arXiv:2303.06164v1 [cs.LG] 10 Mar 2023

理解质量-多样性与深度强化学习之间的协同效应

Bryan Lim*
bryan.lim16@ic.ac.uk
伦敦帝国理工学院
伦敦 , U.K.

Manon Flageat*
manon.flageat18@ic.ac.uk
伦敦帝国理工学院
伦敦 , U.K.

Antoine Cully
a.cully@imperial.ac.uk
伦敦帝国理工学院
伦敦 , U.K.

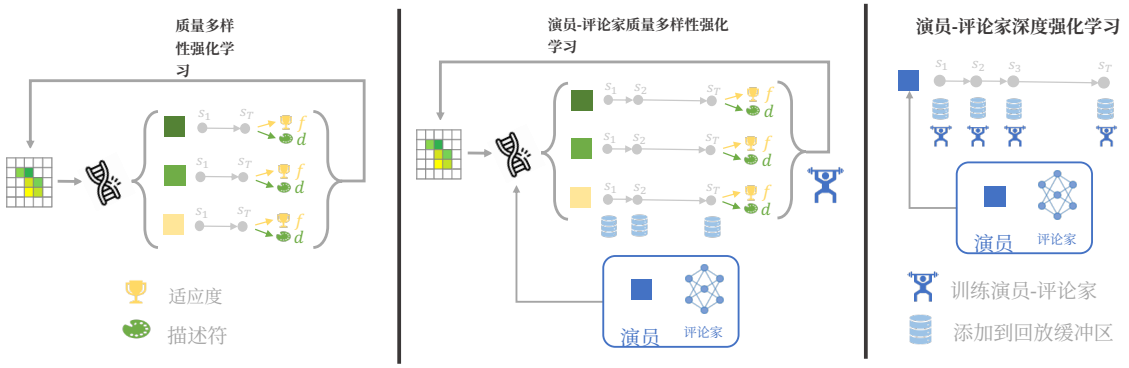

图 1：示意图展示了本文比较的设置：质量多样性强化学习、演员-评论家质量多样性强化学习和演员-评论家深度强化学习。质量多样性强化学习通常指将质量多样性应用于强化学习领域，以寻找多样化策略集合；而演员-评论家质量多样性强化学习方法则将演员-评论家深度强化学习方法中的元素融入质量多样性，以利用深度强化学习领域的特性。该图突出显示了演员-评论家质量多样性强化学习与演员-评论家深度强化学习设置之间的主要区别：（1）演员-评论家模型的更新频率；（2）用于收集数据的策略。

摘要

质量-多样性与深度强化学习之间的协同效应，催生了强大的混合型质量多样性强化学习算法。这些算法展现出巨大潜力，汇集了两个领域的优势。然而，尽管其他强化学习算法取得了显著进展，先前的混合方法却只采用了一种深度强化学习算法（TD3）。此外，质量多样性与强化学习的优化流程存在根本差异，若能采用更具原则性的方法，则有望从中受益。我们提出广义演员-评论家 QD-RL，这是一个统一的模块化框架，适用于 QD-RL 设置中的演员-评论家深度强化学习方法。该框架为研究深度强化学习在 QD-RL 设置中的启示提供了途径，而这正是推动 QD-RL 取得进展的一种重要且高效的方式。我们介绍了两种新算法 PGA-ME（SAC）和 PGA-ME（DroQ），将深度强化学习的最新进展应用于 QD-RL 设置，并成功解决了使用现有 QD-RL 算法无法解决的人形机器人环境问题。然而，我们也发现，并非所有深度强化学习的经验都能有效迁移至 QD-RL。重要的是，本文还表明，QD-RL 中的演员-评论家模型通常训练不足，无需增加环境评估次数也能提升性能。

CCS 概念

• **计算方法** →**进化机器人学**。

关键词

质量-多样性、深度强化学习、神经进化

*两位作者对本研究的贡献相同。

1 引言

深度强化学习（Deep RL）已取得显著进展，并在广泛的应用领域中展现出令人瞩目的突破性成果 [1, 9, 11, 30]。同样，质量多样性（QD）算法也已证明卓有成效，具备多样性和目标切换等互补特性，因此成为强大的优化工具 [7, 10, 16]。QD算法可用于为强化学习问题寻找多样化解集，这通常称为QD-RL。虽然也可以将其视为深度强化学习算法的替代方案，但近期已有研究探索将QD与深度强化学习结合，以发挥协同效应，并展现出应用潜力 [10, 26, 27, 37]。先前的工作将TD3 [15]RL算法与QD（MAP-Elites）结合，除了传统的变异算子外，还引入了梯度引导的变体（PGA-ME [26] 和QD-PG [27]）。事实证明，在QD-RL设置下，这些方法一直是表现最佳的算法 [3, 34]。

鉴于这两个领域都前景广阔，将深度强化学习算法中的洞见有效地应用于质量多样性强化学习至关重要。此外，深度强化学习是一个规模庞大且发展迅速的领域，评估其不断取得的进步，是推动质量多样性强化学习发展的有效途径。然而，现有的质量多样性强化学习研究始终只采用了一类深度强化学习算法，即 TD3，尽管 SAC [18]、REDQ [5]，和 DroQ [19] 等方法在算法上已有其他改进。如果能够将其中一些洞见和进展加以应用，质量多样性强化学习算法很可能还有进一步提升的空间。反过来，如图 1 所示，质量多样性与深度强化学习的设置存在根本差异。若要在质量多样性强化学习中实现深度强化学习的优势，就必须考虑并理解这些差异。

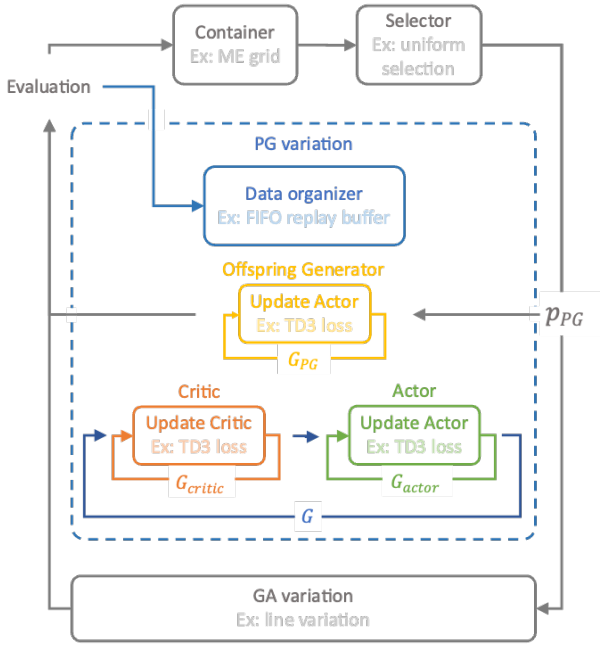

图2：广义演员-评论家 QD-RL 框架示意图。图中的每个实线框都是一个模块，可轻松替换，以集成深度强化学习的最新进展。主循环（灰色）遵循传统的质量多样性流程 [8]。我们提出的广义框架新增了一种 PG变体（蓝色）。多个模块构成了框架的核心：演员（绿色）、评论家（橙色）、后代生成器（黄色），以及用于收集环境状态转移的数据整理器（蓝色）。

在本研究中，我们提出了广义演员-评论家 QD-RL，这是一种模块化框架，如图2所示。该框架旨在通过促进对深度强化学习在质量多样性强化学习算法中的进展进行评估，推动质量多样性强化学习算法的发展。该通用框架的每个模块都可以轻松替换，以整合深度强化学习研究中的新思路和新洞见，并构建新的质量多样性强化学习算法。此外，广义演员-评论家 QD-RL 还是一种强大的分析工具，可用于研究质量多样性强化学习与深度强化学习设置之间的根本差异。

