arXiv:2305.03735v1 [cs.AI] 4 May 2023

用于在竞争性自课程中学习涌现行为的斯塔克尔伯格博弈

Boling Yang[1], **Liyuan zheng**[1],**Lillian J. Ratliff**[1],**Byron Boots**[1],**Joshua R. Smith**[1] 华盛顿大学

摘要： 自课程训练是多智能体强化学习（多智能体强化学习）的一个重要子领域。在机器人学领域，自课程已针对具有物理基础的问题开展了实验，例如鲁棒控制和交互式操纵任务。然而，这些任务的非对称性使得复杂策略的生成颇具挑战。事实上，环境中的非对称性可能会隐式或显式地为一部分智能体提供优势，进而导致低质量均衡。本文提出了一种新颖的博弈论多智能体强化学习算法——斯塔克尔伯格多智能体深度确定性策略梯度（ST-MADDPG）。该算法将双人多智能体强化学习问题构造为斯塔克尔伯格博弈，在层级交互结构中，一名参与者作为“领导者”，另一名作为“跟随者”，其中领导者具有优势。在三个非对称竞争性机器人环境中，我们利用 ST-MADDPG 所赋予领导者的优势，提高自课程训练的质量，并生成更复杂、更先进的自主智能体。

关键词： 多智能体强化学习、Stackelberg 博弈、对抗学习、涌现复杂性

1 引言

多智能体强化学习（MARL）研究多个自主智能体在共同环境中相互交互时所面临的序贯决策问题，每个智能体都旨在优化自身的长期回报 [1]。纯竞争场景构成了 MARL 中一类重要的子问题，通常采用竞争性马尔可夫决策过程框架将其建模为零和双人博弈 [2]。利用竞争性 MARL 方法解决这类问题已取得许多成功，尤其是在对称博弈中，包括有限动作空间上的扩展式博弈，例如国际象棋和电子游戏 [3, 4, 5]。这些算法通常采用协同进化训练方案，让竞争的智能体不断为彼此创造新任务，并通过解决这些新任务逐步改进自身策略。然而，一旦一个或多个进化出的智能体无法对对手构成足够挑战，由于缺乏适应压力，后续训练便不太可能取得进一步进展。协同进化过程的停滞表明智能体已达到均衡。

近年来，竞争性多智能体强化学习方法受到了机器人学界的关注，并已用于解决具有物理基础的问题，例如用于鲁棒控制的对抗学习、自主任务生成和复杂机器人行为学习 [6, 7, 8]。然而，这些问题在实践中通常具有非对称性。与所有智能体拥有相同知识和相同行动能力的对称博弈不同，非对称博弈要求智能体在不平衡的竞争环境中相互耦合，同时各自完成任务。其中一个智能体可能因初始任务更容易而获得优势，并学会利用这一优势迅速主宰比赛 [9]。这会使共同进化过程过早终止，所有智能体都陷入低质量的均衡。例如，在模拟拳击游戏中，如果一名选手的出拳力量明显强于另一名选手，就能轻易将对手击倒。这名选手可能学会在比赛一开始就击倒对手，使对手没有机会探索更好的反制策略，例如通过策略性步法躲避致命一击。

为克服这一挑战，一种常见做法是直接采用基于种群的方法和分布式采样 [10, 7]，生成大量多样化样本。然而，这种方法

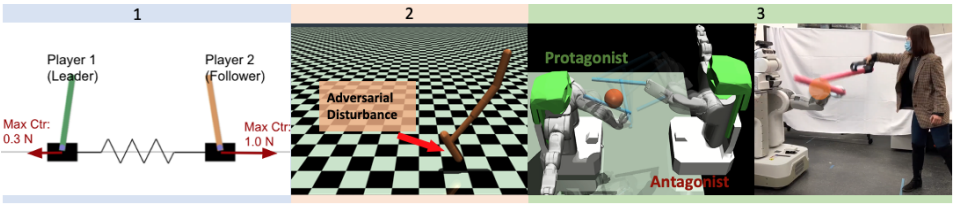


图 1：本工作聚焦于三个具有物理交互的竞争机器人任务。1. **竞争式倒立摆** 是一个简单的一维连续控制环境。2.**带有对抗性扰动的跳跃者** 是一个经典的鲁棒控制问题。3. **击剑博弈** 是一个竞争性人机交互基准。在各种环境中机器人行为的视频演示可在项目网站找到。

只能在一定程度上缓解这一问题。投入足够的工程工作后，可以使用奖励塑形和模仿学习等策略初始化方法，将系统初始化到所需状态 [11, 8]。极小极大遗憾策略是一种风险中性的决策者策略，已被证明能够在对抗学习中防止较强玩家主宰游戏 [6]。更重要的是，这些方法仅仅通过对称的信息结构和同步学习动力学平等对待两名玩家，未能捕捉环境固有的不平衡底层结构。此外，正如 [12, 13, 14], 所述，采用同步梯度下降上升更新的多智能体强化学习方法在实践中可能导致较差的收敛特性。

我们旨在通过直接调整博弈动力学来解决这一问题，从而帮助重新平衡非对称环境中的偏差。近年来，在标准型博弈设置中，针对多智能体学习中的非对称梯度算法已开展了大量研究 [15, 16]。本文利用斯塔克尔伯格博弈结构 [17] 引入层级博弈顺序，从而将一种非对称交互结构引入竞争性马尔可夫博弈 [18]。
在双人斯塔克尔伯格博弈中，领导者知道跟随者会对其公布的策略作出反应。基于这种结构，领导者在优化自身目标函数时会考虑跟随者预期的响应，而跟随者则针对领导者的动作选择短视最佳响应，以优化自身的目标函数。因此，与常规竞争博弈中的均衡收益相比，领导者可以利用斯塔克尔伯格博弈结构获得更高的均衡收益 [19]。当任务中某个智能体是主要智能体（例如，存在对抗者时的鲁棒控制），或者由于非对称博弈环境的影响，某个智能体处于初始或固有劣势，因而需要重新平衡双方的力量时，这一特性尤为理想，正如我们将在实验中展示的那样。本文的主要**贡献**如下：

一种新颖的多智能体强化学习算法：ST-MADDPG。我们将双智能体竞争性多智能体强化学习问题表述为斯塔克尔伯格博弈。通过采用全导数斯塔克尔伯格学习更新规则，我们将当前最先进的多智能体强化学习算法 MADDPG [20] 扩展为一种新颖的斯塔克尔伯格版本，称为斯塔克尔伯格 MADDPG（ST-MADDPG）。

在非对称环境中实现更优的竞争性学习。我们从一个新颖的视角出发，研究信息和施力不对称如何影响多智能体协同进化过程中智能体的表现与行为。首先，我们设计了一个具有连续控制空间的简单竞争性强化学习基准：竞争式倒立摆（图1.1）。在该环境中，我们证明，采用斯塔克尔伯格梯度更新能够为领导者智能体带来优势，从而弥补智能体初始或固有的劣势，并带来更优的表现。

在实际机器人任务中的应用。通过两项任务展示了所提算法的实际效果。在鲁棒控制问题（图 1.2）中，在对抗训练期间赋予领导者优势，使最终的机器人能够更好地抵御对抗性且强烈的随机扰动。在竞技击剑游戏（图 1.3）中，ST-MADDPG 使部分进攻型智能体学会了复杂策略，从而获得更好的表现。值得注意的是，表现最佳的两名攻击者学会了诱使对手移动到较难操控的关节构型，从而暂时限制对手的机动性。

