$ 表示全局状态空间。$\boldsymbol{A} = \{A_i\}_{i=1,\ \cdots,\ n}$ 表示所有智能体的动作空间。$\boldsymbol{O} = \{O_i\}_{i=1,\ \cdots,\ n}$ 表示所有智能体的观测空间。 P: $\boldsymbol{S} \times \boldsymbol{A} \to \boldsymbol{S}$ 表示状态转移概率函数。所有智能体共享相同的奖励，该奖励是状态和动作的函数，适用于智能体 R： $\boldsymbol{S} \times \boldsymbol{A} \to \mathbb{R}$。每个智能体 i 根据观测函数 $\boldsymbol{\Omega}(s,\ i)$： $\boldsymbol{S} \to O_i$接收私有观测 $o_i \in O_i$。 $\gamma \in [0, 1]$ 表示折扣因子。

多臂老虎机。多臂老虎机（MABs）是一种简单却十分强大的框架，可用于在不确定性下反复做出决策。在该框架中，学习者执行一系列动作，并在每次动作后立即观察相应的奖励。目标是在给定的一组 K 动作和特定时间范围 T内最大化总奖励。多臂老虎机的成功程度通常以遗憾来衡量，即多臂老虎机算法的累积奖励与最佳臂基准之间的差值。一个著名的多臂老虎机算法是 Exp3算法（Auer等人，2002年），该算法旨在提高选择优质臂的概率，并在时变奖励分布下实现 $O(\sqrt{KT\log(K)})$ 的遗憾。另一个相关概念是上下文赌博机问题（Hazan 和 Megiddo，2007），在该问题中，学习者根据上下文等先验信息做出决策。

3. 技能群体课程

本节首先对课程增强型Dec-POMDP框架进行形式化定义，该框架在去中心化部分可观测马尔可夫决策过程框架下对带有课程问题的多智能体强化学习进行建模。随后，如图 1 所示，我们提出多智能体主动课程学习算法——技能群体课程（SPC）。在以下小节中，我们将在第 3.1 节建立课程学习框架，然后在第 3.2 节提出一种上下文多臂老虎机算法作为教师，以解决非平稳性问题。最后，我们将在第 3.3 节介绍具备可迁移技能和群体规模不变通信的学生，以应对智能体数量变化和稀疏奖励问题。

3.1. 问题定义

我们考虑的多智能体自动课程学习问题中的环境配备了参数化任务空间，因此可以建模为课程增强型 Dec-POMDP。

定义 3.1（课程增强型Dec-POMDP）。课程增强型 Dec-POMDP由元组$\langle \Phi, \mathcal{M}\rangle$定义，其中 Φ 和 $\mathcal{M}$ 分别表示任务空间和去中心化部分可观测马尔可夫决策过程。给定任务 ϕ，去中心化部分可观测马尔可夫决策过程$\mathcal{M}(\phi)$表示为 $\left\{n^{\phi}, \boldsymbol{S}^{\phi}, \boldsymbol{A}^{\phi}, P^{\phi}, r^{\phi}, O^{\phi}, \boldsymbol{\Omega}^{\phi},\ \gamma^{\phi}\right\}$。上标 ϕ 表示去中心化部分可观测马尔可夫决策过程的各个元素由任务 ϕ决定。请注意，任务 ϕ 可以是环境的若干参数，也可以是有限任务空间中的任务ID。在课程增强型 *Dec-POMDP*中，目标是通过教师提供的一系列训练任务，提升学生在目标任务上的表现。.

[1]源代码和视频演示可在以下网址找到 https://sites.google.com/view/marl-spc/。

令 τ 表示一条轨迹，其无条件分布$\Pr_{\mu}^{\pi,\phi}(\tau)$（在策略 π 、任务 ϕ及初始状态分布为 $\mu(s_0)$的条件下）为$\Pr_{\mu}^{\pi,\phi}(\tau)=\mu(s_0)\sum_{t=0}^{\infty}\pi(a_t\mid s_t)\ P^{\phi}(s_{t+1}\mid s_t,\ a_t)$。我们用 $p(\phi)$表示目标任务的分布，用 $q(\phi)$表示每个任务采样步骤中训练任务的分布。我们考虑分别由 θ 和 ψ参数化的智能体联合策略 $\pi_\theta(a|s)$和 $q_\psi(\phi)$。课程增强型Dec-POMDP中需要最大化的总体目标为：

$$\begin{aligned} J(\theta,\psi) &= \mathbb{E}_{\phi\sim p(\phi),\tau\sim \Pr_{\mu}^{\pi}}\left[R^{\phi}(\tau)\right] \\ &= \mathbb{E}_{\phi\sim q_\psi(\phi)}\left[\frac{p(\phi)}{q_\psi(\phi)}V\left(\phi,\pi_\theta\right)\right] \end{aligned} \tag{1}$$

其中，$R^{\phi}(\tau)=\sum_t\gamma^t r^{\phi}(s_t,a_t;\ s_0)$ 和 $V(\phi,\pi_\theta)$ 表示在去中心化部分可观测马尔可夫决策过程 $\mathcal{M}(\phi)$ 中 π_θ 的价值函数。然而，在优化 $q_\psi(\phi)$ 时，我们无法得到偏导数 $\nabla_\psi J(\theta,\psi)=\nabla_\psi\sum_\tau q_\psi(\phi)\ R^{\phi}(\tau)\Pr_{\mu}^{\pi,\phi}(\tau)$[2]，因为在大多数多智能体强化学习场景中，奖励函数和关于智能体数量的转移概率函数是非参数、不可微且不连续的。

因此，我们使用不可微方法，即多臂赌博机算法，来优化 $q_\psi(\phi)$，并交替使用强化学习算法（学生）来优化 $\pi_\theta(a|s)$。然而，解决这一问题面临三个关键挑战：（1）由于学生的策略不断变化，教师面临非平稳性问题。（2）学生会遗忘旧任务，需要重新学习这些任务。有些任务可以是其他任务的先修任务，有些任务则彼此独立，可以并行进行。（3）目前缺乏通用的学生框架，以应对不同任务中智能体数量的变化以及稀疏奖励问题。

3.2. 教师作为非平稳上下文老虎机

如前所述，由于学生在学习过程中策略不断变化，教师面临非平稳性问题。具体而言，学生在不同学习阶段学习不同任务时，教师在提供相同任务时会观察到学生表现各异，导致教师的奖励分布随时间变化。此外，学生可能会遗忘先前学到的策略。为缓解这一问题，教师应在利用那些已被发现有助于提升学生在目标任务上表现的任务，与探索那些可能无法直接促进学生学习的任务之间取得平衡。

幸运的是，我们发现非平稳性源于学生，而通过将学生策略嵌入上下文的上下文老虎机可以缓解这一问题——老虎机将学生策略嵌入上下文中。如图1 左图所示，教师将学生的策略表示用作上下文，并从训练任务的分布中选择一个任务。具体而言，我们通过两步在线聚类过程（Zhang 等人，1996年）将上下文纳入 Exp3 算法（Zhang 等人，1996年）。上下文以 x表示，即学生的策略表示。教师动作是一个特定任务，记为 ϕ，教师奖励则是学生在目标任务上的回报。教师算法如算法1所示。在采样阶段（步骤 1-5），教师为学生选择一个训练任务。在训练阶段（步骤 6-7），教师根据从学生处获得的评估奖励调整参数。