利用广义演员-评论家 QD-RL，我们开展了广泛研究，以更好地理解深度强化学习与质量多样性强化学习之间的协同效应和差异。首先，我们通过调整广义演员-评论家 QD-RL 中演员和评论家训练模块的组件，引入 PGA-ME（SAC）和 PGA-ME（DroQ）。借助该框架的模块化特性，这些算法分别评估了深度强化学习中的三项关键进展在质量多样性强化学习中的作用：(i) 最大熵强化学习，(ii) 增加评论家网络训练，以及 (iii) 评论家正则化。通过 PGA-ME（SAC），我们首次在质量多样性强化学习设置下解决了人形机器人任务。有趣的是，借助 PGA-ME（DroQ），我们发现深度强化学习中的进展并非都能轻易迁移到质量多样性强化学习中。值得注意的是，质量多样性强化学习中的策略多样性使得深度强化学习中的正则化机制变得冗余。其次，我们框架的严谨形式化促使我们重新审视现有 QD-RL 算法中演员和评论家网络的训练步数。我们发现，先前的 QD-RL 研究显著低估了这一数值，并表明只要充分训练这些网络，无需增加任何模型或评估次数，就能提升 PGA-ME 的性能。我们使用现有 QD-RL 算法作为基线方法，在三个 QD-RL 任务上通过实验展示了这些结果。

2 相关工作

2.1 质量-多样性

与旨在搜索单个目标最优解的传统优化算法不同，质量多样性算法旨在搜索由多种高性能解组成的种群 [4, 8, 28]。质量多样性方法已在众多应用领域展现出优势，例如机器人学 [7]、电子游戏设计 [17]、工程优化 [16]，甚至药物 [35] 和材料 [20] 的发现。质量多样性算法受生物进化启发，在优化过程中明确保留多个行为生态位。该过程为每个解 θ 赋予一个行为描述符 $d(\theta)$ 值，此外还赋予进化计算中通常为解指定的任务适应度或目标 $F(\theta)$ 值。质量多样性算法通过局部竞争，只保留落在同一生态位内（即具有相似行为描述符 d）的适应度最高的解。MAP-Elites (ME) [25], 是一种广受欢迎的质量多样性算法，其优势在于简单；它通过将行为描述符空间离散化为等间距网格，简化了这一竞争过程。

质量多样性优化通常由无梯度突变算子驱动，这是因为它源自进化计算。然而，在处理较大的搜索空间时，突变通常效率低下，例如深度神经进化的应用，其中神经网络的参数（权重和偏置）构成搜索空间。为了扩展到更高维的搜索空间（数千个参数），梯度方法与进化方法相结合的混合方法日益受到重视。ME-ES [6] 通过进化策略探索了自然梯度的应用。另一类样本效率更高的方法，则考虑采用离策略深度强化学习算法中的演员-评论家方法进行策略梯度更新 [26, 27]。我们的工作属于后一类方法，具体而言，是在 PGA-ME [26] 算法的基础上开展研究。

2.2 深度强化学习

强化学习（RL）长期以来一直被用于解决序贯决策问题。随着深度学习的兴起，神经网络强大的函数逼近能力使强化学习方法得以成功应用于范围广泛得多的问题。这就是深度强化学习。此后，深度强化学习方法的成功应用涵盖众多领域，包括在 Atari 游戏套件 [24],中达到超越人类的表现、在围棋 [30] 和星际争霸 [36], 等复杂游戏中击败专家、实现稳健的足式运动 [22] 以及灵巧的机器人操作 [1]。近来，深度强化学习还助力数学 [11] 和能源生产 [9]领域取得科学突破。

尽管取得了诸多成就，深度强化学习仍面临一些挑战，这些挑战源于习得策略的脆弱性，表明其缺乏

算法1： 单策略深度强化学习训练

输入： 评论家步数 G_{Q_ψ}，策略步数 $G_{\pi_{\theta_a}}$
初始化： 评论家 Q_ψ、演员 π_{θ_a}、回放缓冲区 $\mathcal{D}$
对于 $j = 1, \ldots, N$，**执行**
  对于 $t, \ldots, T$，**执行**
    // 在环境中收集数据
    $a_t \sim \pi_{\theta_a}(a_t|s_t)$
    $s_{t+1} \sim p\ (s_{t+1}|s_t,\ a_t)$
    $\mathrm{D} \leftarrow (s_t,\ a_t,\ s_{t+1},\ r_t)$
    // 更新评论家和演员
    for $g_Q = 1, \ldots, G_{critic}$ **do**
      $Q_\psi \leftarrow$ 更新_评论家$(Q_\psi,\ \mathrm{D})$
    for $g_\pi = 1, \ldots, G_{actor}$ **执行**
      $\pi_{\theta_a} \leftarrow$ 更新_演员$(Q_\psi,\ \pi_{\theta_a},\ \mathrm{D})$
返回 π_{θ_a}

算法 2： TD3 评论家和策略更新

定义 更新_评论家$(Q_\psi, \mathcal{D})$:
  // 从 D 中采样 N 个状态转移
  $(s_t, a_t, s_{t+1}, r_t) \sim \mathrm{D}$
  // $Q_{\psi'}$ 和 $\pi_{\theta'}$ 表示目标网络
  $\epsilon \sim$ 裁剪$(\mathrm{N}(0,\ \sigma_p)\ -c,\ c)$
  $y = r(s_t,\ a_t,\ s_{t+1}) +$ 最小值$_{i=1,\ 2}\, Q_{\psi'_i}(s_{t+1},\ \pi_{\theta'}(s_{t+1}) + \epsilon)$
  $\psi_i \leftarrow \frac{1}{N} \sum (y - Q_{\psi i}(s,\ a))^2$
  返回 Q_ψ
定义更新_策略$(Q_\psi, \pi_{\theta_a}, \mathcal{D})$:
  // 从 D 中采样 N 个状态转移
  $(s_t,\ a_t,\ s_{t+1},\ r_t) \sim \mathrm{D}$
  $\theta_a \leftarrow \frac{1}{N} \sum Q_{\psi_1}(s_t,\ a_t)$
  回报 π_{θ_a}

泛化能力。质量多样性和进化领域的理念已展现出应对这些挑战的潜力。例如，通过增加所习得行为以及用于学习的任务和环境的多样性。受质量多样性启发的研究此前已应用于深度强化学习，在困难的探索任务中实现了最先进的性能 [10]，并在开放式学习的自动课程学习领域开展了开创性工作 [37, 38]。

我们的工作利用了深度强化学习领域取得的进展，并研究如何将这一设置中的改进和进展有效地应用于质量多样性强化学习设置。我们还更具体地考察了 SAC [18] 和 DroQ [19] 算法的一些关键改进，这些算法属于离策略深度强化学习算法。下一节将详细介绍这些算法。

3 预备知识

3.1 问题设定

强化学习。强化学习（RL）问题描述的是这样一种情境：智能体与环境交互，并通过奖励以及环境状态的变化获得反馈。智能体的目标是最大化这一奖励。更正式地说，RL 可定义为一个马尔可夫决策过程（MDP）[32]，由元组 (S, A, P, R) 构成，其中 S 和 A 分别是环境中状态和动作的集合。$\mathrm{P}(s_{t+1}|s_t, a_t)$ 表示在执行动作 a_t 的情况下，从状态 s_t 转移到 s_{t+1} 的概率。奖励函数用于表示我们希望智能体学习的任务，并定义在时间步 $r_t = r(s_t, a_t, s_{t+1})$ 执行动作 a_t、从状态 s_t 转移到 s_{t+1} 时获得的奖励。策略 $\pi_\theta(a_t|s_t)$ 决定智能体如何根据当前状态 s_t 选择下一个动作 a_t。因此，RL 的目标是优化策略 π_θ 的参数 θ，使其最大化整个回合轨迹 $\tau = (s_0, a_0, \ldots, s_T, a_T)$ 上的期望累积奖励 $R(\tau) = \sum_{t=0}^{T} r_t$；该轨迹对应于回合期间观测到的状态：$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\mathrm{R}(\tau)]$。

