arXiv:2601.00693v1 [cs.LG] 2 Jan 2026

ARISE：融合群体探索的自适应强化学习

Rajiv Chaitanya M 和 D R Ramesh Babu

[1] 本科生，计算机科学与工程系，Dayananda Sagar 工程学院，印度班加罗尔 rajiv.muttur@gmail.com
https://orcid.org/0009-0007-7159-5610[2] 教授，计算机科学与工程系，Dayananda Sagar 工程学院，印度班加罗尔

摘要。有效探索仍是强化学习面临的一项关键挑战，尤其是在奖励非平稳或策略维度较高的情况下。我们提出 **ARISE**，这是一个轻量级框架，通过在标准策略梯度方法中加入精简的基于群体的探索层来增强强化学习。ARISE 将策略动作与粒子驱动的提议相结合，其中每个粒子代表在动作空间中采样的候选策略轨迹，并利用奖励方差线索自适应地调节探索。对于简单的基准任务，性能提升较小（例如，在 CartPole-v1 上提升 $+\mathbf{0.7\%}$）；而在更具挑战性的任务中，ARISE 则带来显著提升，包括在 LunarLander-v3 上提升 $+\mathbf{46\%}$，以及在 Hopper-v4 上提升 $+\mathbf{22\%}$，同时在 Walker2d 和 Ant 上保持稳定。在奖励发生非平稳偏移时，ARISE 展现出显著的鲁棒性优势，在 CartPole 上以 $+\mathbf{75分}$ 的优势超越近端策略优化，并相应提升了 LunarLander 的表现。消融研究证实，群体组件和自适应机制均有助于提升性能。总体而言，ARISE 提供了一种简单且与架构无关的途径，无需改变核心算法结构，即可使强化学习智能体具备更强的探索能力和韧性。

关键词：强化学习、策略优化、群体智能、自适应算法、非平稳环境

1 引言

强化学习（RL）在控制、机器人学和决策领域取得了显著进展 [17,9]。尽管函数逼近方面已有进步，探索仍然是一项限制。标准方法，如 ϵ-贪心探索、熵奖励、内在好奇心和基于计数的方法 [1,11,15]，往往在奖励稀疏、具有欺骗性或呈多模态结构的环境中失效。因此，单智能体强化学习流程常常收敛到次优的局部最优解，无法发现更广泛的解结构。

基于种群的进化强化学习方法试图通过分布式探索来应对这些挑战。进化策略（ES） [14]，MAP-Elites、质量-多样性算法 [2,10],以及基于种群的训练（PBT） [5] 表明，保持行为多样性至关重要。然而，这些方法通常会带来巨大的计算开销，或依赖进化更新而非基于梯度的改进。即使是多智能体强化学习框架，往往也缺乏能够让智能体在保持多样性的同时协调探索的机制。

我们提出了**ARISE**，即自适应强化学习与群体探索集成，这是一种轻量级的基于梯度的强化学习框架，利用多个智能体探索不同的策略区域，并辅以选择性信息共享和自适应协调。ARISE 建立在三种核心机制之上。

1. **新颖性驱动的探索**：每个智能体都会根据策略空间或轨迹空间偏差获得新颖性奖励，从而鼓励多样化行为并防止过早同质化。2.**知识广播**：表现出色的智能体会定期传播其策略参数，接收方则通过受控的混合系数混合这些权重，从而实现信息共享，同时避免模式崩溃。3. **基于方差的自适应协调**：群体根据智能体间性能方差确定协调温度，并据此动态调整探索与利用之间的平衡。

与进化方法或简单的基于种群的方法相比，ARISE 在保持基于梯度的演员-评论家架构样本效率的同时，还获得了分布式探索所带来的行为鲁棒性和多样性。该方法仅对架构进行少量修改，就能轻松与常见的连续控制强化学习基线协同运行，例如近端策略优化 [15] 和软演员-评论家 [3]。

总而言之，我们的贡献如下：
– 提出了一种基于群体的强化学习新型统一框架，高效整合了传统与前沿的协调方法，以及新颖性激励和多智能体探索。– 我们开发的动态广播机制在实现有用策略信息快速传播的同时，能够保持多样性。–基于标准连续控制基准的实证结果表明，ARISE 不仅提升了样本效率，还凭借在奖励欺骗场景中的鲁棒性确保了最终性能，并超越了强大的基线。

ARISE 证明，可扩展的协同探索并不需要漫长的进化过程或复杂的种群控制方法。量身定制地结合多样性、通信与适应，可以显著提高基于梯度的强化学习系统探索的全面性。

2 相关工作

探索一直是强化学习面临的根本挑战。经典方法依赖随机动作选择、基于价值的不确定性或内在奖励，鼓励智能体访问未访问状态。诸如 ϵ-贪心探索、熵正则化和基于计数的探索 [1,16] 等方法提供了简单而有效的启发式策略，但在奖励稀疏或训练早期由次优吸引子主导时，往往会失效。好奇心驱动方法，包括基于预测误差的内在动机[11]和信息增益公式 [4], ，通过奖励智能体降低不确定性来改进探索。然而，这些方法在高维观测空间中可能不稳定，还会出现与任务成功无关的探索漂移。