算法1教师采样与训练

输入： 上下文 x、簇的数量 N_c、 N_c 个 Exp3 实例，其任务分布为 $w(\phi_k,\ c)$，适用于 $k=1$、.....、K 和 $c=1$、.....、N_c，学习率为 α，以及用于维护历史上下文的缓冲区

输出： $\mathcal{M}(\phi)=\left\{n^{\phi},\boldsymbol{S}^{\phi},\boldsymbol{A}^{\phi},P^{\phi},r^{\phi},O^{\phi},\boldsymbol{\Omega}^{\phi},\ \gamma^{\phi}\right\}$，教师老虎机参数

采样

1. 获取上下文 x，并将其保存到缓冲区 2. 运行在线聚类算法，并获取聚类中心的索引 $c(x)$ 3. 将活动的 Exp3 实例设为索引为$c(x)$的实例 4. 为每个任务 ϕ_k设置概率 $p(\phi_k,c(x))=(1-\alpha)w(\phi_k,c(x))/\sum_{j=1}^{K}w(\phi_k,c(x))+\frac{\alpha}{K}$ 5. 根据分布$p_{\phi_k},\ _c$训练，采样一个新任务 6. 从学生测试 r中获取回报（折扣累积奖励） 7. 通过设置 $w(\phi_k,c(x))=w(\phi_k,c(x))e^{\alpha r/K}$来更新活动的 Exp3 实例

3.2.1. 上下文表示

经过分析，我们发现，必须学习一种有效的表征，将学生的策略作为上下文。一种直接的方法是直接使用学生参数 θ 作为上下文。然而，如果改变学生的架构，参数数量就会过多，无法作为神经网络的输入。因此，我们提出了一种替代方法。

学习良好策略表征的一项原则是预测性表示，这意味着给定状态时，该表征应能准确预测策略动作。遵循这一原则，我们通过监督学习使用模仿函数。监督学习不需要直接获取奖励信号，因此是奖励无关表征学习的一种颇具吸引力的方法。

[2] $p(\phi)$ 不在偏导数中，因为它是一个固定分布。

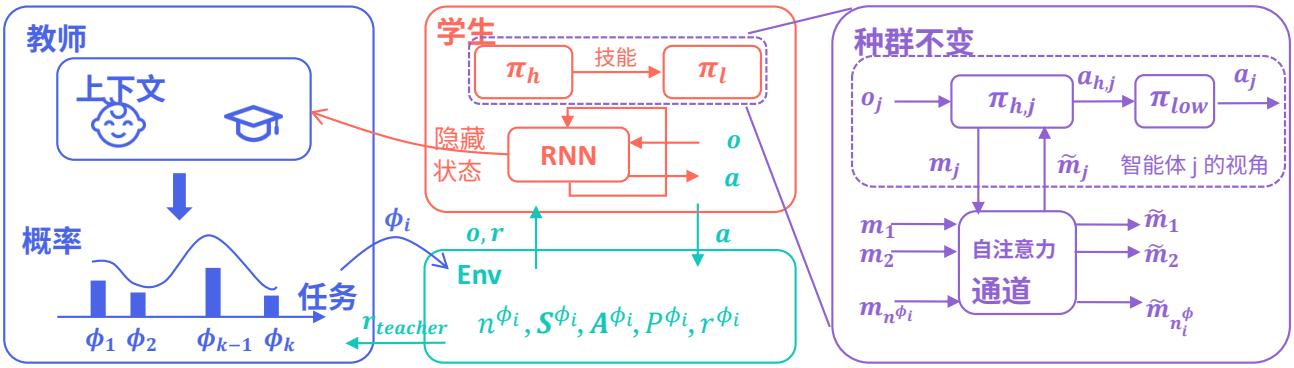


图 *1*。 SPC 的整体框架。它由三个部分组成：可配置环境、教师和学生。 左侧。 教师被建模为上下文多臂老虎机。在每个教师时间步，教师从老虎机动作的分布中选择一个训练任务。 中间。 学生配备了分层技能框架和群体规模不变通信，并使用多智能体强化学习算法在训练任务上进行训练。学生不仅将其循环神经网络模仿模型的隐藏状态作为上下文返回给教师，还会返回测试任务上的平均折扣累积奖励。 右侧。 学生学习分层策略，其中群体规模不变通信在高层进行，并通过自注意力通信通道实现，以处理来自数量不定的智能体的消息。学生中的智能体共享相同的低层策略。

。直观而言，模仿函数试图根据历史行为模仿低层策略。在实践中，我们使用基于循环神经网络的模仿函数f_{im}： $\mathcal{S} \times \mathcal{A} \to [0,1]$。由于循环神经网络在理论上是图灵完备的（Hy¨otyniemi，1996年），其内部状态可作为学生策略的表征。我们使用负交叉熵目标 $\mathbb{E}[\log f_{im}(s,\ a)]$。

3.2.2. 遗憾分析

在本小节中，我们证明所提出的教师算法具有 $\mathbb{E}[R(T)] = O\left(T^{2/3}(LK\log T)^{1/3}\right)$ 的遗憾界，其中 T 是总轮次数， L 是 Lipschitz 常数， K 是臂数（即教师动作的数量）。遗憾分析用于论证在非平稳环境中使用赌博机算法的合理性。遗憾界体现了 SPC 的最优性，因为教师奖励就是学生在目标任务上的回报。

首先，我们介绍任务空间泛化能力的 Lipschitz 假设。

假设 3.2（关于上下文的 Lipschitz 连续性）。不失一般性，将上下文映射到$[0,1]$ 区间，使教师的期望奖励关于上下文满足 Lipschitz 连续性。

$$|r(\phi \mid x) - r(\phi \mid x')| \le L \cdot |x - x'|$$
$$\text{对于任意臂 } \phi \in \Phi \text{ 以及任意一对上下文 } x,\ x' \in \mathcal{X} \tag{2}$$

其中， L 是 Lipschitz 常数， $\mathcal{X}$ 是上下文空间。

这一假设表明，对于在一组任务上训练的任何策略，性能提升的速度不会快于策略变化的速度。这是一个合理的假设，因为我们不能指望学生在新的上下文中仅经过几个训练步骤，就能在某项任务上取得显著进步。我们采用一种适用于有限数量上下文的现有上下文赌博机算法（Auer等人，2002年）（见附录A），并以引理3.3为基础证明定理3.4。

引理 3.3。 算法 *2* 的遗憾界为 $\mathbb{E}[R(T)] = \mathcal{O}(\sqrt{TK|\mathcal{X}|\log K})$。

引理 3.3 表明，若针对每个上下文分别运行 Exp3 的独立副本，遗憾将以平方根的形式依赖于 $|\mathcal{X}|$（Auer 等人，2002年）。这促使我们利用离散化技术来处理庞大的上下文空间。

定理 3.4. 考虑上下文位于 $[0,1]$中的 *Lipschitz* 上下文赌博机问题。算法 *1* 的遗憾为$\mathbb{E}[R(T)] = O\left(T^{2/3}(LK\ln T)^{1/3}\right)$。

证明。 见附录 B。 □

实践中，无法像定理 $[0,1]$ 的证明中那样，使用均匀网格离散化高维上下文空间3.4。为了解决这一问题，我们采用基于层次结构的平衡迭代规约与聚类（BIRCH）在线聚类算法（Zhang 等人，1996年）对上下文空间进行离散化。BIRCH 是一种高效且易于更新的算法，能够有效地对大型数据集进行聚类。在此情况下，它用于对高维的基于 RNN 的策略表示进行聚类。所得簇可视为均匀网格的近似。