2 预备知识

本节将介绍所需的预备数学模型和符号记法。

双人竞争性马尔可夫博弈。 我们考虑一个双人零和、完全可观测的竞争性马尔可夫博弈（即竞争性 MDP）。竞争性马尔可夫博弈是一个由 $(\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, P, r)$ 组成的元组，其中 $\mathcal{S}$ 是状态空间，$s \in \mathcal{S}$ 是一个状态，玩家 $i \in \{1, 2\}$, A^i 是玩家 i 的动作空间，且满足 $a^i \in \mathcal{A}^i$。P: $\mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \to \mathcal{S}$ 是一个转移核，其中 $P(s'|s, a^1, a^2)$ 表示在状态为 s 时，智能体在 s 中同时采取动作 (a^1, a^2) 后转移到状态 s' 的概率。奖励 r: $\mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \to \mathbb{R}$ 是玩家 1 的奖励函数。由于竞争性设置具有零和性质，玩家 2 获得 r 的相反数作为自身的奖励反馈。每个智能体都使用由 θ^i 参数化的随机策略 $\pi^i_{\theta^i}$。

轨迹 $\tau = (s_0, a^1_0, a^2_0, \ldots\ldots, s_T, a^1_T, a^2_T)$ 给出了累积奖励或回报，其定义为 $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a^1_t, a^2_t)$，其中折扣因子 $0 < \gamma \leq 1$ 为不同时间步收到的奖励赋予权重。在状态 s_t 中执行联合动作配置 (a^1_t, a^2_t) 后，$\pi = \{\pi^1, \pi^2\}$ 的期望回报可由以下 Q^π 函数表示:

$$Q^\pi(s_t, a^1_t, a^2_t) = \mathbb{E}_{\tau\sim\pi}\big[\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a^1_{t'}, a^2_{t'}) | s_t, a^1_t, a^2_t\big],$$

其中，$\tau \sim \pi$ 是简写，表示轨迹上的分布取决于 π：$s_0 \sim \rho$，$a^1_t \sim \pi^1(\cdot|s_t)$ $a^2_t \sim \pi^2(\cdot|s_t)$ $s_{t+1} \sim P(\cdot|s_t, a^1_t, a^2_t)$。$\rho$ 是系统的初始状态分布。

博弈目标是期望回报，定义为

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\big[\sum_{t=0}^{T} \gamma^t r(s_t, a^1_t, a^2_t)\big] = \mathbb{E}_{s\sim\rho, a^1\sim\pi^1(\cdot|s), a^2\sim\pi^2(\cdot|s)}\big[Q^\pi(s, a^1, a^2)\big].$$

在竞争性马尔可夫博弈中，玩家 1 旨在寻找最大化博弈目标的策略，而玩家 2 旨在将其最小化。也就是说，他们分别求解 $\max_{\theta_1} J(\pi^1, \pi^2)$ 和 $\min_{\theta_2} J(\pi^1, \pi^2)$。

斯塔克尔伯格博弈预备知识。 斯塔克尔伯格博弈是两个智能体之间的博弈，其中一个智能体被视为领导者，另一个则是跟随者。每个智能体都有一个希望优化的目标函数，该目标函数不仅取决于自身的动作，也取决于另一个智能体的动作。具体而言，领导者假设跟随者将采取最佳响应，并据此优化自己的目标函数。设 $J_1(\theta_1, \theta_2)$ 和 $J_2(\theta_1, \theta_2)$ 分别为领导者和跟随者希望最小化的目标函数（在竞争场景下 $J_2 = -J_1$），其中 $\theta_1 \in \Theta_1 \subseteq \mathbb{R}^{d_1}$ 和 $\theta_2 \in \Theta_2 \subseteq \mathbb{R}^{d_2}$ 是各自的决策变量或策略，而 $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ 是它们的联合策略。领导者和跟随者旨在求解以下问题:

$$\max_{\theta_1\in\Theta_1} \{J_1(\theta_1, \theta_2) \,\big|\, \theta_2 \in \arg\max_{\theta_2\in\Theta_2} J_2(\theta_1, \theta_2)\}, \tag{L}$$

$$\max_{\theta_2\in\Theta_2} J_2(\theta_1, \theta_2). \tag{F}$$

由于领导者假设跟随者会选择最佳响应 $\theta^*_2(\theta_1) = \arg\max_\phi J_2(\theta_1, \theta_2)$，因此，跟随者的决策变量隐式地是领导者决策变量的函数。在推导优化问题 (L) 的充分条件时，领导者利用这一信息来计算其成本的全导数:

$$\nabla J_1(\theta_1, \theta^*_2(\theta_1)) = \nabla_{\theta_1} J_1(\theta) + (\nabla\theta^*_2(\theta_1))^\top \nabla_{\theta_2} J_1(\theta),$$

其中 $\nabla\theta^*_2(\theta_1) = -(\nabla^2_{\theta_2} J_2(\theta))^{-1} \nabla_{\theta_2\theta_1} J_2(\theta)$[1] 根据隐函数定理 [21]。

若 $\theta = (\theta_1, \theta_2)$ 满足 $\nabla J_1(\theta_1, \theta^*_2(\theta_1)) = 0$ 且 $\nabla^2 J_1(\theta_1, \theta^*_2(\theta_1)) > 0$，则它是 (L) 的一个局部解。对于跟随者的问题，最优性的充分条件是 $\nabla_{\theta_2} J_2(\theta_1, \theta_2) = 0$ 且 $\nabla^2_{\theta_2} J_2(\theta_1, \theta_2) > 0$。由此得到以下均衡概念，它刻画了局部 Stackelberg 均衡的充分条件。

*<escape>***定义 1（微分斯塔克尔伯格均衡，Fiez 等人 16）** 联合策略剖面 $\theta^* = (\theta^*_1, \theta^*_2) \in \Theta_1 \times \Theta_2$ 是微分斯塔克尔伯格均衡，当且仅当 $\nabla J_1(\theta^*) = 0$，$\nabla_{\theta_2} J_2(\theta^*) = 0$，
$\nabla^2 J_1\ \theta^* > 0$ $\nabla^2_{\theta_2} J_2\ \theta^* > 0$ ()，且 ()。*</escape>*

[1] $J(\theta_1, \theta_2)$ 关于 θ_i 的偏导数记为 $\nabla_{\theta_i} J(\theta_1, \theta_2)$，而对于某个函数 h，$J(\theta_1, h(\theta_1))$ 的全导数记为 ∇J，其中 $\nabla J(\theta_1, h(\theta_1)) = \nabla_{\theta_1} J(\theta_1, h(\theta_1)) + (\nabla h(\theta_1))^\top \nabla_{\theta_2} J(\theta_1, h(\theta_1))$。

Stackelberg 学习动态源自一阶梯度充分条件，表示为 $\theta_{1,k+1} = \theta_{1,k} - \alpha_1 \nabla J_1(\theta_{1,k}, \theta_{2,k})$，以及 $\theta_{2,k+1} = \theta_{2,k} - \alpha_2 \nabla_{\theta_2} J_2(\theta_{1,k}, \theta_{2,k})$，其中 $\alpha_i, i = 1,2$ 分别是领导者和跟随者的学习率。