质量多样性强化学习。我们将用于强化学习问题的质量多样性算法称为 QD-RL 设置 [13, 26, 27, 34]。在 QD-RL 中，解的适应度或目标值 F 定义为总回合奖励，而描述符 d 定义为状态-动作轨迹 τ 的函数：

$$F(\theta) = J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\mathcal{R}(\tau)] \quad \text{and} \quad d(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[d(\tau)] \quad (1)$$

与深度强化学习只搜索单个目标最大化策略 π_θ 不同，质量多样性强化学习搜索的是一个由多种不同的、Θ 目标最大化策略组成的种群。PGA-ME [26],QD-PG [27]和 CMA-MEGA [34]都是 QD-RL算法的示例，它们以 TD3 [15]的形式使用深度强化学习算法。

3.2 深度强化学习中的演员-评论家学习

当前大多数深度强化学习方法都依赖于近似动作价值函数 $Q(s_t,\ a_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{k=0}^{T-t} \gamma^k r_{t+k+1} \mid s_t, a_t\right]$，该函数编码了在状态 s_t 中执行动作 a_t 后遵循策略 π_θ 所获得的期望回报。此近似通过一个称为评论家网络的深度神经网络 Q_ψ 实现。评论家通过计算梯度来改进策略（或演员）π_θ。同时使用演员网络和评论家网络的方法称为演员-评论家。演员-评论家深度强化学习方法通常遵循算法1所示的过程。它们通常会在以下步骤间迭代：使用演员策略 π_θ 与环境交互，经过 G_{Q_ψ} 个梯度步学习评论家 Q_ψ，然后经过 G_{π_θ} 个梯度步更新演员策略 π_θ。常用的演员-评论家深度强化学习算法尤其包括 TD3 [15]，其评论家和演员的更新过程详见算法 2。此前的质量多样性强化学习研究也几乎一律采用 TD3 作为首选的深度强化学习算法 [26, 27]。本文将探讨从替代性演员-评论家深度强化学习方法中获得的见解。

4 深度强化学习的进展

本节将总体介绍与质量多样性强化学习相关的主要深度强化学习进展。

4.1 SAC：最大熵强化学习

总体而言，SAC 与 TD3 十分相似，因为它同样是无模型、离策略的。软演员-评论家（SAC）[18] 是一种基于最大熵强化学习框架的算法。它在最大化奖励的基础上，在目标函数中引入了额外的熵最大化项（公式 2）。熵最大化的目的是鼓励策略进行探索，并且还可能捕捉到多种最优行为模式。

$$J(\pi_\theta) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}[r(s_t, a_t) + \alpha\mathcal{H}(\pi(.|s_t)] \quad (2)$$

为适应公式 2 中的新目标，SAC 有几项关键区别。首先，SAC 依赖随机策略，

算法 3： SAC 评论家与策略更新

def 更新_评论家(Q_ψ, $\mathcal{D}$):
  // 从 D 中采样 N 状态转移
  $(s_t, a_t, s_{t+1}, r_t) \sim$ D
  // $Q_{\psi'}$ 表示目标网络
  $y = r_t +$ 最小值$_{i=1,2}\, Q_{\psi'_i}(s_{t+1}, \pi_\theta(.|s_{t+1}))$
  $\psi_i \leftarrow \frac{1}{N} \sum (y - Q_{\psi_i}(s, a))^2$
  返回 Q_ψ
定义更新_策略(Q_ψ, π_{θ_a}, $\mathcal{D}$):
  // 从 D 中采样 N 状态转移
  $(s_t, a_t, \{v7, r_t) \sim$ D
  $\theta_a \leftarrow \frac{1}{N} \sum (\min_{i=1,2} Q_\psi(s_t, a_t) +)$
  返回 π_θ a

以实现熵最大化，而 TD3 使用确定性策略。这意味着动作的每个维度都被建模为高斯分布 $N(\mu, \sigma)$，其中策略 $\pi(|s)$ 输出该分布的参数（μ和 σ）。SAC 也不需要目标策略平滑；在 TD3 中，目标策略平滑是指人为地向策略输出的动作注入噪声。这是因为最大熵目标和随机策略具有相同的正则化效果。算法 3 给出了 SAC 的策略和评论家更新过程。

在性能方面，SAC 以在更广泛的任务中表现更出色、更可靠而著称。例如，在大多数基准连续控制 Gym 任务中，SAC 和 TD3 的表现相近。然而，SAC 成功使人形机器人能够行走，而 TD3 在这个更复杂的环境中则以失败告终。因此，SAC 在强化学习社区中更常被用作许多算法和应用的默认基准与起点 [5, 19, 21]。

4.2 DroQ：评论家网络训练与正则化

尽管 SAC 仍然常被用作最先进的深度强化学习基线，但在其基础上也取得了一些改进，显著提升了样本效率。这些进展主要集中在改进评论家网络训练上。

第一个可以利用的洞见是在训练评论家网络时提高更新与数据比率（UTD） [5]。UTD比率指的是在环境中收集到一个新的状态转移（数据）时，评论家接受训练的程度。因此，它对应于每收集一个新的状态转移所执行的评论家梯度步数。为清晰起见，UTD仅指评论家，而不包括演员。因此，下文将UTD称为C-UTD（评论家UTD）。具体而言，SAC通常会在收集到一个新的状态转移后，以固定步数（通常为一步）同时训练演员和评论家。因此，提高C-UTD比率是指仅增加评论家步数，同时保持演员步数不变。当C-UTD为20时，每当一个新的状态转移进入回放缓冲区，评论家的训练次数将是演员的20倍。在算法1中，这对应于 G_{critic}。

然而，关键在于，必须将 C-UTD 比率的提高与某种网络正则化方法结合起来，以防止评论家因执行更多梯度更新而过拟合。REDQ [5] 通过使用大型 Q 网络集成来解决这一问题。然而，众所周知，这种方法计算开销很大，而且耗时很长。DroQ [19] 通过使用经随机失活和层归一化正则化的Q网络 Q_ψ 来降低计算成本，并以此作为正则化手段，保留高C-UTD 比率带来的优势。Smith等人 [31] 随后指出，在演员-评论家强化学习方法中加入正则化，是性能大幅提升的关键；而且可以采用多种正则化方法来获得同样的优势。

算法4： 广义演员-评论家 QD-RL 算法 - 高亮 可轻松 替换以构成新算法的模块化部分

输入： 代数 J，批量大小 B
初始化 , 论家 Q_ψ，演员 π_{θ_a}
对 $j = 1, ..., J$ **执行**
  对于 $b = 1, ..., B$，**执行**
    // 变体
    $\pi_{\theta_b} \leftarrow$
    如果 $b > B*$ ，**则**
      $\pi_{\theta_b} \leftarrow$ Θ)
    else
      对于 $g = 1, ...,$ ，**执行**
        $\pi_{\theta_b} \leftarrow$
    // 评估
    针对 $t, ..., T$ **do**
      $a_t \sim \pi_{\theta_b}(a_t|s_t)$
      $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
      $\tau_{\theta_b} \leftarrow (s_t, a_t, s_{t+1}, r_t)$
    $F(\pi_{\theta_b}) \; d(\pi_{\theta_b}) \leftarrow$评估($\tau_{\theta_b}$)
    // 更新数据整理器和容器
    D $\leftarrow$
    Θ $\leftarrow$
  // 更新评论家和演员
  对于 $g = 1, ...,$ **执行**
    对于 $g = 1, ...,$ **执行**
      $Q_\psi \leftarrow$
    对于 $g = 1, ...,$ **执行**
      $\pi_{\theta_a} \leftarrow$
返回 Θ

5 广义演员-评论家QD-RL

我们提出了广义演员-评论家 QD-RL，这是一种模块化框架，如图2和算法4所示。该框架提出了一种广义的演员-评论家 QD-RL方法视角，用于统一QD-RL设置中使用的离策略演员-评论家深度强化学习算法。它基于QD框架 [8]，该框架将QD算法定义为容器、选择器和变异的组合。图2中的每个组件框（算法4中以高亮线条标出 轻松替换的模块。广义演员-评论家 QD-RL引入了一种策略梯度（PG）变体，由以下模块组成：