3.3. 具备种群不变技能的学生 ls

我们提出了一种种群不变技能框架，以应对智能体数量变化和稀疏奖励问题带来的挑战。该框架允许智能体通过自注意力通道进行通信，使其能够跨不同任务学习可迁移技能。学生模块的设计与算法无关，并且与任何最先进的多智能体强化学习算法正交。尽管已有一些研究尝试解决智能体数量变化的问题（Iqbal 等人，2021年；Hu 等人，2021年），但这些方法严重依赖对环境的先验知识。

种群不变团队协作通信。为了实现种群不变属性并学习智能体之间的战术，我们引入了通信。利用 Transformer 架构处理不同长度输入的能力（Vaswani 等人，2017年），我们将自注意力融入通信机制中。如图 1右侧所示，每个智能体 j 接收一个观测 o_j 并将其编码为消息向量 $m_j = f(o_j)$，随后将其发送至自注意力通道，其中 f是一个观测编码器函数。

通道聚合所有消息，并通过自注意力机制发送新的消息向量 $\tilde{m}_j$。具体而言，给定通道输入 $\mathbf{M} = [m_1, m_2, \cdots, m_n] \in R^{n \times d_m}$，以及通道的可训练权重 $\mathbf{W}^Q$，$\mathbf{W}_K$，$\mathbf{W}_V \in R^{d_m \times d_m}$，我们得到三种不同的表征：$\mathbf{Q} = \mathbf{M}\mathbf{W}^Q$，$\mathbf{K} = \mathbf{M}\mathbf{W}_K$，$\mathbf{V} = \mathbf{M}\mathbf{W}_V$。随后，输出消息为

$$\tilde{\mathbf{M}} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_m}}\right)\mathbf{V} \quad (3)$$

其中 d_m 是消息的维度。由于可训练权重的维度与智能体数量无关，我们的学生模型可以利用种群不变性，有效地学习战术。

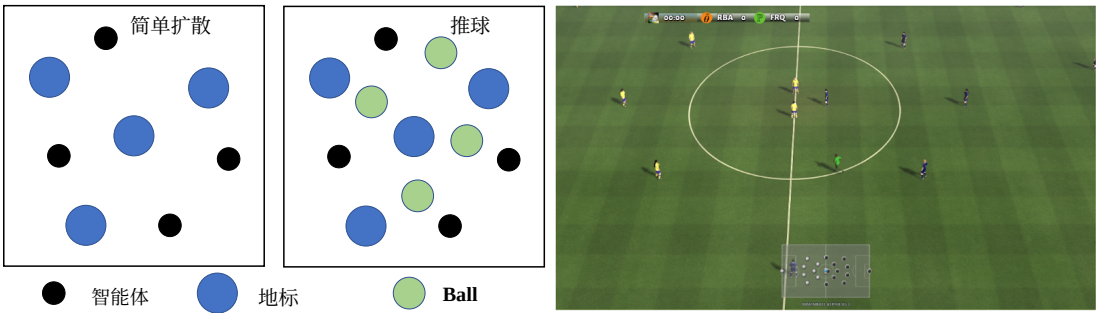
可迁移层次技能。如图中点线框所示，1 右侧，在接收到来自通道的新消息后，$\tilde{m}_j$每个智能体都会采用一个高层动作（技能）$a_{h,\ j} = \pi_{h,\ j}(o_j,\ \tilde{m}_j)$，以执行低层策略$a_j = \pi_{low}(o_j,\ a_{h,\ j})$。在本研究中，我们将高层动作（技能）$a_{h,\ j}$ 推广到连续嵌入空间，使技能既可以是 DIAYN（Eysenbach 等人，2018）中的潜在连续向量，也可以是用于采样离散选项的类别分布（Bacon 等人，2017年）。

实现。我们使用近端策略优化（PPO）（Sc hulman 等人，2017年）在学生中实现高层和低层策略。遵循（Fu 等人，2022）提出的常见做法，每个智能体的高层策略独立学习，而低层策略共享参数，因为不同智能体的基本动作模式应保持一致。低层级智能体由环境提供奖励，而高层策略则训练为以固定间隔采取动作。在此间隔内，累积低层级奖励用作高层级奖励。使用类别分布来实现离散技能时，我们从该分布中采样一个“选项”，并向低层策略提供相应的独热嵌入。

4. 进一步的相关工作

多智能体强化学习中的自动课程学习。课程学习是一种训练策略，它通过根据任务难度组织任务，模拟人类的学习过程（Portelas 等人，2020b）。任务选择被建模为课程马尔可夫决策过程（CMDP）（Narvekar 和 Stone，2018）。自动课程学习机制旨在根据过去的交互学习任务选择函数，例如 ADR（Akkaya 等人，2019年；Mehta 等人，2020年）、ALP-GMM（Portelas 等人，2020a）、SPCL（Jiang 等人，2015）、GoalGAN（Florensa 等人，2018）、PLR（Jiang 等人，2021b；a）、SPDL（Klink 等人，2020年）、CUR-ROT（Klink 等人，2022）和图课程（Svetlik 等人，2017）。近期提出了若干多智能体强化学习课程学习框架，例如开放式进化（Banzhaf 等人，2016；Lehman 等人，2008年；Standish，2003年）、基于种群的训练（Jaderberg 等人，2019年；Liu 等人，2019年）、元学习（Gupta 等人，2021年；Portelas 等人，2020c）以及利用涌现课程进行训练（Baker 等人，2019年；Leibo 等人，2019年；Portelas 等人，2020b）。总之，这些框架遵循一个共同原则：通过自我优化智能体种群之间的选择压力，持续生成性能更优的智能体，从而实现自动课程。

分层多智能体强化学习与通信。分层强化学习（HRL）已得到广泛研究，旨在解决稀疏奖励问题并促进迁移学习。单智能体分层强化学习侧重于学习任务的时间分解，方法包括学习子目标（Nachum 等人，2018b；a；Sukhbaatar 等人，2018；Nair 和 Finn，2019；Wang 等人，2021）或发现可复用技能（Daniel 等人，2012；Gregor 等人，2016；Shankar 和 Gupta，2020；Sharma 等人，2020）。分层多智能体强化学习的最新进展已在第 1 节讨论。在多智能体设置中，通信已被证明能够有效促进智能体之间的合作（Foerster 等人，2016；Das 等人，2019；Sukhbaatar 等人，2016；Singh 等人，2018；Jiang 和 Lu，2018；Kim 等人，2019；Wang 等人，2020a）。然而，目前将分层强化学习扩展到多智能体系统或利用通信的方法，都局限于固定数量的智能体，且缺乏迁移到不同智能体数量设置的能力。



(a) (b)

图2。(a) 多智能体粒子环境。(b) Google Research Football。

5. 实验

为展示我们方法的有效性，我们在两个环境中的若干任务上进行了实验：多智能体粒子环境（MPE）（Lowe 等人，2017年）中的简单扩散和推球任务，以及 Google Research Football（GRF）环境（Kurachet 等人，2019年）中具有挑战性的 5 对 5 任务。我们旨在探究以下研究问题：