2.1 基于种群与进化强化学习

基于种群的方法通过维护多个随时间演化的智能体来应对探索问题。进化策略（ES） [14] 采用无梯度黑盒优化，大规模探索策略空间；而 MAP-Elites 和质量—多样性（QD）算法 [2,10] 则显式搜索多样化的行为档案。这些方法具有强大的全局探索能力，但会牺牲精细的梯度更新，且往往产生高昂的计算成本。基于种群的训练（PBT） [5] 通过定期用更强的智能体替换部分较弱的智能体，改进了超参数适应；但它依赖进化选择压力，若不仔细调优，多样性可能会迅速丧失。

近期的混合方法将策略梯度与进化概念结合起来。例如，ERL 和 CERL 框架 [7,12] 利用进化种群来增强基于梯度的学习器。尽管这些系统效果显著，但往往采用复杂的消息传递结构或辅助流程，增加了不必要的系统复杂性和计算开销。

2.2 多智能体与群体启发方法

多智能体强化学习框架通过引入智能体之间的协调来提升探索能力。去中心化和集中式训练范式（例如 MADDPG[8]和 QMIX [13]）能够提升合作学习和竞争学习的效果，但主要聚焦于多智能体任务，而非将多智能体结构用作探索工具。基于群体的优化方法（例如粒子群优化 [6], ）使用简单的通信规则，在种群中传播有用信息，展现了分布式局部更新的优势。不过，这些方法很少与基于梯度的强化学习直接结合。

2.3 ARISE 的定位

ARISE 位于基于梯度的强化学习、种群多样性和自适应信息共享的交汇处。与进化方法不同，ARISE 保留了完整的基于反向传播的优化。与传统的基于种群的训练不同，ARISE 不依赖替换或进化选择，因此种群能够保持稳定的多样性。最后，与标准的多智能体强化学习不同，ARISE 并不聚焦于多智能体环境；相反，它将多个智能体用作在单个任务中进行结构化探索的机制。

通过融合新颖性驱动行为、受控策略广播和基于方差的自适应协调，ARISE 提供了一种轻量而强大的进化式和基于种群的强化学习替代方案，在拓展探索范围的同时保持样本效率。

3 方法

ARISE（自适应强化学习与群体探索集成）是一种混合学习框架，将同策略强化学习与动作空间中的粒子群探索相结合。该系统由 M 个并行演员-评论家智能体组成，每个智能体作为一个群体粒子运行，并配有自己的策略、记忆缓冲区和适应度信号。与传统的多智能体近端策略优化不同，ARISE 在推理时注入基于粒子群优化的探索性动作，利用根据粒子间距离计算的新颖性指标增强奖励信号，并定期广播全局最优策略，以协调整个群体的利用。由此形成的机制将策略梯度学习与种群驱动搜索相结合，在保持同策略更新稳定性的同时，实现高多样性的探索。

3.1 智能体架构

每个智能体 $i \in \{1, \ldots, M\}$ 都维护一个独立的演员-评论家策略 $\pi_{\theta_i}(a \mid s)$ 和价值函数 $V_{\theta_i}(s)$，它们由共享主干网络 ϕ 后接智能体专属头部生成：

$$h_i = \phi(s), \qquad \pi_{\theta_i}(a \mid s) = \mathrm{Actor}(h_i), \qquad V_{\theta_i}(s) = \mathrm{Critic}(h_i).$$

演员会为连续控制生成高斯动作分布，或为离散任务生成分类 logits。智能体将状态转移存储在各自的近端策略优化缓冲区中，其中包含 $(s_t,\ a_t,\ r_t,\ d_t,\ \log \pi_{\theta_i}(a_t \mid s_t)\ V_{\theta_i}(s_t))$ 以及对应的智能体标识符。

动作选择采用偏置采样规则，优先选择表现出色的智能体：

$$i^* \sim \begin{cases} \text{best agent} & 0.70, \\ \text{second best} & 0.20, \\ \text{uniform over } \{1, \ldots, M\} & 0.10. \end{cases}$$

这种偏置在保持多样性的同时，融入了自然的进化压力。

3.2 粒子群动作混合

与环境交互时，每个智能体都会生成两个动作提议：

1. 强化学习动作 $a_i^{\mathrm{RL}}, t \sim \pi_{\theta_i}(\cdot \mid s_t)$；2. 粒子群优化动作 a_i^{PSO}, t 根据粒子在动作空间中的位置和速度计算得出。