- **数据整理器：** 存储和管理收集到的状态转移。
- **更新评论家：** 评论家的训练过程，可根据所用的深度强化学习算法进行调整。
- **更新演员：** 演员的训练过程，同样可根据所用的深度强化学习算法进行调整。

此外，我们还定义了以下参数，这些参数对于涵盖现有算法的全部范围至关重要：

- G: 并行回合评估结束时（每一代）的评论家和演员更新循环总数。

- G_{critic}：指定一次集体G更新中的评论家步数（对应于深度强化学习算法1中的 G_{critic}）。
- G_{actor}：指定一次集体G更新中的演员步数（对应于算法1中深度强化学习的 G_{actor}）。
- G_{PG}：指定为生成策略梯度后代而针对每个亲本执行的演员步数。

- p_{PG}：确定策略梯度变体占整个批量大小的比例。该值通常由变体 [26, 27]平均分配。

在以下各节中，我们首先说明选择这些组件的理由，然后阐述如何利用该框架创建新算法。最后，我们展示该框架如何帮助我们发现当前质量多样性强化学习训练实践中的一项局限。

5.1 组件模块的动机

我们主要通过考察深度强化学习与质量多样性强化学习设置之间的差异，来构建广义演员-评论家 QD-RL（见图 1）。此处我们重点介绍其中一些根本差异，以及框架中引入的各个组件和参数如何旨在提升泛化能力。

离策略行动者-评论家训练。 QD-RL 设置会在优化过程中评估大量不同的策略。因此，在 QD-RL 设置中无法采用 TRPO 或 PPO [29], 等同策略演员-评论家方法，因为它们只能利用当前演员收集的近期数据来更新策略。更新演员和更新评论家这两个模块，以及参数 G_{actor}和 G_{critic}，封装了常见的离策略演员-评论家深度强化学习算法。这些组件可用于适应损失函数和架构的变化。此外，它们还支持以统一方式实现延迟策略更新（通过更改 G_{critic}和G_{actor}）或不同深度强化学习算法的高 C-UTD 设置（通过修改 G_{critic}），例如 DroQ。

时间尺度：单步与回合式。 大多数深度强化学习算法通常以时间步为单位运行，这也称为单次转移。例如，在单策略深度强化学习设置中，SAC 和 TD3 算法会在环境运行一个或几个时间步后更新评论家和策略（见算法1，第9行）。另一方面，起源于进化计算的基于种群的质量-多样性算法会进行回合评估，仅在一个回合的完整轨迹展开结束后才更新评论家和策略（见算法4，第21行）。主要原因在于，质量多样性中使用的行为描述符和适应度值被定义为整个轨迹的函数。因此，对于长度为 T 的回合，传统深度强化学习算法至少会更新评论家和策略 T 次。相反，质量多样性强化学习只会在策略并行评估结束时更新网络，导致训练步数总量存在差异。

在我们的框架中引入参数 G，旨在通过控制每代之后的更新循环次数来弥补这一点。这使 G_{critic}和 G_{actor}的值能够保持不变，并遵循所选深度强化学习算法在每个时间步采用的相同训练流程。

数据分布：单一策略与策略种群。如图 1 所示，在质量多样性强化学习中，数据来自 B 个不同的策略，它们都表现出不同的行为。相反，在大多数深度强化学习方法中，状态转移由单个演员收集。因此，与深度强化学习相比，质量多样性强化学习中的回放缓冲区 D 数据分布存在显著差异，需要加以考虑。前一点进一步加剧了这一问题：在深度强化学习中，演员通常会在状态转移之间更新。这会使数据分布偏向奖励最高的演员策略。相较之下，在质量多样性强化学习中，经过评估的批次B 策略在整个回合期间保持不变，导致 D 中的数据分布变化更加剧烈。此外，质量多样性强化学习中的回放缓冲区 D 所含数据被替换的频率高于深度强化学习，并且仅经过几代就会被完全替换。

数据组织器模块可以通过所用回放缓冲区的类型（例如先进先出）和大小，以及对状态转移来源进行排序，来处理这一问题。它还支持灵活采用深度强化学习中的各种采样流程，例如重要性采样。

5.2 使用广义演员-评论家质量多样性强化学习整合深度强化学习的进展

借助广义演员-评论家框架，我们采用两种成熟且经验证的深度强化学习算法及其改进方法，这些方法不同于 TD3（详见第 4.1 节和第 4.2 节），即 SAC 和 DroQ，并提出 PGA-ME 的变体，分别称为 PGA-ME（SAC）和 PGA-ME（DroQ）。

5.2.1 PGA-ME（TD3）。通过在更新评论家和演员更新模块中采用 TD3 流程（见算法 2），即可恢复原始的 PGA 算法。G_{critic}设为 2，G_{actor}设为 1，以实现延迟策略更新；G 设为 150，从而恢复作者实现中的 300 个评论家步数 [26]。

5.2.2 PGA-ME（SAC）。PGA-ME（SAC）通过在算法 3 中用 SAC 更新过程替换 TD3 过程中的更新评论家和更新演员来实现。这包括新的损失函数和最大熵目标，同时也使用随机策略（见第 4.1 节）。

5.2.3 PGA-ME（DroQ）。由于 DroQ 基于 SAC 算法构建，因此可以使用 PGA（SAC）中的大多数相同构建模块。在 PGA-ME（DroQ）中，通过在每一代使用$G_{actor} = 1$和 $G_{critic} = C - UTD > 1$对评论家进行更多训练，实现高 C-UTD 设置。此外，评论家更新模块经过修改，采用随机失活和层归一化。

5.3 对质量多样性强化学习方法的影响

在本节中，我们依据深度强化学习与质量多样性强化学习设置之间差异的原则性考量（第 5.1 节），重点说明通用演员-评论家框架在研究当前质量多样性强化学习实践方面的优势。

对于演员-评论家质量多样性强化学习算法，$T \times B$ 个状态转移会由 B 种截然不同的策略在环境中收集，并在每次更新评论家 Q_ψ和演员π_{θ_a}（即每代）之间添加到回放缓冲区中。在广义演员-评论家框架中，参数 G 对应于 Q_ψ和π_{θ_a}的更新循环总数，并允许（1）对

回放缓冲区中的状态转移进行更多次采样，以及（2）对评论家和演员进行更多训练（即执行更多梯度步）

T的常见取值为 $1,000$ [14, 33]，而 B的常见取值为 128 [7, 26, 34]，这意味着在更新评论家和演员之前，大约需要进行 $\sim 100,000$ 次状态转移。在常见的深度强化学习设置中，这需要累计进行 $G \sim 50,000$ 次评论家和演员更新循环。相比之下，作者对 PGA-ME 的实现所选用的超参数值对应于 $G = 150$，小了好几个数量级。这表明，先前的质量多样性强化学习研究在很大程度上低估了更新总次数的价值 G。我们在接下来的实验中通过增加 G来研究这一差异的影响。这项研究不同于对 PGA-ME（DroQ）中高 C-UTD 的研究，因为 C-UTD 对应于每次收集的状态转移所进行的评论家更新次数，因此由 G_{critic}控制；而我们在此提出研究评论家和演员步骤累计总数 G的影响。

6 实验

为更好地理解质量多样性与深度强化学习算法之间的关系，我们通过实验研究广义演员-评论家框架所支持的方法及其带来的启示。根据前文的分析，我们旨在回答以下问题：

(1) 深度强化学习领域的进展能否应用于质量多样性强化学习？具体而言：• 通过 PGA-ME（SAC）实现最大熵强化学习（第 5.2.2 节）？• 通过 PGA-ME（DroQ - 正则化 + C-UTD）实现 C-UTD（第 5.2.3 节）？• 通过 PGA-ME（DroQ - 正则化）实现评论家正则化（第 5.2.3 节）？

(2) 我们的形式化分析所带来的直觉是否正确，即质量多样性强化学习中的行动者-评论家网络通常训练不足？
这些研究凸显了我们的框架对于将深度强化学习算法的未来进展应用于质量多样性强化学习的重要性。