MADDPG。Lowe 等人 [20]表明，在简单的多智能体连续控制任务中，朴素的策略梯度方法表现不佳，并提出了一种更先进的多智能体强化学习算法，称为 MADDPG，它是最先进的多智能体控制算法之一。MADDPG 的核心思想是采用集中式训练与分散式执行框架。具体而言，他们使用集中式评论家网络 Q_w 来近似 Q^π 函数，并使用全局评论家更新每个智能体的策略网络 π_θ^i。考虑确定性策略设定，其中每位玩家的策略为 μ_{θ_i}，参数为 θ_i。[2] （玩家 1）的博弈目标为 $J(\theta_1, \theta_2) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[Q_w(s, \mu_{\theta_1}(s)\ \mu_{\theta_2}(s))\right]$，其中 $\xi = (s, a^1, a^2, r, s')$，$\mathcal{D}$ 是经验回放缓冲区。每位玩家的策略梯度可计算为 $\nabla_{\theta_1} J(\theta_1, \theta_2) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[\nabla_{\theta_1} \mu_{\theta_1}(s) \nabla_{a^1} Q_w(s, a^1, a^2)|_{a^1 = \mu_{\theta_1}(s)}\right]$，以及 $\nabla_{\theta_2} J(\theta_1, \theta_2) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[\nabla_{\theta_2} \mu_{\theta_2}(s) \nabla_{a^2} Q_w(s, a^1, a^2)|_{a^2 = \mu_{\theta_2}(s)}\right]$。

评论家的目标函数定义为均方贝尔曼误差 $L(w) = \mathbb{E}_{\xi \sim \mathcal{D}}[(Q_w(s,\ a^1,\ a^2) - (r + \gamma Q_{w'}(s',\ \mu_{\theta_1'}(s')\mu_{\theta_2'}(s')))^2]$，其中 $Q_{w'}$ 和 $\mu_{\theta_1'},\ \mu_{\theta_2'}$ 是通过在训练过程中对 Q_w 和 $\mu_{\theta_1},\ \mu_{\theta_2}$ 网络参数进行 Polyak 平均得到的目标网络。

在竞争性场景中使用 MADDPG 时，集中式评论家通过梯度下降进行更新，而两个智能体的策略则通过同时梯度上升与下降进行更新 $\theta_1 \leftarrow \theta_1 + \alpha^1 \nabla_{\theta_1} J(\theta_1, \theta_2), \theta_2 \leftarrow \theta_2 - \alpha^2 \nabla_{\theta_2} J(\theta_1, \theta_2)$。

3 Stackelberg MADDPG算法

在本节中，我们介绍新颖的 ST-MADDPG 算法。ST-MADDPG 的一个核心特点是，领导者智能体在推导基于梯度的更新时，会利用跟随者将对其动作作出响应这一知识。也就是说，全导数学习更新通过在学习过程中预判跟随者的更新，使领导者占据优势，并促使其在生成对抗网络和演员-评论家网络等广泛应用中收敛至 Stackelberg 均衡 [16, 14]。根据 Bas.ar 和 Olsder [19, 第 4], 章，在跟随者具有唯一最佳响应的双人博弈中，领导者在 Stackelberg 均衡下的收益优于纳什均衡，这在许多应用中是人们所期望的。完整的 ST-MADDPG 算法见附录1 中的算法A.1。

将玩家 1 设为领导者后，两个玩家的 ST-MADDPG 策略梯度更新规则如下：

$$\theta_1 \leftarrow \theta_1 + \alpha^1 \nabla J(\theta_1, \theta_2),$$
$$\theta_2 \leftarrow \theta_2 - \alpha^2 \nabla_{\theta_2} J(\theta_1, \theta_2),$$

其中，领导者更新中的全导数为

$$\nabla J(\theta_1, \theta_2) = \nabla_{\theta_1} J(\theta_1, \theta_2) - \nabla_{\theta_1 \theta_2} J(\theta_1, \theta_2)(\nabla^2_{\theta_2} J(\theta_1, \theta_2))^{-1} \nabla_{\theta_2} J(\theta_1, \theta_2). \tag{1}$$

(1) 中全导数的二阶项可通过应用链式法则计算：

$$\nabla_{\theta_1 \theta_2} J(\theta_1, \theta_2) = \mathbb{E}_{\xi \sim \mathcal{D}}[\nabla_{\theta_1} \mu_{\theta_1}(s) \nabla_{a^1 a^2} Q_w(s, a^1, a^2)(\nabla_{\theta_2} \mu_{\theta_2}(s))^T|_{a^1 = \mu_{\theta_1}(s), a^2 = \mu_{\theta_2}(s)}],$$
$$\nabla^2_{\theta_2} J(\theta_1, \theta_2) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[\nabla^2_{\theta_2} \mu_{\theta_2}(s) \nabla_{a^2} Q_w(s, a^1, a^2)|_{a^2 = \mu_{\theta_2}(s)}\right].$$

为获得全导数的估计值 $\nabla J(\theta_1,\ \theta_2)$，通过从经验回放缓冲区中采样来计算（1）中的各项。逆海森向量积可通过共轭梯度高效计算 [14]。

隐式映射正则化。 Stackelberg 梯度动力学中的全导数需要计算跟随者海森矩阵 $\nabla^2_{\theta_2} J(\theta_1,\ \theta_2)$ 的逆矩阵。由于实际强化学习问题中的策略网络可能具有高度非凸性，因此 $(\nabla^2_{\theta_2} J(\theta_1,\ \theta_2))^{-1}$ 可能存在病态条件。

[2]遵循原始 DDPG 算法中的设置和符号约定，我们使用 μ 表示确定性策略，以区别于随机策略。 [20],

因此，在实践中，我们不直接计算这一项，而是计算其正则化变体，形式为 $(\nabla^2_{\theta_2} J(\theta_1, \theta_2) + \lambda I)^{-1}$。这种正则化方法可以解释为：领导者认为跟随者正在优化一个正则化代价函数 $J(\theta_1, \theta_2) + \frac{\lambda}{2}\|\theta_2\|^2$，而跟随者实际上优化的是 $J(\theta_1, \theta_2)$。正则化项 λ 在领导者的 Stackelberg 梯度更新与个体梯度更新之间进行插值。

命题 *1* 考虑一个 *Stackelberg* 博弈，其中领导者使用正则化总梯度 $\nabla^\lambda J_1(\theta) = \nabla_{\theta_1} J_1(\theta) - \nabla^\top_{\theta_2\theta_1} J_2(\theta)(\nabla^2_{\theta_2} J_2(\theta) + \lambda I)^{-1}\nabla_{\theta_2} J_1(\theta)$ 进行更新。以下极限条件成立：*1)* $\nabla^\lambda J_1(\theta) \to \nabla J_1(\theta)$，当 $\lambda \to 0$ 时；*2)* $\nabla^\lambda J_1(\theta) \to \nabla_{\theta_1} J_1(\theta)$，当 $\lambda \to \infty$ 时。