Q1：在复杂的大规模多智能体强化学习问题中，课程学习是否必要？（第 5.2节）

Q2：*SPC* 能否优于以往基于课程的多智能体强化学习方法？如果能，*SPC* 的哪些组成部分对性能提升贡献最大？（第 5.3节）

Q3：*SPC* 能否有效地为学生智能体学习课程？（第 5.4 节）

5.1. 环境、基线与指标

环境。 在 GRF 5对5场景中，我们控制四个智能体（不包括守门员），与内置人工智能对手进行比赛。每个智能体接收一种紧凑的编码，其中包含一个 115 维向量，用于概括比赛的各个方面，例如球员坐标、控球权和方向、活跃球员以及比赛模式。单个智能体可用的动作集合包含 19 种离散动作，例如待机、移动、传球、射门、带球等。GRF 提供两种奖励：得分奖励和检查点奖励，以鼓励智能体向前推进球并成功射门。此外，我们还在具有挑战性的 GRF 5对5任务中加入了射门奖励。对于课程，我们从 GRF 中选取了若干基础场景，包括 3对3、传球-射门、3对1 和空门。

在 MPE 中，我们研究了简单扩散和推球（见图 2a）。在简单扩散中，有 n 个智能体需要覆盖所有 n 个地标。智能体会因碰撞而受到惩罚，只有覆盖所有地标后才能获得正奖励。在推球中，有 n 个智能体、 n 个球，以及 n 个地标。智能体必须推动球来覆盖每个地标。覆盖所有地标后即可获得成功奖励。

基线。 我们在表1 中将以下方法作为基线进行比较[3]：

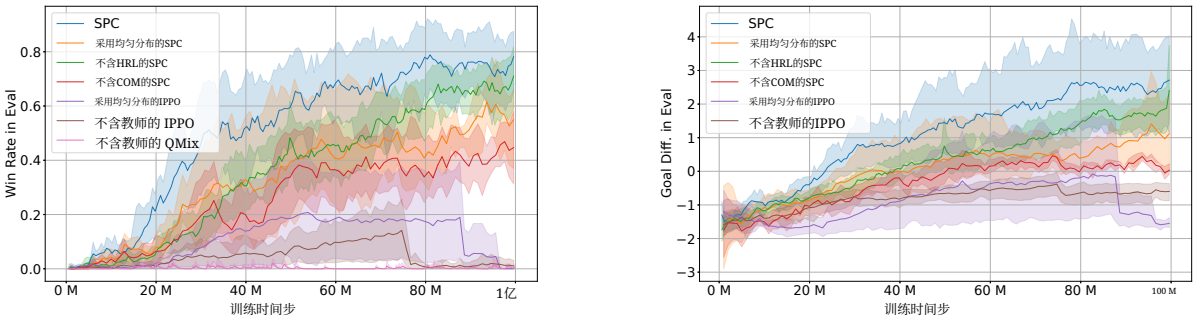
表*1*。 基线算法。

类别	方法
MARL	QMIX（Rashid 等，2018） IPPO（de Witt 等，2020年）
基于课程	采用均匀任务采样的 IPPO VACL (Chen 等，2021年)
消融实验	采用均匀任务采样的 COST 不含分层强化学习和 COM 的 COST

指标。为了评估我们的方法在 GRF 5对5场景中的表现，我们采用了除平均回合奖励之外的其他指标，因为仅凭这一指标可能无法准确反映智能体的表现。具体而言，我们使用胜率和平均净胜球；平均净胜球的计算方式是多智能体强化学习智能体的进球数减去对方球队的进球数。

我们评估多智能体强化学习算法的性能，以证明在复杂的大规模多智能体强化学习问题中采用课程学习的必要性。为确保公平比较，我们移除了MPE任务中VACL的集中式评论家。所有实验均采用独立近端策略优化（IPPO）作为后端多智能体强化学习算法。为确保结果的稳健性，我们在一个30节点集群上开展实验，其中一个节点配备128核CPU和四块A100 GPU。每次实验均使用五个随机种子重复进行，运行时间为1至2天。

[3]我们还运行了 CDS（Li 等人，2021年）和 CMARL（Wu 等人，2021年），但我们未纳入它们的表现，因为 CMARL（Wu 等人，2021年）报告的净胜球相比我们的方法较低。



(a) 胜率　　(b) 净胜球

图*3.* 各种方法在5对5足球比赛中的评估表现。

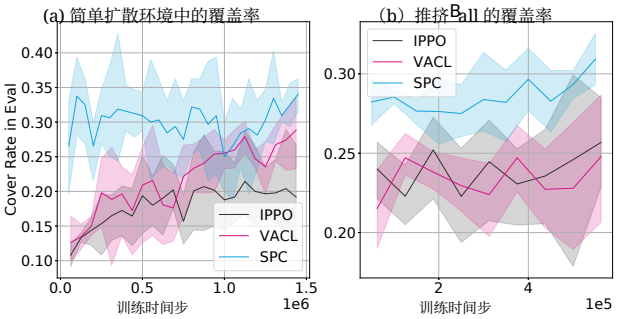


图*4.* 各种方法在MPE上的评估表现。

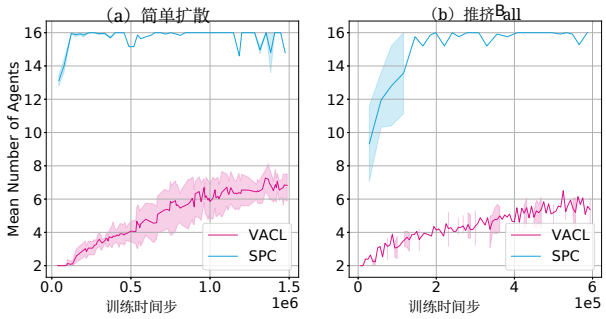


图*5.* MPE上智能体数量的变化。

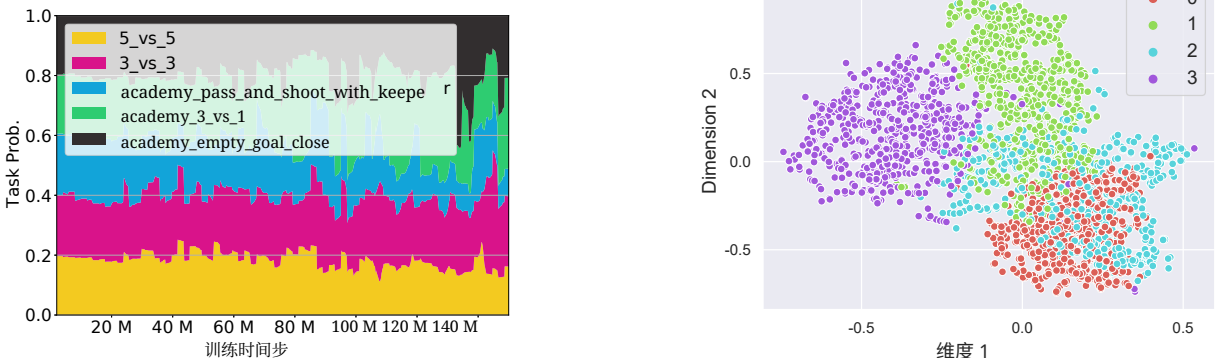
5.2. 课程学习的必要性

我们的实验首先表明，在简单的环境中（例如MPE），学生无需课程即可直接学会完成任务。在MPE实验中，我们随机选择初始状态，并在达到固定的最大步数后结束回合。具体而言，任务空间由 n 个智能体组成，其中 $n \in \{2, 4, 8, 16\}$，最大允许步数设为25。所有评估均在目标任务上进行，$n = 16$。IPPO直接在目标任务上进行训练和评估，图4 中的结果表明，其表现与VACL算法相近。此外，我们观察到SPC方法其覆盖率仅略高于基线方法。此外，我们研究了不同群体规模的概率变化，如图 5所示。我们观察到，SPC 提供的课程正逐渐接近目标任务。这些结果表明，在学生能够直接学会完成任务的简单环境中，课程学习可能并非必需。