6.1 实验设置

在本研究中，我们使用连续控制质量多样性强化学习基准任务 [13]，任务形式为在 Brax [14] 模拟器中运行半猎豹、蚂蚁和人形机器人（见图 3）。原始 PGA-ME 论文 [26]也使用了这些任务。不过，我们还加入了人形机器人任务，以评估所提出的方法。所有这些任务都是单向的，目标是找到多样化的步态策略 Θ，使其尽可能快速地向前移动。这一定义的灵感来自快速适应和损伤恢复 [7]方面的应用，并已证明在这些应用中有效。对于一个策略 $\theta \in \Theta$，适应度 $F(\theta)$和描述符$d(\theta)$定义如下：

$$d(\theta) = \frac{1}{T}\sum_{t}^{T}\begin{pmatrix} C_1^\theta(t) \\ \vdots \\ C_I^\theta(t) \end{pmatrix}, \text{where } I \text{ is the number of feet.} \quad (3)$$

$$F(\theta) = \sum_{t=0}^{T} r_{forward}(\theta) + r_{survive}(\theta) + (-r_{torque}(\theta)) \quad (4)$$

其中 C_i^θ表示一个布尔值：当脚 i 与地面接触时为 1，否则为 0。选择这一行为描述符可以产生多样的步态，因为它量化了行走时每条腿的平均使用情况。所有环境的回合长度均为 $T = 1000$。对于所有纳入考量的 PGA-ME 变体，我们使用



图 3：半猎豹、蚂蚁和人形机器人环境。

批量大小 B 为 128，并对每次运行进行 5 次重复试验。我们基于 Wilcoxon秩和检验计算所有指标的 p值，并采用 Bonferroni校正。

为了能够切实可行地实现大量梯度步数，以训练模型，我们使用 JAX [2], 这一机器学习框架。它支持即时编译（jit），可在 GPU 上执行操作以加快训练速度，并显著优化执行效率。我们使用 QDax 库 [23] 来实现这一点，并对演员-评论家模型的训练进行即时编译。

6.2 在质量多样性强化学习中整合强化学习领域的进展

在本节中，我们使用广义演员-评论家 QD-RL 框架进行实验，研究将深度强化学习中的新进展融入质量多样性强化学习的益处，具体包括熵正则化（SAC）、较高的 C-UTD（DroQ）和评论家正则化（DroQ）。在实验结果中，我们按照第 6.3 节的结果，采用了比以往文献更多的评论家更新次数，以最大限度地提升每种算法的性能。

为了更直观地了解在传统深度强化学习设定中，SAC 和 DroQ 相较于 TD3 所带来的性能提升，我们首先在相同环境中运行这三种深度强化学习算法，并且只考虑奖励。图 4 展示了这一比较，并印证了文献中的结果 [19, 31]。我们看到，SAC 在半猎豹和蚂蚁任务上的表现与 TD3 相当。在复杂的人形机器人任务中，TD3 完全失败，而 SAC 则成功实现了高性能。这一结果表明，动作空间探索与最大熵目标激励相结合，对于使人形机器人任务变得可行至关重要。

*6.2.1 PGA-ME（SAC）的优势——最大熵强化学习。*我们首先通过 PGA-ME（SAC）（第 5.2.2 节）评估最大熵目标在质量多样性强化学习设置中的有效性。图 5 展示了最大适应度和 QD 评分指标。最值得关注的是，我们发现 PGA-ME（SAC）能够在人形机器人任务中达到最高适应度；根据以往文献，这一适应度意味着任务已得到解决 [18]。在此之前，没有任何质量多样性强化学习算法（包括 PGA-ME（TD3））达到过这样的分数。深度强化学习领域的研究（图 4 中的实验也证实了这一点）已表明，SAC 能够解决人形机器人任务，而 TD3 无法做到；如今，仅通过更换框架中的强化学习算法，便在质量多样性强化学习设置中取得了相同结果，这充分显示了其中的潜力。与深度强化学习的结果一致，这也表明 SAC 的动作空间探索和最大熵目标使 PGA-ME（SAC）能够解决人形机器人任务。我们在此强调，只有对演员-评论家模型进行充分训练，才能取得这一结果；第 6.3 节将对此进行说明。

半猎豹的结果也遵循深度强化学习的趋势，PGA-ME（SAC）和 PGA-ME（TD3）的表现相当。

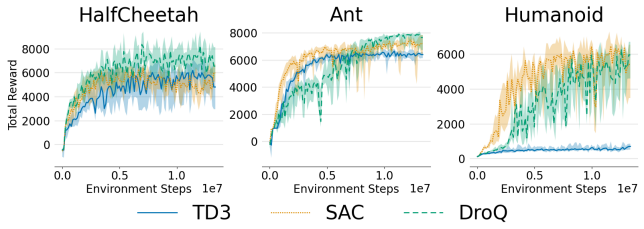

图 4：深度强化学习基线方法对比。粗线表示 10 次重复试验的中位数，阴影区域表示四分位数。

然而，在蚂蚁任务中，PGA-ME（SAC）在最大适应度和 QD 评分方面不如 PGA-ME（TD3）（$p < 5.10^{-5}$）。然而，将最大适应度与深度强化学习 TD3 基线的总奖励进行比较后可以发现，即使是 PGA-ME（TD3），其表现也未达到强化学习基线的水平。蚂蚁任务的描述符空间有 4 个维度，而人形机器人和半猎豹任务的描述符空间则有 2 个维度。我们推测，由于维度更高，档案中的解彼此差异更大。因此，在该任务中从深度强化学习切换到质量多样性强化学习设置时，数据分布偏移更大。此外，SAC 会促使策略表现出较高的熵，从而使数据的多样性更高。因此，SAC 中的更新评论家似乎比 TD3 中的更新评论家更容易受到这种偏移的影响。

6.2.2 PGA-ME（DroQ - 正则化）的局限性 + C-UTD）。此处，我们研究了集成到 PGA-ME（DroQ）中的更高 C-UTD 比率（见第 4.2 节）：DroQ 使用的 C-UTD 比率为 20，这意味着每收集一次状态转移后，评论家训练次数是 SAC 的 20 倍。在广义演员-评论家 QD-RL 中，这将使评论家更新次数增加 $G_{critic} = 20$，并达到 $G_{actor} = 1$，同时保持为 PGA-ME（SAC）选定的 G 值。具体而言，这意味着每代总共执行 $\sim 10^5$ 次评论家步数，而 PGA-ME（SAC）每代执行 $\sim 10^4$ 次评论家步数。

训练量的大幅增加有一个主要缺点：执行大量梯度步所需的时间。根据现有速度最快的开源DroQ实现（在实体机器人上可实时执行每秒 2400 个梯度步）[31]，，我们可以估算，在最佳情况下，执行 10_6 个评论家步需要在 GPU 上耗时 7 分钟。若进行 10_6次评估，批量大小为 128，则需要在 GPU 上训练近 900 小时，这显然不切实际。我们仍在图5中运行 PGA-ME（DroQ - 正则化 + C-UTD），并限制其训练时间。考虑到所需预算，性能提升似乎微乎其微（$p < 5.10^{-3}$）。这一维度首先揭示了一个有趣的观点：设置上的差异可能导致深度强化学习中的变更在质量多样性强化学习框架下难以切实应用。

6.2.3 PGA-ME（DroQ - 正则化）的局限性 DroQ 提出的第二个见解是通过随机失活和层归一化对评论家进行正则化。考虑到 PGA-ME（DroQ - 正则化 + C-UTD）的不可行性和无效性，我们将正则化与 C-UTD 比率分开进行评估。我们在图 5 中将这种方法命名为 PGA-ME（DroQ - 正则化）。