4 实验

本节将介绍三个实验环境，并探讨以下问题：（**Q1**）：训练环境中的不同非对称性会如何影响智能体的性能和行为？（**Q2**）：ST-MADDPG 带给领导者的优势能否弥补较弱智能体的先天劣势？（**Q3**）：ST-MADDPG 使用的全导数更新方法是否优于另一种近似方法？（**Q4**）：如何利用 Stackelberg 信息结构解决现实世界中的机器人学问题？

请注意，与训练良好的单智能体或合作式多智能体强化学习环境不同，在大多数竞争式多智能体强化学习环境中，学习过程的累积奖励趋势并非单调上升。因此，为了评估智能体的表现，我们选择收集对局数据：让训练好的智能体与训练过程中共同进化的搭档或人工设计的参考对手进行多场对局。更多执行细节见第 4.1、4.2 和4.3节。

竞争式倒立摆。为了回答**Q1** 和**Q2**，我们提出了一种双人零和竞争游戏，其中每个智能体都要完成一项一维控制任务。如图2所示，该环境包含两个常规倒立摆智能体。两个智能体的动力学通过一根弹簧耦合，弹簧的两端分别连接到一个智能体的主体上。当两个智能体同时将各自的杆保持在直立位置时，双方都不会获得奖励。如果其中一个智能体失去平衡，那么从此之后的每个时间步，直到游戏结束，该智能体都会获得 -1 奖励。仍保持平衡的智能体则会在每个时间步获得 $+1$ 奖励，直到它也失去平衡并导致游戏结束。因此，每个智能体的目标都是防止自己的杆倒下，同时利用弹簧施加扰动力，设法破坏对手的平衡。

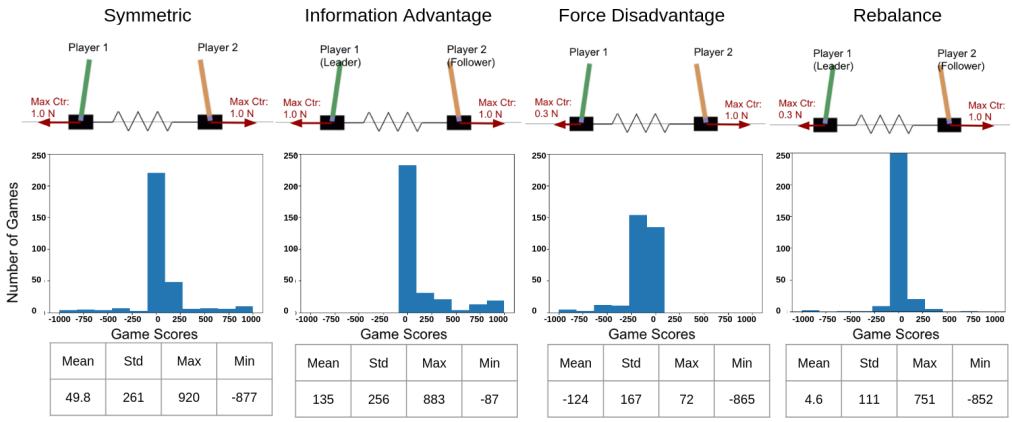
带有对抗性扰动的单足跳跃机器人。为了研究**Q3**，我们将首先聚焦于使用对抗训练 [22, 9]，为经典单足跳跃机器人环境创建鲁棒控制策略。在这里，第一个智能体控制经典的单足跳跃机器人，该机器人由四个刚性连杆和三个驱动关节组成。第二个智能体则学习对单足跳跃机器人的足部施加对抗性的二维力。

击剑博弈。为了进一步考察**Q3**，我们考虑 Yang 等人提出的一种零和竞争博弈。 [8]。该博弈体现了更具现实性的机器人学挑战，例如复杂的机器人运动学、高度不确定的状态转移动力学，以及高度非对称的博弈机制。这是一种双人攻防博弈：攻击者试图用剑攻击预先设定的目标区域，在不触碰防守者之剑的前提下最大化游戏得分；防守者则通过守护目标区域，力求将攻击者的得分降至最低。游戏规则详见附录A.4。

4.1 非对称优势下的学习

本实验明确研究了在对称设置和非对称环境下，训练所得智能体在竞争性双杆环境中的性能与行为。我们首先展示 ST-MADDPG 如何为某个智能体提供优势并提升其性能。随后我们表明，在一个智能体相较于另一个智能体具有施力优势的非对称环境中，可以使用 ST-MADDPG 保持智能体之间的性能平衡。

领导者优势。本实验从一个对称的竞争性双杆环境开始，在该环境中，两个智能体的行动能力相同。为了了解领导者优势这一固有特性



Mean	Std	Max	Min
49.8	261	920	-877

Mean	Std	Max	Min
135	256	883	-87

Mean	Std	Max	Min
-124	167	72	-865

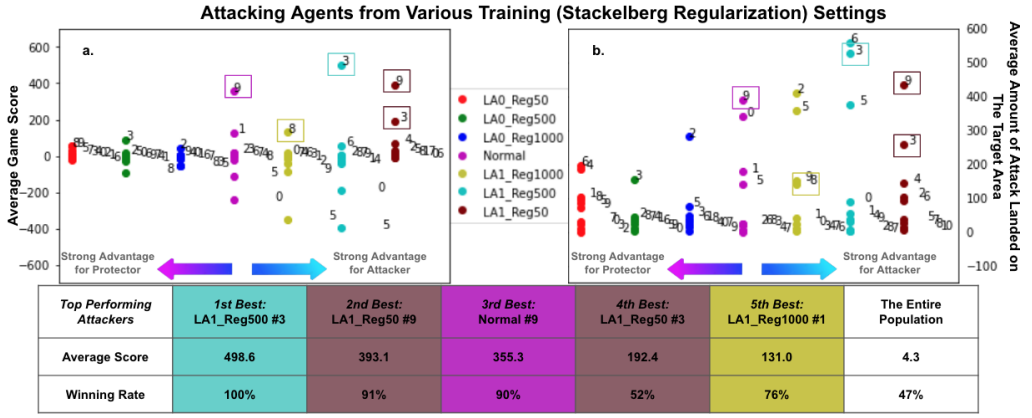
Mean	Std	Max	Min
4.6	111	751	-852

Fig图 2：四种不同变体中学习到的策略性能的统计分析 e
co竞争性 cartpoles 环境。游戏得分指玩家 1 的得分。

如何影响斯塔克尔伯格博弈结构中的自动课程学习过程，我们在竞争性双杆环境中分别运行了 MADDPG 和 ST-MADDPG 方法。MADDPG 训练代表对称环境，而 ST-MADDPG 训练则赋予玩家 1 领导者优势。对于这两种方法，我们分别使用四个不同的随机种子创建了四对智能体。为了比较两种训练方法下智能体的表现，我们进行了一场锦标赛，每种方法均获得 320 个游戏得分和轨迹。锦标赛的详情见附录 A.2。图 2 的前两列总结了两场锦标赛的统计数据。请注意，本节中的锦标赛游戏得分指玩家 1 的得分。因此，玩家 1 获胜时得分为正，玩家 2 获胜时得分为负，两名玩家平局时得分为零。