然而，对于更复杂的场景，例如 GRF 中的 5对5任务，我们的结果表明课程学习是一种很有前景的解决方案。如图 3a所示，在没有课程学习的情况下，QMix 和 IPPO 无法在 5对5场景中取得良好表现，且 IPPO 略优于 QMix。在图 3b中，我们省略了 QMix 的曲线，因为其平均分较低，会影响图表的呈现。原因可能是 QMix 属于离策略多智能体强化学习算法，因此高度依赖回放缓冲区。然而，在此类稀疏奖励场景中，回放缓冲区中可供 QMix 学习的有效样本要少得多。例如，回放缓冲区中可能包含大量零分样本，导致表现不尽如人意。与此同时，IPPO 凭借其同策略特性，能够获得更高的样本效率，并在此类场景中优于 QMix 等离策略算法。尽管多智能体强化学习方法在 GRF 的基础场景中表现良好，但它们无法解决 5对5任务等复杂场景。因此，课程学习是解决复杂的大规模多智能体强化学习问题的一种很有前景的方案。

5.3. 性能与消融研究

我们的研究表明，SPC在MPE任务中的表现优于VACL。与VACL中通过种群规模变量的连续松弛进行训练不同，我们的多臂老虎机教师在测试时取得了更高的成功率，因为种群规模本质上是一个离散变量。此外，如图5所示，当任务相对简单、无需课程时，SPC提供的课程能够有效探索任务空间并收敛至目标任务。



(a) SPC 在训练期间的任务分布。

(b) 上下文的可视化。

图6. 学习到的课程的可视化。

在GRF实验中，我们没有将VACL纳入GRF的基线方法，因为其实现严重依赖特定场景的先验知识，例如用于划分学习过程的阈值。图6 表明，在5对5比赛中，SPC采用均匀任务采样时的胜率和净胜球均高于I PPO。这些实验表明，当教师根据学生的表现获得奖励时，基于赌博机的教师能够利用学生所处的学习阶段，并提供合适的训练任务。

在消融实验中，我们考察了 SPC 算法两个关键组成部分的影响：上下文多臂老虎机教师，以及学生框架的层次结构。我们分别将前者替换为均匀任务采样，并移除后者。如图 3a和图 3b所示，SPC 的胜率和得分差均高于采用均匀采样的 SPC 以及不含分层强化学习的 SPC。此外，采用均匀任务采样的 SPC 优于采用均匀任务采样的 IPPO。这凸显了分层强化学习在 5对5足球比赛中的重要性，并表明上下文多臂老虎机和层次结构对 SPC 的性能贡献相当。移除分层强化学习和多臂老虎机后，性能相对于 SPC 的下降幅度相近。然而，值得注意的是，采用均匀任务采样的 SPC 的性能方差大于不含分层强化学习的 SPC，这表明均匀采样可能会为学生训练引入更多不合适的任务。总体而言，这些结果进一步证明了 SPC 对复杂的大规模多智能体强化学习问题的必要性[4]。

5.4 学习到的课程的可视化

我们基于选定的一次试验，对训练过程中 SPC 的任务采样分布进行可视化，如图所示。 6a。在训练开始时，任务概率分布看起来近乎均匀，因为教师会探索任务空间并跟踪学生的学习状态，起到防止遗忘的作用。随着训练的推进，概率分布会随时间变化。例如，随着学生逐渐熟练掌握这些场景，3对1和空目标任务的占比会逐渐下降。我们还使用 t-SNE（Van der Maaten 和 Hinton，2008年）对图 6b 中的上下文分布进行可视化；这些上下文被收集并存储在缓冲区中。我们根据索引将上下文分为四类，不同部分代表最终学生策略表示中的不同上下文。

6. 讨论

结论。 我们提出了技能群体课程（SPC），这是一种新颖的多智能体主动课程学习算法，旨在解决多智能体系统中的可扩展性和稀疏奖励问题。SPC 通过引入种群不变的多智能体通信框架，并采用层次化方案让智能体学习技能，从零开始学习复杂行为。此外，SPC 将教师建模为上下文老虎机，并以学生的策略表示来表征上下文，从而缓解非平稳性。尽管我们的设计选择着眼于解决 GRF 5对5任务，但我们认为，分析并解决这些问题对于多智能体主动课程学习算法的进一步发展至关重要。虽然 SPC 包含多个组件，实施起来可能较为复杂，但我们提供了简洁且组织良好的代码，便于使用。

局限性。 我们承认，本算法存在局限性。由于我们的目标是解决困难任务，因此 SPC 对简单任务而言设计过度。此外，了解智能体数量变化对环境动态的影响也很有意思。

[4]我们还展示了 SPC 在 GRF 11对11完整比赛中的表现（见附录 C）。

参考文献

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., 等。《利用机器人手操控魔方》。*arXiv* 预印本 *arXiv:1910.07113*，2019年。Auer, P., Cesa-Bianchi, N., Freund, Y., 和 Schapire, R. E.《非随机多臂老虎机问题》。《*SIAM*计算杂志》，32(1):48–77，2002年。

Bacon, P.-L.、Harb, J. 和 Precup, D. 选项评论家架构。载于*AAAI*人工智能会议论文集，第31卷，2017年。

Baker, B.、Kanitscheider, I.、Markov, T.、Wu, Y.、Powell, G.、McGrew, B. 和 Mordatch, I. 从多智能体自动课程中涌现的工具使用。*arXiv* 预印本 *arXiv:1909.07528*，2019年。

Banzhaf, W.、Baumgaertner, B.、Beslon, G.、Doursat, R.、Foster, J. A.、McMullin, B.、De Melo, V. V.、Miconi, T.、Spector, L.、Stepney, S. 等。定义和模拟开放式新颖性：要求、指南和挑战。《生物科学理论》，135(3)：131–161，2016年。

Bernstein, D. S., Givan, R., Immerman, N., 和 Zilberstein, S. 马尔可夫决策过程去中心化控制的复杂性。运筹学数学，27(4):819–840，2002年。

Chen, J., Zhang, Y., Xu, Y., Ma, H., Yang, H., Song, J., Wang, Y., 和 Wu, Y. 面向稀疏奖励合作多智能体问题的变分自动课程学习。神经信息处理系统进展，34，2021年。

Daniel, C., Neumann, G., 和 Peters, J. 分层相对熵策略搜索。载于人工智能与统计学，第273–281页，2012年。

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., 和 Pineau, J. Tarmac：定向多智能体通信。载于国际机器学习会议，第1538–1546页。PMLR，2019年。

de Witt, C. S., Gupta, T., Makoviichuk,D., Makoviychuk, V., Torr, P. H., Sun, M. 和 Whiteson, S. 在《星际争霸》多智能体挑战赛中，独立学习是否已满足所有需求？*arXiv* 预印本 *arXiv:2011.09533*，2020年。