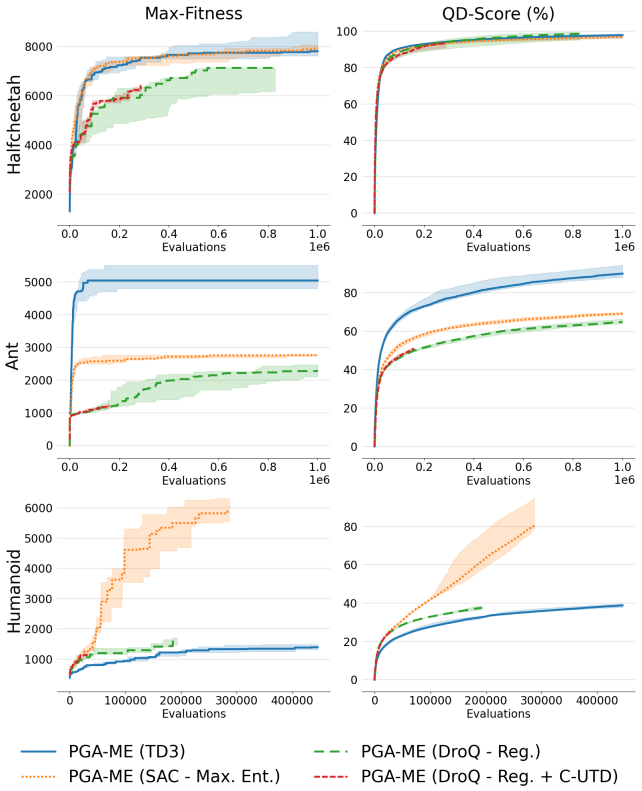

图5：采用所提出的广义演员-评论家 QD-RL 框架构建的 PGA 变体在使用 TD3、SAC 和 DroQ 强化学习算法时的性能。

我们观察到，在所有环境中使用 PGA-ME（DroQ - 正则化）（$p < 5.10^{-3}$）都会导致性能下降。这一结果再次表明，由于设置不同，深度强化学习中的见解未必能很好地迁移到质量多样性强化学习中。在 DroQ 中，评论家正则化旨在防止评论家因较高的 C-UTD 设置而过拟合。我们推测，这一见解无法有效迁移，是因为在质量多样性强化学习中，数据由大量不同的策略收集，而非单一策略。相比之下，尽管采用离策略设置，深度强化学习算法所观察到的转移仍高度偏向最新的演员。因此，由不同策略产生的更多样化数据分布不太容易导致过拟合。看来，在这一过程中进一步对评论家进行正则化会适得其反。

6.3 质量多样性强化学习中充分训练评论家和演员的重要性

在本节中，我们通过实验研究第 5.3 节基于原理的分析所得出的观察结果：在质量多样性强化学习中，演员-评论家模型通常训练不足。为此，我们研究了改变每个环境中表现最佳的 PGA-ME 变体的 G（更新循环次数）所产生的影响。

6.3.1 取值 G 的依据。首先，我们旨在估算 G的合理范围。给定批量大小 $B = 128$ 和回合长度

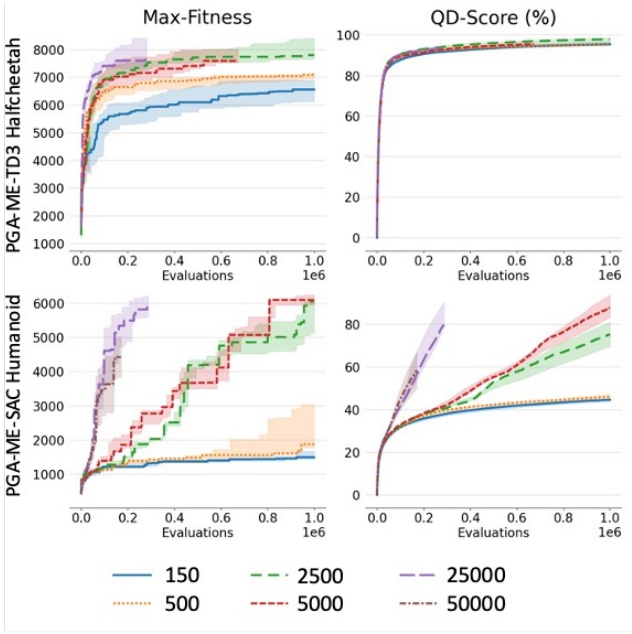

图 6：结果表明，在 PGA-ME-TD3 和 PGA-ME-SAC 中增加 G均会产生积极影响。

算法	PGA-TD3（半猎豹）		PGA-SAC（人形机器人）	
更新循环次数，G	150	2,500	150	25,000
演员添加	0.032	**0.041**	0.058	**0.28**
PG 添加	0.244	**0.377**	0.760	**2.38**

表 1：G值较小和较大时，演员和 PG 变体的平均添加效用；更新循环数。

$T = 1$，000 已使用，每一代会向回放缓冲区 D 添加 128，000 个状态转移。根据第 5.3 节讨论的原则性方法，如果遵循深度强化学习设置，即每向 D 添加一个状态转移就执行一次梯度步，那么 QD-RL 设置同样应在每一代后执行 128，000 个梯度步。我们研究了逐步增大 G 的情况，起点为 $G = 150$原作者的实现，该实现对应 300 个评论家步数；终点为 $G = 50$，000，对应 100，000 个评论家步数。

6.3.2 增加 G。尽管通过即时编译优化了代码，但随着每代所采用的更新循环次数G 增加，训练演员-评论家模型所需的时间也随之增加，进而延长了整个算法的运行时间。为便于实际应用，我们在每个算法运行 21 小时后停止运行。因此，梯度步数较多的变体无法达到评估预算次数。图 6 展示了这些实验的 QD评分和最大适应度曲线。我们观察到，增加更新循环次数 G能够提升 PGA-ME 的性能。PGA-ME（TD3）和 PGA-ME（SAC）均呈现这一趋势。我们还发现，PGA-ME（SAC）要解决人形机器人任务，必须有足够大的G。这表明演员-评论家模型的训练严重不足，因为最优值似乎至少比先前工作中使用的值（$p < 5.10^{-3}$）高出 10× 。此外，这种增加并不需要额外的环境评估，从而提高了样本效率。

6.3.3 解释增加 G 对梯度导向变体效用的影响。为进一步理解增加 G所带来的改进，我们在表 1 中展示了变体算子指标 [12]。该指标对应于算法中每种不同变体在每一代向档案中添加的个体数量。这里，我们给出了演员以及由梯度导向（PG）变异产生的后代的平均效用。我们观察到，增加 G时，这两种效用均显著提高。这表明，增加 G 会带来更好的演员和评论家。更好的演员可以通过变异 [12], 促进这一优秀个体的传播，而更好的评论家则可以改进 PG 变体，从而整体上提高质量多样性分数。

7 讨论与结论

本文提出了广义演员-评论家 QD-RL，这是一种模块化框架，旨在通过促进对 QD-RL 算法中深度强化学习进展的评估，推动 QD-RL 算法的发展。该框架的每个模块都可以轻松替换，以整合深度强化学习领域的新思想和新见解。因此，它是促进 QD-RL 进展的一种关键且高效的方式。我们通过提出 PGA-ME（SAC）和 PGA-ME（DroQ）这两种新算法展示了这种泛化能力，它们利用该框架将深度强化学习领域的进展融入 PGA-ME 原先采用的 TD3 算法中。PGA-ME（SAC）能够解决人形机器人环境中的任务，而现有的 QD-RL 方法无法做到这一点。然而，由于时间限制，PGA-ME（DroQ）中的高 C-UTD 设置在 QD-RL 中不太实用；此外，PGA-ME（DroQ）中的评论家正则化也难以从深度强化学习迁移到质量多样性领域。有趣的是，这表明并非深度强化学习领域的每一项进展都能迁移到质量多样性领域。此外，我们还发现，演员-评论家模型的训练步数对 QD-RL 算法的性能至关重要。我们的研究表明，以往工作中的这些模型训练不足，这可以通过分析深度强化学习与质量多样性设置之间的差异来解释。

总的来说，我们认为，广义演员-评论家 QD-RL 是理解质量多样性与深度强化学习之间强大协同效应并有效利用它们的重要一步。我们的研究表明，并非所有来自深度强化学习的见解都能有效迁移并应用于 QD-RL 设置。更重要的是，对这两种设置之间差异进行有原则的分析，可以解释这些见解为何有效或无效。