在对称设置下（即 MADDPG），玩家 1 和玩家 2 的表现相近。锦标赛的平均得分为 49。8。虽然大多数游戏的得分介于−90 至 90之间，但其余游戏的得分几乎覆盖了从 −877 至 920的整个得分范围。这表明，尽管两名玩家在大多数情况下的表现相近，但他们中的任何一方偶尔都能大幅胜过另一方。相比之下，在训练期间给予领导者优势时（即 ST-MADDPG），玩家 1 赢得了更多游戏，平均得分也更高，为135。玩家 2 在其获胜表现最佳的一场中仅获得 −87，这意味着跟随者从未能够显著胜过领导者。因此，领导者的整体表现优于跟随者。通过回放收集到的轨迹来观察智能体的行为时，我们发现，来自对称环境的两名玩家通常会通过弹簧相互推拉，展开激烈竞争。虽然他们能够保持自己的杆直立，但无法破坏另一智能体的平衡，因此在大多数比赛中都无法获胜。与此同时，对于来自 ST-MADDPG 的智能体，领导者学会了一种将跟随者拉出框架以赢得比赛的策略。机器人的行为演示视频可在项目网站上找到。

重新平衡非对称环境。 鉴于 ST-MADDPG 能够创造优势，从而提升领导者的表现，我们希望测试这种优势能否用于弥补非对称环境赋予某个智能体的劣势。我们通过赋予玩家 1 力劣势，创建了一个非对称竞争性双杆环境。在该环境中，玩家 1 的最大控制力有所降低，仅为玩家 2 最大控制力的30% 。随后，我们再次分别使用 MADDPG 和 ST-MADDPG（以玩家 1 为领导者）训练智能体，每种方法采用四个随机种子，并通过锦标赛生成评估数据。如图 2所示，在 MADDPG 训练后，面对显著的力劣势，玩家 1 的表现明显逊于玩家 2。然而，应用斯塔克尔伯格梯度更新后，两名玩家的表现趋于一致。平均得分为4.57，最高得分为 751，最低得分为 −852，**领导者优势能够弥补玩家 1 所面临的力劣势，**并产生了与上述对称环境相似的得分分布。



Top Performing Attackers	*1st Best:* LA1_Reg500 #3	*2nd Best:* LA1_Reg50 #9	*3rd Best:* Normal #9	*4th Best:* LA1_Reg50 #3	*5th Best:* LA1_Reg1000 #1	The Entire Population
Average Score	498.6	393.1	355.3	192.4	131.0	4.3
Winning Rate	100%	91%	90%	52%	76%	47%

图3：可视化了来自七种训练设置的攻击者的性能和行为。每个组的名称提供了有关训练设置的信息，其中第一部分表示领导者智能体（即 LA0：防守者，LA1：攻击者），第二部分表示正则化值。每种设置下有十个智能体，每个智能体的ID均标注在两个图表中相应数据点的**右侧**。对于对应的Y值较小的数据点，其ID的位置更靠右。**a.** 展示了每个攻击者与一个未参与训练的强基准保护者之间进行100场比赛的平均游戏得分。**b.** 展示了攻击者在单场游戏中执行的平均攻击次数（即攻击性）。在两个图表中，均突出显示了表现最佳的五个智能体（根据其获得的平均得分确定）。

4.2 单足跳跃机器人应对对抗性随机扰动

本研究着重推导并实现一种 MADDPG 算法变体，使用全导数学习更新来构建 Stackelberg信息结构。尽管其他个体学习算法在自动课程学习相关研究中也很受欢迎 [6, 7], ，但它们通过代理函数对去中心化价值网络进行估计，使得直接计算全导数不可行。另一种方法是在个体学习设置中，为领导者和跟随者设置不同的更新步数，以近似 Stackelberg信息结构 [23]。在 Stackelberg 博弈中，通过允许跟随者针对每个训练数据批次执行远多于领导者的更新步数，来近似跟随者的最佳响应更新。

在这个鲁棒控制问题中，我们首先比较 ST-MADDPG 和 MADDPG。随后，我们还在两种基于 PPO 的训练设置下评估了近似斯塔克尔伯格更新。第一种 PPO 方法以分散式方式使用原始 PPO 训练所有智能体，这种方法在自动课程学习文献中很常见 [6, 7, 8]。之后，我们通过让跟随者（对抗性扰动）采取比领导者（单足跳跃机器人）多十倍的更新步数，构建了一种带有近似 Stackelberg 动态的 PPO 变体 ST-PPO。

如表1所示，经过 ST-MADDPG 训练的单足跳跃机器人智能体，在对抗攻击（ST-MADDPG：平均奖励 5113.6；MADDPG：平均奖励 1601.6；平均**提升 319.3%**）和多种强度级别的随机扰动下（ST-MADDPG 的平均奖励至少高出**2.22× 倍**于 MADDPG）均显著优于经 MADDPG 训练的单足跳跃机器人智能体。相比之下，近似的 Stackelberg 信息结构在对抗训练中只能为领导者带来有限的提升（平均**11%**）。类似地，在随机扰动测试中，ST-PPO 训练的单足跳跃机器人智能体表现略优于 PPO 训练的智能体（最多高出**1.14**× 倍）。因此，**在对抗训练中赋予机器人领导者优势，可以进一步提高机器人控制策略的鲁棒性**。在计算复杂度相近的情况下（详见A.3.2），**在构建领导者优势方面，全导数更新比近似斯塔克尔伯格更新更有效**。关于该实验的详细讨论见附录A.3。

4.3 复杂环境下的协同进化

上述实验基于两个机器人动力学和游戏规则都较为简单的环境，但实际机器人应用场景可能要复杂得多。为了进一步了解 Stackelberg 结构下复杂环境中的竞争性协同进化，我们在击剑博弈中评估了多种训练设置。与先前的环境相比，击剑博弈体现了实际机器人问题中更为真实的挑战。两个人形智能体具有七个自由度，运动学十分复杂；此外，游戏中大量的接触也使环境转移模型具有很高的复杂度和不确定性。此外，该游戏机制也具有非对称性。防守者只有在目标区域内与攻击者接触时才能获得奖励，因此，攻击者通过发起攻击动作来主导协同进化过程。如果攻击者学会停止攻击以避免受到惩罚，那么双方智能体的策略更新都可能变得低效，因为它们会陷入在任务空间中探索高度次优区域的困境。因此，攻击者在游戏中扮演着更重要的角色，本实验将重点评估攻击者的行为和表现。

为了研究不同 Stackelberg 设置如何影响协同进化过程的质量，我们比较了来自 MADDPG 的攻击者和来自六种不同 ST-MADDPG 设置的攻击者。对于这六种 ST-MADDPG 设置，我们分别将防守者和攻击者指定为领导者，并采用小型（λ=50）、中型（λ=500）和大型（λ=1000）正则化值。实验测试了协同进化过程，其中两个智能体分别具有强、中、弱的领导者优势。对于七种设置中的每一种，我们使用 10 个不同的随机种子训练 10 对智能体。随后，我们让这 70 个经过训练的攻击者分别与一个精心设计的基于启发式的防守者策略进行 100 场对局评估。通过将防守者的剑置于目标区域与攻击者的剑上最接近目标区域的点之间，基于启发式的策略利用对游戏规则的内在知识来执行强大的防御策略。采用这一启发式基线策略，可以在攻击者之间进行公平比较。收敛到更高质量均衡的协同进化过程，应能产生更加稳健的攻击者策略，使其在与这一强大且未见过的防守者竞争时表现更佳。该基于启发式的策略的设计以及实验细节见附录A.4。