Eysenbach, B., Gupta, A., Ibarz, J. 和Levine, S. 多样性就是你所需要的一切：无需奖励函数的技能学习。*arXiv* 预印本 *arXiv:1802.06070*，2018。

Florensa, C.、Held, D.、Geng, X. 和 Abbeel, P. 强化学习智能体的自动目标生成。载于国际机器学习会议，第 1515–1528 页。PMLR，2018。

Foerster, J., Assael, I. A., De Freitas, N，和Whiteson, S. 学习通过深度多智能体强化学习进行通信。神经信息处理系统进展，29，2016年。Fu, W., Yu, C., Xu, Z., Yang, J. 和 Wu, Y. 重新审视协作多智能体强化学习中的一些常见实践。*arXiv* 预印本 *arXiv:2206.07505*，2022年。Gregor, K., Rezende, D. J. 和 Wierstra, D. 变分内在控制。*arXiv* 预印本 *arXiv:1611.07507*，2016年。

Gupta, A.、Lanctot, M. 和 Lazaridou, A. 基于动态群体的元学习：用于自然语言多智能体通信。神经信息处理系统进展，34:16899–16912，2021年。

Hazan, E. 和 Megiddo, N. 利用先验知识进行在线学习。载于国际计算学习理论会议，第 499–513 页。Springer，2007 年。

Hochreiter, S. 和 Schmidhuber, J. 长短期记忆。神经计算，9(8):1735–1780，1997 年。

Hu, S.、Zhu, F.、Chang, X. 和 Liang, X. UPDET：通过使用 Transformer 进行策略解耦实现的通用多智能体强化学习。*arXiv* 预印本 *arXiv:2101.08001*，2021年。

Hu, Y.、Wang, W.、Jia, H.、Wang, Y.、Chen, Y.、Hao, J.、Wu, F. 和 Fan, C. 学习利用塑形奖励：一种新的奖励塑形方法。*arXiv* 预印本 *arXiv:2011.02669*，2020年。

Huang, S.、Chen, W.、Zhang, L.、Li, Z.、Zhu, F.、Ye, D.、Chen, T. 和 Zhu, J. Tikick：通过单智能体演示实现多智能体足球完整比赛。*arXiv* 预印本 *arXiv:2110.04507*，2021年。

Hy¨otyniemi, H. 图灵机是循环神经网络。载于*STeP '96*/芬兰人工智能学会出版物，1996年。

Iqbal, S.、De Witt, C. A. S.、Peng, B.、B¨ohmer, W.、Whiteson, S. 和 Sha, F. 面向多智能体强化学习的随机实体级分解。载于 国际机器学习会议，第 4596–4606 页。PMLR，2021年。

Jaderberg, M.、Czarnecki, W. M.、Dunning, I.、Marris, L.、Lever, G.、Castaneda, A. G.、Beattie, C.、Rabinowitz, N. C.、Morcos, A. S.、Ruderman, A. 等。基于种群的强化学习在3D多人游戏中达到人类水平的表现。*Science*，364(6443)：859–865，2019年。

Jiang, J. 和 Lu, Z. 学习用于多智能体合作的注意力通信。神经信息处理系统进展，31，2018。

Jiang, L., Meng, D., Zhao, Q., Shan, S., 和 Hauptmann, A. G. 自定进度课程学习。发表于第二十九届*AAAI*人工智能会议，2015年。

Jiang, M., Dennis, M., Parker-Holder, J., Foerster, J., Grefen- stette, E., 和 Rockt¨aschel, T. 回放引导的对抗性环境设计。神经信息处理系统进展，34:1884–1897，2021a。

Jiang, M., Grefenstette, E., 和 Rockt¨aschel, T. 优先级关卡回放。发表于国际机器学习会议，第4940–4950页。PMLR，2021b。

Kim, D., Moon, S., Hostallero, D., Kang, W. J., Lee, T., Son, K., 和 Yi, Y. 学习在多智能体强化学习中调度通信。*arXiv*预印本 *arXiv:1902.01554*，2019年。

Klink, P., D'Eramo, C., Peters, J. R., 和 Pajarinen, J. 自主节奏深度强化学习。神经信息处理系统进展，33:9216–9227，2020年。

Klink, P., Yang, H., D'Eramo, C., Peters, J., 和 Pajarinen, J. 通过受约束的最优传输实现课程强化学习。载于国际机器学习会议，第 11341–11358 页。PMLR，2022。

Kurach, K., Raichuk, A., Sta´nczyk, P., Zajkac, M., Bachem, O., Espeholt, L., Riquelme, C., Vincent, D., Michalski, M., Bousquet, O., 等。Google Research Football：一种新颖的强化学习环境。*arXiv* 预印本 *arXiv:1907.11180*，2019年。

Lehman, J., Stanley, K. O. 等。通过寻求新颖性利用开放式演化来解决问题。载于*ALIFE*，第 329–336 页。Citeseer，2008年。

Leibo, J. Z.、Hughes, E.、Lanctot, M. 和 Graepel, T. 自动课程与社会互动中创新的涌现：多智能体智能研究宣言。*arXiv* 预印本 *arXiv:1903.00742*，2019年。

Li, C.、Wu, C.、Wang, T.、Yang, J.、Zhao, Q. 和 Zhang, C. 在共享式多智能体强化学习中庆祝多样性。*arXiv* 预印本 *arXiv:2106.02195*，2021年。

Liu, S.、Lever, G.、Merel, J.、Tunyasuvunakool, S.、Heess, N. 和 Graepel, T. 通过竞争实现涌现协调。*arXiv* 预印本 *arXiv:1902.07151*，2019年。

Long, Q., Zhou, Z., Gupta, A., Fang, F., Wu, Y., 和 Wang, X. 面向多智能体强化学习扩展的进化群体课程。*arXiv* 预印本 *arXiv:2003.10423*，2020年。

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., 和 Mordatch, I. 面向混合协作竞争环境的多智能体演员 - 评论家。神经信息处理系统进展，2017年。Mehta, B., Diaz, M., Golemo, F., Pal, C. J., 和 Paull, L. 主动域随机化。载于机器人学习会议，第1162–1176页。PMLR，2020年。Nachum, O., Gu, S., Lee, H., 和 Levine, S. 用于分层强化学习的近最优表征学习。*arXiv* 预印本 *arXiv:1810.01257*，2018a。

Nachum, O., Gu, S. S., Lee, H., andLevine, S. 数据高效的分层强化学习。神经信息处理系统进展，31，2018b。

Nair, S.andFinn, C. 分层前瞻：通过视觉子目标生成对长时域任务进行自监督学习。*arXiv* 预印本 *arXiv:1909.05829*，2019年。

Narvekar, S. and Stone, P. 学习用于强化学习的课程策略。*arXiv* 预印本 *arXiv:1812.00285*，2018。

OpenAI。OpenAI Five。https://openai.com/blog/openai-five/，2019年。访问日期：2019年3月4日。

Pang, Z.-J., Liu, R.-Z., Meng, Z.-Y., Zhang, Y., Yu, Y., and Lu, T. 关于星际争霸完整对局的强化学习。载于*AAAI*人工智能会议论文集，2019年。

Portelas, R.、Colas, C.、Hofmann, K. 和 Oudeyer, P.-Y. 面向连续参数化环境中深度强化学习课程学习的教师算法。载于机器人学习会议，第 835–853 页，2020年a。

Portelas, R.、Colas, C.、Weng, L.、Hofmann, K. 和 Oudeyer, P.-Y. 深度强化学习的自动课程学习：简短综述。*arXiv* 预印本 *arXiv:2003.04664*，2020年b。