虽然我们发现，直接将某些方法（DroQ 正则化和 C-UTD）简单应用于质量多样性强化学习算法并不可行，但深度强化学习领域中考虑离策略和离线数据的其他见解也可以加以利用。本研究尚未探讨的另一项局限是所学习策略的可复现性 [12]。SAC 在评估期间使用随机策略，而 TD3 使用确定性策略，因此前者在不同评估中的表现可能有所不同。我们希望在未来的工作中进一步研究这些问题。

致谢

本研究得到了工程与物理科学研究理事会（EPSRC）项目 REcoVER 的资助，项目编号为 EP/V006673/1。

参考文献

[1] Ilge Akkaya、Marcin Andrychowicz、Maciek Chociej、Mateusz Litwin、Bob Mc- Grew、Arthur Petron、Alex Paino、Matthias Plappert、Glenn Powell、Raphael Ribas 等。2019年。使用机器人手求解魔方。arXiv预印本 arXiv:1910.07113（2019年）。[2] James Bradbury、Roy Frostig、Peter Ha wkins、Matthew James Johnson、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、Skye Wanderman-Milne 和 Qiao Zhang。2018年。JAX：用于 Python+NumPy 程序的可组合变换。http://github.com/google/jax[3] Felix Chalumeau、Raphael Boige、Bryan Lim、Valentin Macé、Maxime Allard、Arthur Flajolet、Antoine Cully 和 Thomas Pierrot。2022年。神经进化是技能发现中强化学习的一种有竞争力的替代方案。arXiv预印本 arXiv:2210.03516（2022年）。[4] Konstantinos Chatzilygeroudis、Antoine Cully、Vassilis Vassiliades 和 Jean- Baptiste Mouret。2021年。质量-多样性优化：随机优化的一个新分支。载于《黑箱优化、机器学习与无免费午餐定理》。施普林格，109–135。[5] Xinyue Chen、Che Wang、Zijian Zhou 和 Keith Ross。2021年。随机集成双重Q学习：无需模型即可快速学习。arXiv预印本 arXiv:2101.05982（2021年）。[6] Cédric Colas、Vashisht Madhavan、Joost Huizinga 和 Jeff Clune。2020年。将 MAP-Elites 扩展至深度神经进化。载于《2020年遗传与进化计算会议论文集》。67–75。[7] Antoine Cully、Jeff Clune、Danesh Tarapore 和 Jean-Baptiste Mouret。2015年。能够像动物一样适应的机器人。《Nature》521卷，7553期（2015年），503–507。[8] Antoine Cully 和 Yiannis Demiris。2017年。质量与多样性优化：一个统一的模块化框架。《IEEE进化计算汇刊》22卷，2期（2017年），245–259。[9] Jonas Degrave、Federico Felici、Jonas Buchli、Michael Neunert、Brendan Tracey、Francesco Carpanese、Timo Ewalds、Roland Hafner、Abbas Abdolmaleki、Diego de Las Casas 等。2022年。通过深度强化学习控制托卡马克等离子体。《Nature》602卷，7897期（2022年），414–419。[10] Adrien Ecoffet、Joost Huizinga、Joel Lehman、Kenneth O Stanley 和 Jeff Clune。2021年。先获取首次回报，再进行探索。《Nature》590卷，7847期（2021年），580–586。[11] Alhussein Fawzi、Matej Balog、Aja Huang、Thomas Hubert、Bernardino Romera- Paredes、Mohammadamin Barekatain、Alexander Novikov、Francisco J R Ruiz、Julian Schrittwieser、Grzegorz Swirszcz 等。2022年。利用强化学习发现更快的矩阵乘法算法。《Nature》610卷，7930期（2022年），47–53。[12] Manon Flageat、Felix Chalumeau 和 Antoine Cully。2022年。不确定领域中用于神经进化的 PGA-MAP-Elites 实证分析。《ACM进化学习汇刊》（2022年）。[13] Manon Flageat、Bryan Lim、Luca Grillotti、Maxime Allard、Simón C. Smith 和 Antoine Cully。2022年。对用于强化学习神经进化的质量-多样性算法进行基准测试。https://doi.org/10 48550 2211 02193 . /[14] ARXIV. . C. Daniel Freeman、Erik Frey、Anton Raichuk、Sertan Girgin、Igor Mordatch 和 Olivier Bachem。2021年。Brax——用于大规模刚体模拟的可微物理引擎。http://github.com/google/brax [15] Scott Fujimoto、Herke Hoof 和 David Meger。2018年。解决演员-评论家方法中的函数近似误差。载于《国际机器学习大会》。PMLR，1587–1596。[16] Adam Gaier、Alexander Asteroth 和 Jean-Baptiste Mouret。2018年。通过代理辅助照明实现数据高效的设计探索。《进化计算》26卷，3期（2018年），381–410。[17] Daniele Gravina、Ahmed Khalifa、Antonios Liapis、Julian Togelius 和 Georgios N Yannakakis。2019年。通过质量多样性进行程序化内容生成。载于《2019年IEEE游戏会议（CoG）》。IEEE，1–8。[18] Tuo mas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。2018年。软演员-评论家：采用随机演员的离策略最大熵深度强化学习。载于《国际机器学习大会》。PMLR，1861–1870。[19] Takuya Hiraoka、Takahisa Imagawa、Taisei Hashimoto、Takashi Onishi 和 Yoshi- masa Tsuruoka。2021年。用于双重高效强化学习的 Dropout Q函数。arXiv预印本 arXiv:2110.02034（2021年）。[20] Yibin Jiang、Daniel Salley、Abhishek Sharma、Graham Keenan、Margaret Mullin 和 Leroy Cronin。2022年。一种由人工智能驱动的化学合成机器人，用于纳米材料的探索与优化。《Science Advances》8卷，40期（2022年），eabo2626。https://doi.org/10.1126/sciadv.abo2626 arXiv:https://www.science.org/doi/pdf/10 1126 2626 . /sciadv.abo Aviral[21] Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。2020年。面向离线强化学习的保守Q学习。《神经信息处理进展》处理系统 33 (2020), 1179–1191.[22] Joonho Lee、Jemin Hwangbo、Lorenz Wellhausen、Vladlen Koltun 和 Marco Hutter。2020。学习在复杂地形上进行四足运动。*Science robotics* 5, 47 (2020), eabc5986。[23] Bryan Lim、Maxime Allard、Luca Grillotti 和 Antoine Cully。2022。通过大规模并行加速质量-多样性以应用于机器人学。*arXiv* 预印本 *arXiv:2202.01258* (2022)。[24] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Bellemare、Alex Graves、Martin Riedmiller、Andreas K Fidjeland、Georg Ostrovski 等。2015。通过深度强化学习实现人类水平的控制。*Nature* 518, 7540 (2015), 529–533。[25] Jean-Baptiste Mouret 和 Jeff Clune。2015。通过映射精英来探索搜索空间。*arXiv* 预印本 *arXiv:1504.04909* (2015)。[26] Olle Nilsson 和 Antoine Cully。2021。策略梯度辅助 MAP-Elites。载于遗传与进化计算会议论文集。866–875。[27] Thomas Pierrot、Valentin Macé、Felix Chalumeau、Arthur Flajolet、Geoffrey Cideron、Karim Beguir、Antoine Cully、Olivier Sigaud 和 Nicolas Perrin-Gilbert。2022。用于样本高效质量-多样性优化的多样性策略梯度。载于遗传与进化计算会议论文集。1075–1083。[28] Justin K Pugh、Lisa B Soros 和 Kenneth O Stanley。2016。质量-多样性：进化计算的新前沿。*Frontiers in Robotics and AI* 3 (2016), 40。[29] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347* (2017)。[30] David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershel- vam、Marc Lanctot 等。2016。利用深度神经网络和树搜索掌握围棋。*Nature* 529, 7587 (2016), 484–489。[31] Laura Smith、Ilya Kostrikov 和 Sergey Levine。2022。漫步公园：通过无模型强化学习在20分钟内学会行走。*arXiv* 预印本 *arXiv:2208.07860* (2022)。[32] Richard S Sutton 和 Andrew G Barto。2018。强化学习：导论。MIT出版社。[33] Yuval Tassa、Yotam Doron、Alistair Muldal、Tom Erez、Yazhe Li、Diego de Las Casas、David Budden、Abbas Abdolmaleki、Josh Merel、Andrew Lefrancq 等。2018。DeepMind 控制套件。*arXiv* 预印本 *arXiv:1801.00690* (2018)。[34] Bryon Tjanaka、Matthew C Fontaine、Julian Togelius 和 Stefanos Nikolaidis。2022。近似梯度以实现强化学习中的可微质量多样性。*arXiv* 预印本 *arXiv:2202.03666* (2022)。[35] Jonas Verhellen 和 Jeriek Van den Abeele。2020。探索化学空间中的精英片区。*Chemical science* 11, 42 (2020), 11485–11491。[36] Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Michaël Mathieu、Andrew Dudzik、Junyoung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等。2019。利用多智能体强化学习达到《星际争霸II》宗师级水平。*Nature* 575, 7782 (2019), 350–354。[37] Rui Wang、Joel Lehman、Jeff Clune 和 Kenneth O Stanley。2019。POET：环境及其优化解的开放式协同进化。载于遗传与进化计算会议论文集。142–151。[38] Rui Wang、Joel Lehman、Aditya Rawal、Jiale Zhi、Yulun Li、Jeffrey Clune 和 Kenneth Stanley。2020。增强版 POET：通过不受限制地发明学习挑战及其解实现开放式强化学习。载于国际机器学习大会。PMLR, 9940–9951。