攻击者攻击性的提升。 在全部七种训练设置中，很大一部分攻击者收敛到一组保守策略，执行的攻击动作较少，游戏得分也相对较低（无论是正分（获胜）还是负分（失败））。鉴于发起攻击动作也意味着面临被防守者惩罚的巨大风险，这一结果并不令人意外。然而，如图3.b. 所示，随着攻击者在训练中逐渐占据更多优势，少数智能体变得越来越激进，在目标区域发动了更多攻击。

涌现复杂性。 竞争参与度的提升有助于形成更好的攻击策略。事实上，图3.a. 显示，表现最好的五名攻击者都位于图的右半部分，在那里，它们的共同进化对手并非斯塔克尔伯格博弈中的领导者。表现最好的两名攻击者来自两种 ST-MADDPG 设置，在这两种设置中，攻击者分别以 500 和 50 的正则化值担任领导者。它们都学会了诱使防守者移动到较难操控的联合配置，然后趁防守者部分受困、忙于脱离该配置时，以较低风险攻击目标区域。当然，一些其他激进策略也可能导致过度投入竞争，表现却不佳（例如图3中的 LA1 Reg1000 和 LA1 Reg500 的智能体 #5）。

5 结论与局限性

本研究探讨了多智能体强化学习在非对称、具有物理基础的竞争性游戏中的应用。我们提出了 Stackelberg-MADDPG 算法，将双人多智能体强化学习问题表述为斯塔克尔伯格博弈，并赋予系统中的一个智能体优势。我们证明，一个智能体相对于另一个智能体所固有的优势可能会使训练过程偏向不利的均衡，而我们提出的算法能够重新平衡此类环境，从而提升两个智能体的表现。我们的实验还表明，一些问题可能对随机种子更加敏感，这可能使所提方法的效果不那么明显。其他

局限性包括仅在双智能体竞争性场景中进行实验，以及训练期间需要访问对手的参数。

参考文献

[1] K. Zhang、Z. Yang 和 T. Bas¸ar。多智能体强化学习：理论与算法精选综述。*arXiv* 预印本 *arXiv:1911.10635*，2019年。[2] J. Filar 和 K. Vrieze。竞争性马尔可夫决策过程。Springer Science & Business Media，2012年。[3] D. Silver、T. Hubert、J. Schrittwieser、I. Antonoglou、M. Lai、A. Guez、M. Lanctot、L. Sifre、D. Kumaran、T. Graepel 等。通过自我对弈和通用强化学习算法精通国际象棋与将棋。*arXiv* 预印本 *arXiv:1712.01815*，2017年。[4] C. Berner、G. Brockman、B. Chan、V. Cheung、P. P. Debiak、C. Dennison、D. Farhi、Q. Fischer、S. Hashme、C. Hesse 等。大规模深度强化学习下的《Dota 2》。*arXiv* 预印本 *arXiv:1912.06680*，2019年。[5] O. Vinyals、I. Babuschkin、W. M. Czarnecki、M. Mathieu、A. Dudzik、J. Chung、D. H. Choi、R. Powell、T. Ewalds、P. Georgiev 等。运用多智能体强化学习达到《星际争霸 II》宗师级水平。*Nature*，575(7782):350–354，2019年。[6] M. Dennis、N. Jaques、E. Vinitsky、A. Bayen、S. Russell、A. Critch 和 S. Levine。通过无监督环境设计实现涌现复杂性与零样本迁移。*arXiv* 预印本 *arXiv:2012.02096*，2020年。[7] B. Baker、I. Kanitscheider、T. Markov、Y. Wu、G. Powell、B. McGrew 和 I. Mordatch。多智能体自课程中涌现的工具使用。*arXiv* 预印本 *arXiv:1909.07528*，2019年。[8] B. Yang、G. Habibi、P. Lancaster、B. Boots 和 J. Smith。通过竞争性人机交互促进身体活动。载于第五届机器人学习年会，2021年。[9] L. Pinto、J. Davidson、R. Sukthankar 和 A. Gupta。鲁棒对抗性强化学习。载于国际机器学习会议，第2817–2826页。PMLR，2017年。[10] T. Bansal、J. Pachocki、S. Sidor、I. Sutskever 和 I. Mordatch。多智能体竞争带来的涌现复杂性。*arXiv* 预印本 *arXiv:1710.03748*，2017年。[11] J. Won、D. Gopinath 和 J. Hodgins。用于控制参与双人竞技体育运动的物理模拟角色的策略。*ACM* 图形学汇刊（*TOG*），40(4):1–11，2021年。[12] J. Foerster、R. Y. Chen、M. Al-Shedivat、S. Whiteson、P. Abbeel 和 I. Mordatch。具备对手学习意识的学习方法。载于第*17*届自主智能体与多智能体系统国际会议论文集，AAMAS '18，第122–130页，2018年。[13] M. Prajapat、K. Azizzadenesheli、A. Liniger、Y. Yue 和 A. Anandkumar。竞争性策略优化。*arXiv* 预印本 *arXiv:2006.10611*，2020年。[14] L. Zheng、T. Fiez、Z. Alumbaugh、B. Chasnov 和 L. J. Ratliff。斯塔克尔伯格行动者-评论家：博弈论强化学习算法。*arXiv* 预印本 *arXiv:2109.12286*，2021年。[15] T. Lin、C. Jin 和 M. I. Jordan。近乎最优的极小极大优化算法。载于学习理论会议，第2738–2779页。PMLR，2020年。[16] T. Fiez、B. Chasnov 和 L. J. Ratliff。斯塔克尔伯格博弈中的隐式学习动态：均衡刻画、收敛性分析与实证研究。载于国际机器学习会议，2020年。[17] H. Von Stackelberg。市场结构与均衡。Springer Science & Business Media，2010年。[18] Y. Yang 和 J. Wang。从博弈论视角概述多智能体强化学习。*arXiv* 预印本 *arXiv:2011.00583*，2020年。[19] T. Bas¸ar 和 G. J. Olsder。动态非合作博弈论。SIAM，1998年。

[20] R. Lowe、Y. WU、A. Tamar、J. Harb、O. Pieter Abbeel 和 I. Mordatch。面向混合合作竞争环境的多智能体行动者-评论家。神经信息处理系统进展，30:6379–6390，2017年。[21] S. G. Krantz 和 H. R. Parks。隐函数定理：历史、理论与应用。Springer Science & Business Media，2002年。[22] Y. Duan、X. Chen、R. Houthooft、J. Schulman 和 P. Abbeel。连续控制深度强化学习基准测试。载于国际机器学习会议，第1329–1338页。PMLR，2016年。[23] A. Rajeswaran、I. Mordatch 和 V. Kumar。基于模型的强化学习的博弈论框架。载于国际机器学习会议，第7953–7963页。PMLR，2020年。[24] B. A. Pearlmutter。海森矩阵的快速精确乘法。神经计算，6(1):147–160，1994年。[25] A. Rajeswaran、C. Finn、S. M. Kakade 和 S. Levine。基于隐式梯度的元学习。神经信息处理系统进展，32，2019年。[26] T. P. Lillicrap、J. J. Hunt、A. Pritzel、N. Heess、T. Erez、Y. Tassa、D. Silver 和 D. Wierstra。深度强化学习下的连续控制。*arXiv* 预印本 *arXiv:1509.02971*，2015年。