Portelas, R.、Romac, C.、Hofmann, K. 和 Oudeyer, P.- Y. 元自动课程学习。*arXiv* 预印本 *arXiv:2011.08463*，2020c。

Rashid, T.、Samvelyan, M.、Schroeder, C.、Farquhar, G.、Foerster, J. 和 Whiteson, S. QMix：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议，第 4295–4304 页，2018。

RoboCup。RoboCup联合会官方网站。https://www.robocup.org/，2019年。访问日期：2019年4月10日。

Schulman, J.、Wolski, F.、Dhariwal, P.、Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。Shankar, T. 和 Gupta, A. 通过时间变分推断学习机器人技能。载于第*37*届国际机器学习大会论文集。JMLR.org，2020年。Sharma, A.、Gu, S.、Levine, S.、Kumar, V. 和 Hausman, K. 动力学感知的无监督技能发现。载于国际学习表征会议，2020年。Singh, A.、Jain, T. 和 Sukhbaatar, S. 学习在大规模多智能体合作与竞争任务中何时进行通信。*arXiv* 预印本 *arXiv:1812.09755*，2018。Standish, R. K. 开放式人工进化。《国际计算智能与应用期刊》，3(02):167–175，2003年。Sukhbaatar, S.、Fergus, R. 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，29，2016年。Sukhbaatar, S.、Denton, E.、Szlam, A. 和 Fergus, R. 通过自我博弈学习目标嵌入，实现分层强化学习。*arXiv* 预印本 *arXiv:1811.09083*，2018。Svetlik, M.、Leonetti, M.、Sinapov, J.、Shah, R.、Walker, N. 和 Stone, P. 为强化学习智能体自动生成课程图。载于*AAAI*人工智能会议论文集，第31卷，2017年。Van der Maaten, L. 和 Hinton, G. 使用 t-SNE 可视化数据。《机器学习研究杂志》，9(11)，2008年。Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A. N.、Kaiser, Ł. 和 Polosukhin, I. 注意力机制就是一切所需。神经信息处理系统进展，30，2017年。Vezhnevets, A. S.、Wu, Y.、Leblond, R. 和 Leibo, J. Z. 选项作为响应：在多智能体强化学习中建立行为层级的基础。*arXiv* 预印本 *arXiv:1906.01470*，2019年。Wang, R.、He, X.、Yu, R.、Qiu, W.、An, B. 和 Rabinovich, Z. 学习高效多智能体通信：一种信息瓶颈方法。载于国际机器学习会议，第9908–9918页。PMLR，2020a。Wang, R.、Yu, R.、An, B. 和 Rabinovich, Z. I2HRL：基于交互影响的分层强化学习。载于第二十九届国际人工智能联合会议论文集，第3131–3138页，2021年。Wang, T., Gupta, T., Mahajan, A., Peng, B., Whiteson, S., and Zhang, C. Rode：学习角色以分解多智能体任务。*arXiv*预印本 *arXiv:2010.01523*，2020b。Wang, W., Yang, T., Liu, Y., Hao, J., Hao, X., Hu, Y., Chen, Y., Fan, C., and Gao, Y. 从少到多：大规模动态多智能体课程学习。载于*AAAI*人工智能会议论文集，第34卷，第7293–7300页，2020c。Wu, S., Wang, T., Li, C., and Zhang, C. 容器化分布式基于价值的多智能体强化学习。*arXiv*预印本 *arXiv:2110.08169*，2021。Yang, J., Borovikov, I., and Zha, H. 结合技能发现的分层协作多智能体强化学习。*arXiv*预印本 *arXiv:1912.03558*，2019。Yang, Y. and Wang, J. 从博弈论视角概述多智能体强化学习。*arXiv*预印本 *arXiv:2011.00583*，2020。Zhang, K., Yang, Z., and Bas¸ar, T. 多智能体强化学习：理论与算法精选综述。《强化学习与控制手册》，第321–384页，2021。Zhang, T., Ramakrishnan, R., and Livny, M. Birch：一种面向超大数据库的高效数据聚类方法。*ACM Aigmod Record*，25(2):103–114，1996年。

A. 上下文数量有限时的上下文老虎机

算法2适用于少量上下文的上下文赌博机算法

1: **初始化：**对于每个上下文 x，创建一个算法实例 Exp3$_x$，算法为
Exp32: **对于**每一轮，**执行**3: 调用算法 Exp3$_x$，并传入 $x = x_t$4: 执行 Exp3$_x$
所选的动作5: 将奖励 r_t 返回给 Exp3$_x$6: **结束循环**

B. 定理 3.4 的证明

定理 3.4. 考虑上下文取值于 $[0,1]$的 *Lipschitz* 上下文老虎机问题。算法 *1* 的遗憾为 $\mathbb{E}[R(T)] = O\left(T^{2/3}(LK\ln T)^{1/3}\right)$。

证明。令 S_m 为 $[0,1]$, 上的 ϵ-均匀网格，即 $[0,1]$ 中所有坐标为 ϵ的整数倍的点构成的集合。取$\epsilon = 1/(d-1)$，其中整数 d 表示 S_m中的点数，后续分析中将对其进行调整。
我们将算法 2 应用于上下文空间 S_m。令$f_{S_m}(x)$ 为一个映射，将上下文 x 映射到 S_m 中距离最近的点：

$$f_{S_m}(x) = \min\left(\operatorname*{argmin}_{x' \in S_m} |x - x'|\right)$$

在每一轮 t，我们用 $f_{S_m}(x_t)$ 替换上下文 x_t，并调用 Exp3$_S$。遗憾界由两部分构成：Exp3$_S$ 的遗憾界，以及离散化误差（其适当定义）。形式化地，对每个上下文 $x \in \mathcal{X}$，定义"离散化最佳响应" $\pi^*_{S_m}$：$\mathcal{X} \to \Phi$: $\pi^*_{S_m}(x) = \pi^*(f_{S_m}(x))$。

我们将算法 Alg 的总奖励定义为 Reward(Alg) $= \sum_{t=1}^{T} r_t$。然后，Exp3$_S$ 的遗憾和离散化误差定义如下：

$$R_S(T) = \text{Reward}\left(\pi^*_S\right) - \text{Reward}\left(\text{Exp3}_S\right)$$
$$\text{DE}(S) = \text{Reward}\left(\pi^*\right) - \text{Reward}\left(\pi^*_S\right).$$

由此可知，遗憾等于 $R(T) = R_S(T) + \text{DE}(S)$。根据引理 $\mathbb{E}\left[R_S(T)\right] = \mathcal{O}(\sqrt{TK\log K})$ 可得 3.3，因此只需对离散化误差给出上界，并调整离散化步长 ϵ。
对于每一轮 t 以及相应的上下文 $x = x_t$, $r(\pi^*_S(x) \mid f_S(x)) \geq r(\pi^*(x) \mid f_S(x)) \geq r(\pi^*(x) \mid x) - \epsilon L$。第一个不等式由 π^*_S 的最优性决定，第二个不等式则由利普希茨性决定。将此结果对所有轮次 t求和，可得 $\mathbb{E}$ [奖励（π^*_S）] ≥ 奖励 $[\pi^*] - \epsilon LT$。