对于连续动作，ARISE 使用线性混合规则：

$$a_{i,t} = (1-\alpha)a_{i,t}^{\mathrm{RL}} + \alpha\, a_{i,t}^{\mathrm{PSO}},$$

其中 $\alpha \in [0,1]$ 为群体混合系数（经验值为 $\alpha = 0.12$）。对于离散策略，ARISE 以概率 α 切换到粒子群优化动作，否则遵循强化学习策略采样结果。

这种机制使群体能够探索仅靠梯度驱动更新无法到达的动作空间盆地。

3.3 新颖性驱动的奖励增强

为鼓励分布式探索，ARISE 引入了一个新颖性项，其计算方式是智能体的动作提议与群体中各粒子位置之间的最小欧几里得距离：

$$n_{i,t} = \tanh\left(\min_{j \neq i} \|a_{i,t} - p_j\|\right),$$

其中，p_j 表示粒子 j 的位置。内在奖励通过以下方式并入环境奖励

$$r_{i,t}^{\mathrm{aug}} = r_{i,t} + \beta n_{i,t},$$

其中新颖性系数为 $\beta = 0.01$。新颖性也被纳入群体的适应度信号：

$$F_i = \bar{r}_i + \bar{n}_i,$$

从而将探索性能与粒子群优化更新动态联系起来。

3.4 近端策略优化的在线策略优化

每个智能体仅使用自身的轨迹数据进行近端策略优化更新。给定增强奖励后，ARISE 计算广义优势估计优势：

$$A_{i,t} = \sum_{l=0}^{T-t} (\gamma\lambda)^l \big(r_{i,t+l}^{\mathrm{aug}} + \gamma V_{\theta_i}(s_{t+l+1}) - V_{\theta_i}(s_{t+l})\big),$$

并返回：

$$R_{i,t} = A_{i,t} + V_{\theta_i}(s_t).$$

近端策略优化（PPO）的替代目标为

$$\mathcal{L}_i^{\mathrm{PPO}} = \mathbb{E}_t \left[\min \left(\rho_{i,t} A_{i,t}, \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) A_{i,t} \right) \right],$$

比率为

$$\rho_{i,t} = \frac{\pi_{\theta_i}(a_{i,t} \mid s_{i,t})}{\pi_{\theta_i}^{\mathrm{old}}(a_{i,t} \mid s_{i,t})}.$$

完整的优化目标结合了策略、价值和熵项：

$$\mathcal{J}_i = \mathcal{L}_i^{\mathrm{PPO}} - \eta\, H(\pi_{\theta_i}) + c_v \left\| R_{i,t} - V_{\theta_i}(s_t) \right\|^2.$$

每次更新后，ARISE都会记录平均回报、策略熵和适应度，并更新每个粒子的个体最优（$p$best），以参与全局最优（$g$best）的更新。

3.5 性能方差自适应粒子群优化更新

粒子群优化模块为每个粒子维护速度和位置向量：

$$v_i \leftarrow w v_i + c_1 r_1 (p\mathrm{best}_i - p_i) + c_2 r_2 (g\mathrm{best} - p_i), \qquad p_i \leftarrow p_i + v_i,$$

其中惯性为 w，认知系数为 c_1，社会系数为 c_2，其初始化方式如下

$$w = 0.7, \qquad c_1 = 1.5, \qquad c_2 = 1.5.$$

ARISE 根据奖励方差调整这些参数：

$$\mathrm{VarReward} = \mathrm{Var}\left(\{F_i\}_{i=1}^{M} \right).$$

方差较高表示利用不足，因此会增加 c_2 （社会项）；方差较低则会增加 c_1，以加强自主优化。惯性权重 w 会随着训练进度逐渐衰减，从而稳步降低探索速度。

3.6 利用的策略广播

每轮近端策略优化（PPO）周期结束后，ARISE 会将表现最佳的全局策略参数广播给整个群体：

$$\theta_i \leftarrow \theta_{g^*}, \qquad \forall i \in \{1, \ldots, M\}.$$

这一同步利用步骤确保快速收敛，同时允许后续由粒子群优化驱动的混合回合重新生成行为多样性。广播式利用 与 粒子群优化介导的探索 交替进行，构成了 ARISE 的核心自适应动态。