A 算法与实验详情

A.1 ST-MADDPG

算法 1介绍了本文提出的 ST-MADDPG 算法。在本研究中，我们通过网格搜索为第一个和第二个环境选择隐式映射正则化超参数 λ。一般而言，神经网络可能具有高度非凸性，海森矩阵逆也可能病态。较大的正则化可以防止梯度爆炸，并使学习动态更加平滑，这一点在我们的实验以及其他斯塔克尔伯格学习应用中均有所体现 [16, 14]。如何通过最优或自适应地选择正则化，在斯塔克尔伯格学习与普通梯度学习之间进行权衡，是未来的研究方向。

算法 1： ST-MADDPG 算法

对于 回合 $k = 1, 2, \ldots, K$ **执行**

接收初始状态 s_0;

对于 $t = 1, 2, \ldots, T$ **do**

对于每个智能体 i，根据当前策略选择动作$a^i = \mu_{\theta_i}(s)$;

执行动作 (a^1, a^2)，并观测奖励 r 和新状态 s';

将 (s, a^1, a^2, r, s') 存储到经验回放缓冲区 $\mathcal{D}$中;

$s \leftarrow s'$;

从 $\mathcal{D}$中随机抽取一个包含 N状态转移的小批量 $(s_j, a_j^1, a_j^2, r_j, s_j')$;

设置 $y_j = r_j + Q_{w'}(s_j, \mu_{\theta_1'}(s_j)\, \mu_{\theta_2'}(s_j))$;

通过最小化损失来更新评论家：

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} [\left(Q_w(s_j, a_j^1, a_j^2) - y_j\right)^2]$$

使用通过（1）计算得到的总梯度更新领导者策略：

$$\theta_1 \leftarrow \theta_1 + \alpha^1 \nabla^\lambda J(\theta_1, \theta_2)$$

使用策略梯度更新跟随者策略：

$$\theta_2 \leftarrow \theta_2 - \alpha^2 \nabla_{\theta_2} J(\theta_1, \theta_2)$$

更新目标网络：

$$w' \leftarrow \tau w + (1 - \tau) w'$$
$$\theta_i' \leftarrow \tau \theta_i + (1 - \tau) \theta_i'$$

end

end

A.1.1 计算复杂度

与常规策略梯度更新不同，领导者的策略更新需要计算 $\nabla_{\theta_1 \theta_2} J(\theta_1, \theta_2)$ 项的雅可比向量积以及逆海森向量积，$\nabla^2_{\theta_2} J(\theta_1, \theta_2) \nabla_{\theta_2} J(\theta_1, \theta_2)$，见方程 1。PyTorch 中的自动微分引擎会计算雅可比向量积。逆海森向量积项通过迭代共轭梯度（CG）方法计算，该方法同样由我们使用自动微分引擎实现。因此，斯塔克尔伯格梯度更新步骤的所有计算都在图形处理器上完成。每次 CG 迭代都会计算一个海森向量积，其计算成本是梯度的 ~ 1.5 倍 [24]。根据经验，进行五次 CG 迭代即可达到足够的数值精度，因此领导者更新的成本是 ~ 7.5 常规梯度的倍数 [25, 16]。然而，在实践中，ST-MADDPG 所需时间仅约为 MADDPG 的两倍，因为强化学习的主要瓶颈在于轨迹采样，而非梯度计算。

A.2 竞争式倒立摆

竞争式倒立摆游戏在训练和实验中的最长时间为 1000 个时间步。在训练开始时，每个智能体都会执行 10^4 步均匀随机动作选择，以便更好地进行探索 [26]。经验回放缓冲区的大小设为 10^6, , 学习率设为 10^{-3}，并以 10^2为批大小更新每个智能体。我们进行了两场锦标赛来采集评估数据，一场针对 MADDPG 得出的智能体，另一场针对 ST-MADDPG 得出的智能体。在一场锦标赛中，四个玩家 1 智能体（由四个随机种子训练得出）各自与四个玩家 2 智能体中的每一个进行 20 场比赛，共得到 320 个游戏得分和轨迹。在所有 ST-MADDPG 训练中，正则化值 λ 均设为 1，这样既能防止梯度爆炸，又能在训练中形成很强的领导者优势。这一评估过程使智能体不仅能与其原本的共同进化对手进行比赛，还能与使用不同随机种子训练出的对手交手，从而为每组训练提供更全面的总结。

A.3 单足跳跃机器人

在训练过程中，为了缩短训练时间，游戏的最大长度被限制为 1000 个时间步。然而，在评估实验中，所有试验的最大长度均为 3000 个时间步，以便更好地区分智能体的性能。与竞争性倒立摆环境类似，该环境中的每个训练过程都以 10^4 步的均匀随机动作选择开始，以便更好地进行探索 [26]。经验回放缓冲区的大小设置为 10^6, , 学习率为 8×10^{-5}，每个智能体均使用批大小为 10^2 的批次进行更新。在该环境中，所有 ST-MADDPG 训练中的正则化值λ 均设置为 5000。

A.3.1 实验结果

我们分别使用 MADDPG、ST-MADDPG、PPO 和 ST-PPO 算法，在 20 个随机种子下创建了 20 对单足跳跃机器人和对抗者。每对智能体均进行了 10 场游戏评估，因此每种方法获得了 200 个游戏得分。

对抗攻击。我们发现，在对抗性攻击下，经过 ST-MADDPG 训练的单足跳跃机器人存活时间显著长于经过 MADDPG 训练的机器人（平均奖励分别为 **5113.6** 和 **3333.0** ）。然而，ST-PPO 训练出的单足跳跃机器人智能体（平均奖励为**522.0** ）仅比 PPO 训练出的智能体（平均奖励为**473.6** ）存活时间略长。

高强度随机扰动。在三种环境下测试了所有单足跳跃机器人智能体的泛化能力：一种环境没有扰动，另外两种环境分别具有不同等级的强随机扰动（例如， $0N$、 $0.1N$和 $10.N$）。每个智能体在每种环境中运行 100 次试验，表。1 对比了采用两种训练方法的智能体的表现。尽管训练期间对抗者的最大强度被限制在 $0.001N$，但部分智能体（例如，MADDPG：50%，ST-MADDPG：100%）在最大强度为 $10N$的随机扰动下，平均奖励仍超过 1000。经过 ST-MADDPG 训练的智能体在三个强度等级下的表现均远超经过 MADDPG 训练的智能体。因此，ST-MADDPG 策略具有足够的鲁棒性，能够在未见过的场景中保持高性能。

A.3.2 全导数斯塔克尔伯格更新与近似斯塔克尔伯格更新

如第A.1.1节所述，全导数 Stackelberg 更新步需要执行常规梯度步的 $\sim$ 7.5 倍。在我们的实验中，跟随者的近似斯塔克尔伯格更新所需的更新步数是领导者的十倍。例如，如果常规策略梯度方法在每个训练轮次中为单个智能体更新策略需要 t 时间（不包括轨迹采样时间），则全导数更新方法对领导者的更新需要 $7.5t$ ，对跟随者的更新需要 t 。而近似更新方法对领导者的更新需要 t ，对跟随者的更新需要 $10t$ 。因此，在我们的实验设置中，全导数更新的计算复杂度略低于近似更新。在计算复杂度相近的情况下，全导数更新在构建 Stackelberg信息结构方面明显更加有效。通过进一步增加跟随者的更新步数，可以在计算开销与效果之间进行权衡，从而提升近似方法的有效性。正如 Rajeswaran