因此，遗憾为

$$\mathbb{E}[R(T)] \leq \epsilon LT + O\left(\sqrt{\frac{1}{\epsilon}TK\log T}\right) = O\left(T^{2/3}(LK\log T)^{1/3}\right) \tag{4}$$

对于最后一个不等式，我们选择 $\epsilon = \left(\frac{K\,\text{对数}\,T}{TL^2}\right)^{1/3}$。 □

C. GRF 11对11完整比赛中的 SPC

我们还在稀疏奖励的 GRF 11对11完整比赛场景中进行了实验。如图 7所示，SPC 在经过 2 亿个时间步的训练后，在目标任务中对阵内置人工智能的胜率约为 50%。这并非易事，因为这是多智能体强化学习领域最具挑战性的基准之一，而且如果没有手工设计的工程方案，大多数当前的多智能体强化学习方法都难以取得进展。

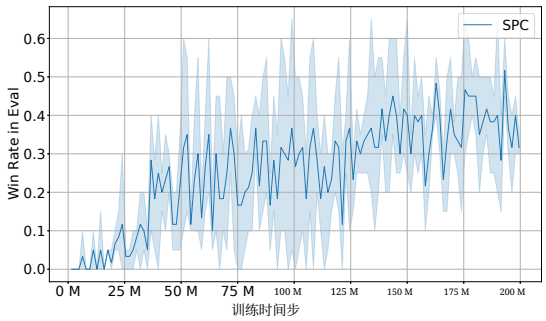


图7. SPC在11对11场景中的表现。

D. 低层技能的定性分析

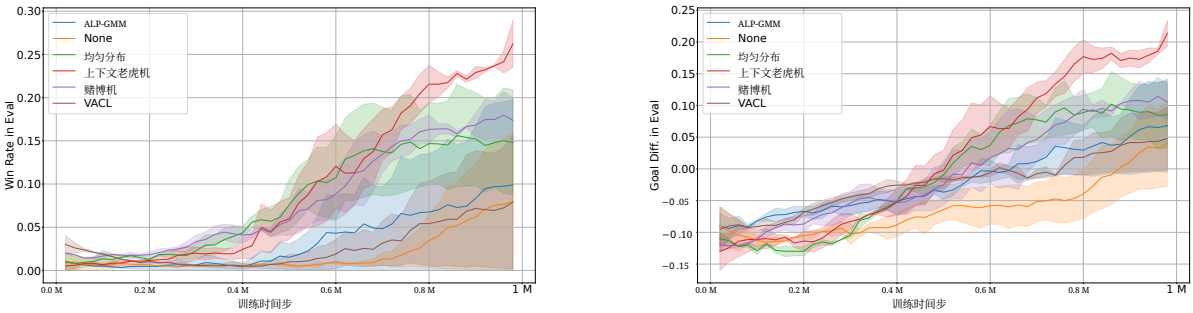
我们展示了不同高层动作下的比赛统计数据。例如，在 GRF 中，每场比赛的射门、传球和跑动动作次数。这些不同的低层策略由高层动作诱导产生。我们通过固定一个智能体的高层动作，同时让其他智能体继续使用 SPC，来评估这些统计数据。表 2 中的结果是在 5对5场景下进行五次运行后取平均得到的。

表2. 低层技能统计。

	每场比赛射门次数	每场比赛传球次数	每场比赛跑动次数
技能1	7.9倍	0.5倍	2254个时间步
技能2	2.3倍	26.4 倍	2149 个时间步
技能 3	1.6 倍	3.9 倍	2875 个时间步

E. 在 GRF corner-5 场景中比较不同教师算法

为进一步说明 SPC教师模块的有效性，我们在 GRF 的 corner-5场景上进行了实验，其中目标任务是控制 11 名球员中的 5 名在 GRF Corner场景中取得进球。实验旨在确定 SPC 中的上下文老虎机在安排训练中的智能体数量方面是否优于其他课程学习方法。我们将 SPC教师与非课程训练（无）、均匀任务采样（均匀分布）、一种最先进的课程学习方法（ALP-GMM）以及一种多智能体课程学习方法（VACL）进行比较。训练任务空间由 n 个智能体组成，其中 $n \in \{1, 3, 5\}$。所有教师都采用相同的基础



(a) 胜率　　(b) 净胜球

图8. 不同教师算法在 GRF corner-5 场景中的评估表现。

架构，不采用 Transformer 架构和分层强化学习。我们还研究了基于 RNN 的上下文的消融效果（参见上下文老虎机和赌博机）。图 8 展示了在一百万个时间步的训练后，SPC 上下文赌博机相较于其他自动课程学习方法的优势。

F. 实现细节

我们使用 Ray RLlib 中近端策略优化（PPO）的默认实现，该实现通过多个工作进程进行经验收集，从而实现扩展。这使我们能够在训练期间利用并行工作进程生成大量轨迹采样，以减轻高方差并促进探索。我们使用分布式工作进程并行进行多次轨迹采样，并为每个智能体采用参数共享。训练器会在工作进程完成同步采样后，向其广播新权重。

F.1. Google Research Football

我们为训练 GRF 5对5场景设置了五项任务，分别是 5对5、3对3、传球-射门、3对1 和空门。在空门任务中，一个智能体需要向前移动，并对着空门射门。在传球-射门和 3对3任务中，由两个智能体对阵一名守门员和三名球员，初始位置各不相同。在 3对1任务中，由三个智能体对阵一名中后卫和一名守门员。在 5对5任务中，由四个智能体对阵五名球员。一般而言，我们将所有球员初始化在固定位置，并设定为中前卫角色。

高层策略同时采用多层感知机和自注意力机制，低层策略则采用多层感知机。对于高层策略，输入首先经过两个各含 256 个单元、使用 ReLU 激活函数的隐藏层，投影为嵌入；随后输入多头自注意力（8 个头，每个头 64 个单元）。接着，输出经过另一个含 256 个单元的全连接层，投影为动作和值。对于低层策略，我们采用含两个各有 256 个单元的隐藏层的多层感知机，即 RLlib 中策略网络的默认配置。

表 *3.* SPC 超参数。

（a）GRF 中使用的 SPC 超参数。

Name	值
折扣率	0.99
GAE参数	1.0
KL 系数	0.2
采样片段长度	1000
训练批次大小	100000
SGD 小批量大小	10000
SGD 迭代次数	60
学习率	1e-4
熵系数	0.0
裁剪参数	0.3
价值函数裁剪参数	10.0

(b) MPE 中使用的 SPC 超参数。

Name	值
折扣率	0.99
GAE参数	1.0
KL 系数	0.5
SGD 迭代次数	10
学习率	1e-4
熵系数	0.0
裁剪参数	0.3
价值函数裁剪参数	10.0

F.2. MPE

在 MPE任务中，智能体必须通过物理动作相互协作，抵达一组地标。智能体可以观测其他智能体和地标的相对位置，并根据任一智能体与各个地标的距离获得集体奖励。换言之，智能体必须覆盖所有地标。此外，智能体彼此碰撞时会受到惩罚。智能体需要推断要覆盖的地标并前往该处，同时避免与其他智能体碰撞。

MPE 中 SPC 的超参数如表 3b 所示。在 MPE 中，采样片段长度、训练批次大小和 SGD 小批量大小等超参数会根据场景的时域进行调整，以便在回合结束后更新策略。我们采用与 GRF 相同的神经网络架构，但所有多层感知机隐藏层均包含 128 个单元。其他未列出的超参数均采用 RLlib PPO实现中的默认配置。