A 附录

A.1 补充结果

A.1.1 变体。 表 3 展示了增加参数 G时的完整变体效用。

算法 更新循环次数， G	PGA-TD3（半猎豹） 150	500	2500	5000	25000
演员添加	0.032	0.034	0.405	0.509	0.917
PG 添加	0.244	0.247	0.377	0.432	0.746
GA添加	2.26	2.03	1.99	2.91	5.20

表2：在蚂蚁任务中，演员、PG 变体和遗传算法变体在 G取小值和大值时的平均添加效用；更新循环次数。

算法 更新循环次数， G	PGA-SAC（人形机器人） 150	500	2500	5000	25000	50000
演员添加	0.058	0.062	0.190	0.173	0.29	0.24
PG 添加	0.760	0.765	0.840	0.866	2.38	3.48
GA添加	1.22	1.20	1.45	1.52	3.90	5.23

表 3：随着人形机器人任务的更新循环次数 G增加，演员、PG 变体和遗传算法变体的平均添加效用

A.1.2 档案可视化。 图 7 展示了在我们的框架中启用不同深度强化学习基线算法时，各种 PGA-ME 变体对应的精英档案。

A.2 实验与实现细节

A.2.1 超参数。 表4展示了所使用的不同 PGA-ME 变体的超参数列表。行块用于区分特定算法所使用的超参数。

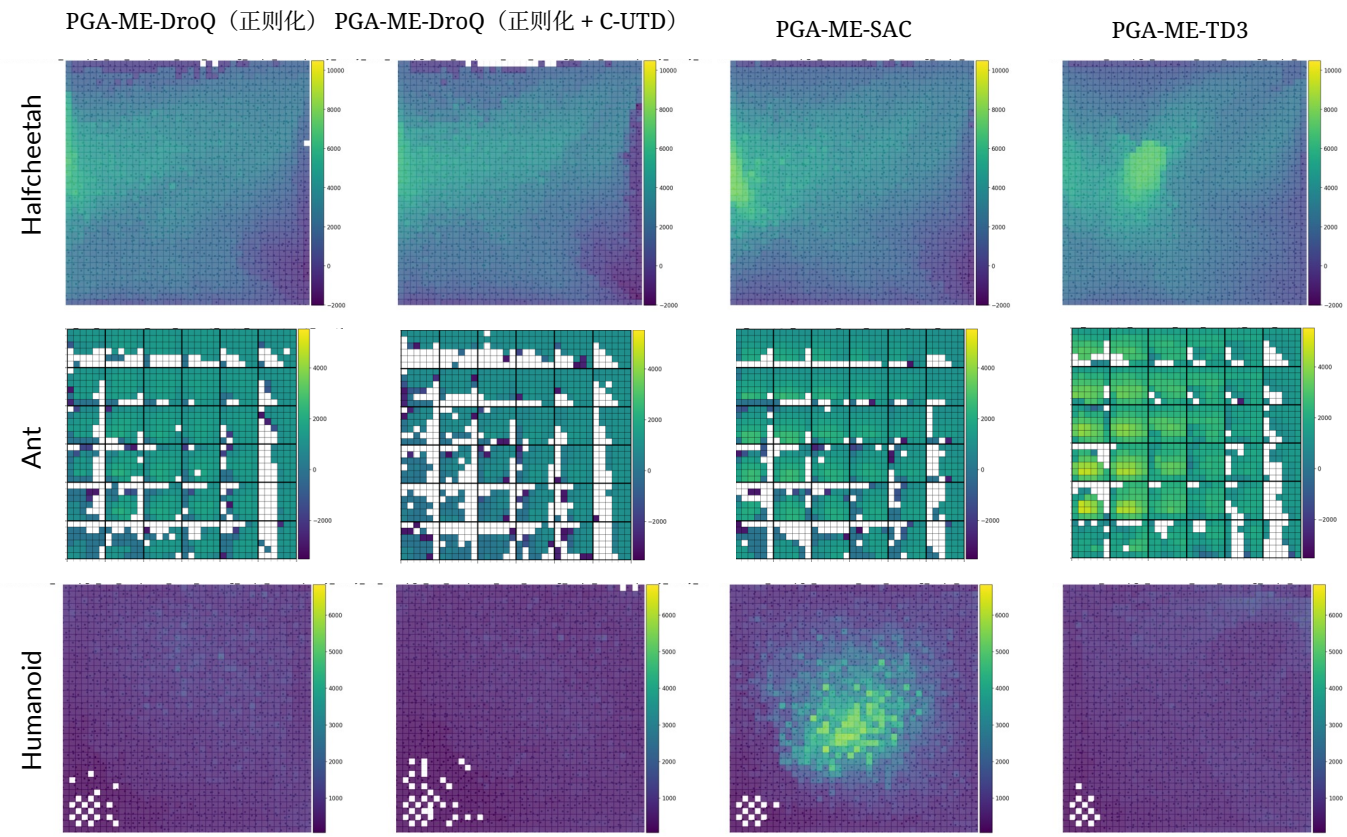

图 7：档案库　可视化了在不同任务中考察的 PGA-ME 变体（TD3、SAC、DroQ）　s.

超参数	PGA-ME-TD3	PGA-ME-SAC	PGA-ME-DroQ
策略隐藏层大小	[64, 64]	[64, 64]	[64, 64]
批量大小, B	128	128	128
Iso 系数，σ_1	0.01	0.01	0.01
线系数，σ_2	0.1	0.1	0.1
策略梯度比例 p_{PG}	0.5	0.5	0.5
策略梯度训练步数 G_{PG}	100	100	100
PGA经验回放缓冲区大小	10^6	10^6	10^6
评论家网络隐藏层大小	[256, 256]	[256, 256]	[256, 256]
评论家网络学习率	0.0003	0.0003	0.0003
贪心学习率	0.0003	0.0003	0.0003
策略梯度变体学习率	0.001	0.001	0.001
状态转移批次大小	256	256	256
噪声裁剪	0.5	-	-
策略噪声	0.2	-	-
软 τ 更新	0.005	-	-
熵系数初始化 α_{init}	-	1.0	1.0
τ 更新	-	0.005	0.005
丢弃率	-	-	0.01

表 4：PGA-ME 变体的超参数——TD3、SAC 和 DroQ。