等人 [23], 所观察到的，当跟随者比领导者多执行25×更新步时，近似方法的性能更加令人满意。

对抗性扰动				
	ST-MADDPG	MADDPG	ST-PPO	PPO
平均得分	5113.6	1601.6	521.8	473.6
Std	3333.0	753.2	147.1	91.4
统计显著性	$p < 0.01$		$p > 0.05$	
无随机扰动				
	ST-MADDPG	MADDPG	ST-PPO	PPO
平均得分	5415.2	2436.8	510.0	495.7
Std	3730.8	2265.5	84.6	102.0
统计显著性	$p < 0.05$		$p > 0.05$	
0.1N 随机扰动				
	ST-MADDPG	MADDPG	ST-PPO	PPO
平均得分	5463.3	2423.6	530.9	486.0
Std	3575.6	2178.2	101.3	75.7
统计显著性	$p < 0.05$		$p > 0.05$	
10 N 随机扰动				
	ST-MADDPG	MADDPG	ST-PPO	PPO
平均得分	4928.3	1991.2	540.1	474.3
Std	3588.0	2084.6	151.3	80.8
统计显著性	$p < 0.05$		$p > 0.05$	

表1：在三个等级的随机扰动下训练的 ST-MADDPG、MADDPG、ST-PPO 和 PPO 单足跳跃机器人智能体的性能比较。通过统计显著性检验（即 u检验）比较两种算法的 Stackelberg 版本与常规版本之间的评估数据。

A.4 击剑游戏

游戏规则。 算法 2 概述了击剑博弈的计分机制。攻击者（或图1中的对手）将剑置于两个智能体之间的橙色球形（目标）区域内，即可获得一分。但如果攻击者的剑置于目标区域内，并同时接触到防守者（或图1中的对手方）的剑，攻击者将受到十分的得分惩罚。同时，左侧防守者智能体的目标是通过给予攻击者得分惩罚，尽可能降低攻击者的得分。此外，如果防守者的剑置于目标区域内，并被动等待攻击者进攻超过两秒，攻击者将获得十分奖励。每个智能体都有一个七维连续控制空间。每场博弈持续1000个时间步（即10秒）

算法2： 击剑博弈计分机制

初始化： 游戏得分 $s = 0$；时间步长 $= 0.01$ 秒；游戏时域 $= 10$ 秒
球拍 a →对手的球拍
球拍 p →主角的球拍
目标 →目标区域
对于 此游戏中的每个时间步**执行**
  如果球拍 a 位于目标内**则**
    如果球拍 a 与球拍 p 接触**则**
      s-= 10
    else
      s += 1
    end
  如果球拍 p 连续超过 *200* 个时间步长处于目标内**则**
    s += 10
end

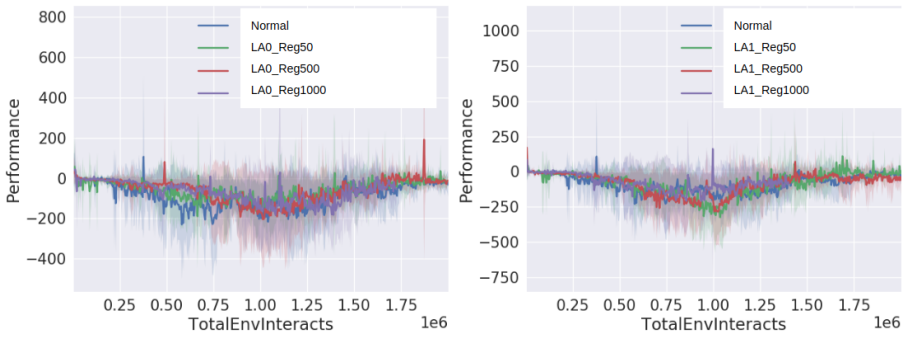


图 4：两幅图表展示了在六种不同设置下，共同进化过程中防守者（即智能体 0）的平均游戏得分。

基于启发式的主角策略。 我们的目标是设计一种强大的基准启发式策略，以营造激烈的人机游戏体验。根据对世界的观测，机器人将其球拍调整至与人类球拍垂直的方向，并在 x、y 和 z 轴上分别施加从 -25 度到 25 度均匀随机采样的角度偏移。为了确保机器人始终执行具有竞争性的防守，该策略指令机器人将球拍中心定位在目标区域与人类球拍上距离目标区域最近的点之间：

$$\bar{b_p} = \bar{tar} + (\bar{h_{close}} - \bar{tar}) \cdot uniform(0.5, 1)$$
$$\bar{h_{close}} = \bar{h_{low}} + ht \cdot (\bar{h_{up}} - \bar{h_{low}})$$
$$ht = \max\left(0, \min\left(1, (\bar{tar} - \bar{h_{low}}) \cdot (\bar{h_{up}} - \bar{h_{low}})/(2 \cdot L_{sword})\right)\right)$$

其中，$\bar{b_p}$、$\bar{tar}$、$\bar{h_{up}}$ 和 $\bar{h_{low}}$ 分别表示机器人球拍坐标系的位置、目标区域的中心、人类球拍的上端和人类球拍的下端。$\bar{h_{close}}$ 表示人类球拍上距离目标区域中心最近的点，L_{sword} 表示球拍的长度。函数 $uniform(0.5, 1)$ 随机决定机器人球拍与人类球拍之间应保持的距离。此外，机器人有 50% 的概率执行根据上一个时间步计算出的期望球拍位置，而不是最新的期望位姿。加入的不确定性为机器人的行为引入了随机性。当机器人移动速度快于或等于对手时，这种启发式方法能让机器人在击剑博弈中占据优势。在本实验中，对手智能体和主角智能体的身体能力相同。

预训练。 由于击剑环境相比另外两个实验环境复杂得多，我们选择使用同一对预训练策略初始化全部七种设置（即图3中的 LA0 Reg50、LA0 Reg500、LA0 Reg1000、常规、LA1 Reg50、LA1 Reg500 和 LA1 Reg1000）的协同进化过程。预训练策略通过使用 MADDPG，对每个智能体迭代执行两轮最佳响应个体更新而创建，这与 Yang 等人的预训练过程类似。 [8]每次最佳响应更新都从 10^4 步均匀随机动作选择开始。经验回放缓冲区大小设置为 10^6,，学习率为 8×10^{-4}，每个智能体的批大小设置为 100。

协同进化。 在全部 10 个随机种子下，七种训练设置中的防守者（即智能体 0）和攻击者（即智能体 1）智能体均使用预训练策略进行初始化。在每个协同进化过程开始时，不进行均匀随机动作选择。经验回放缓冲区大小设置为 10^6,，学习率为 8×10^{-5}，每个智能体的批大小设置为 1024。如图4 所示，智能体偏离了预训练策略，并在协同进化过程的中期进行了剧烈的策略更新，同时伴随着显著的性能波动。随后，在协同进化过程结束时，智能体收敛到了新的均衡。

$\pi^2_{\theta_2}$