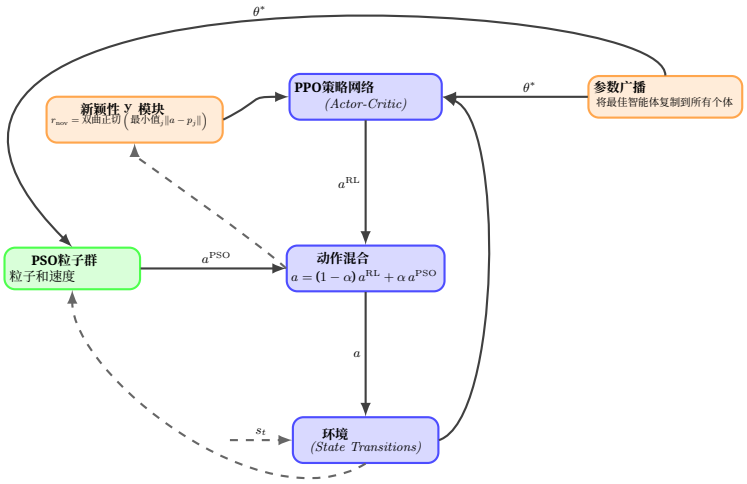


图 1. ARISE 框架的架构。该系统将基于近端策略优化的策略网络与 PSO 粒子群优化相结合。实线箭头表示主要数据流；虚线箭头表示辅助信号。

3.7 训练流程概述

在每次训练迭代中，ARISE 都会执行：

1. 策略与粒子群优化混合的动作采样；2. 新颖性增强的奖励收集；
3. 逐智能体近端策略优化；4. 适应度评估和粒子群优化粒子更新；
5. 全局策略广播；6. 基于性能方差的自适应参数调度。

这一循环反馈机制使 ARISE 能够利用基于梯度的学习，同时持续引入群体层面的多样性，从而产生稳定且高度探索性的学习动态。

4 实验设置

4.1 环境

所有实验均在标准化的连续控制和离散控制基准上进行，以评估基线智能体和 ARISE 智能体：

– **连续控制：** MuJoCo 运动任务（HalfCheetah-v4、Ant-v4）– **离散控制：** OpenAI Gym 经典控制任务（CartPole-v1、LunarLander-v2）

这些环境在探索、稀疏奖励和高维动作空间方面带来了各种挑战。

4.2 基线

为确保比较公平，我们将 ARISE 与三种成熟的强化学习基线进行评估：

1. **近端策略优化：**基于策略的优化，采用优势函数进行更新；2. **深度Q网络：**基于价值的迭代，采用离策略学习和经验回放；3. **优势演员-评论家：**同步多步更新；

所有基线均采用一致的网络架构、超参数和训练方案实现，以确保结果可复现。

4.3 ARISE 配置

ARISE智能体将逐粒子PPO策略与受粒子群优化启发的群体模块相结合。关键参数：

– 群体规模：3 个粒子– 强化学习–粒子群优化动作混合系数： $\alpha = 0.12$– 新颖性系数：0.01– PPO 特定参数： $\gamma = 0.99$、 $\lambda = 0.95$、裁剪 $\epsilon = 0.2$、熵系数 $= 0.01$–训练时域：连续环境每次更新 512 步，离散环境 2048 步– 自适应粒子群优化参数：惯性 $w = 0.7$、认知/社会系数 $c_1 = c_2 = 1.5$

策略架构：演员和评论家均采用带双曲正切激活函数的 2 层多层感知机；正交权重初始化。

4.4 训练方案

– 运行次数：每个环境使用 5 个独立随机种子**– 评估频率：**
每 50 个训练回合**– 指标：**

- 平均回合回报（适应度）
- 粒子间的奖励方差
- 群体多样性（动作空间中的平均两两距离）
- 收敛速度（达到最大回报 90% 所需的回合数）**– 计算设置：**NVIDIA A100 GPU；固定 CPU线程以确保可复现性

4.5 消融研究

为验证 ARISE 的各个组件，我们进行了有针对性的消融实验：

1. **无粒子群优化：** 仅强化学习的PPO智能体 2. **无新颖性奖励：** 标准PPO与群体结合，但没有探索奖励 3.**无广播：** 粒子独立更新，不进行最佳智能体复制

5 实验结果

我们在不同的经典控制、连续控制和非平稳任务中评估 ARISE 框架，并将其与标准强化学习基线进行比较，例如近端策略优化、深度Q网络和优势演员-评论家。指标包括平均回合回报，反映性能和学习稳定性。

5.1 经典控制任务

表 1 展示了 CartPole-v1 和 LunarLander-v3 的性能。ARISE 在这两种环境中均取得了更高的回报，展现出稳定的策略优化能力和有效的探索能力。

表 1. 经典控制任务上的性能（平均回合回报）

环境	ARISE	近端策略优化	深度Q网络	优势演员-评论家
CartPole-v1	333.25	330.96	130.53	64.39
LunarLander-v3	241.51	165.10	220.77	8.72

5.2 连续控制任务

表 2 总结了 MuJoCo 环境上的结果。ARISE 始终取得具有竞争力或更优的回报，尤其是在 Hopper-v4 和 Ant-v4 上，这凸显了强化学习–粒子群优化混合探索在高维连续动作空间中的优势。

表 2。连续控制任务上的性能（平均回合回报）

环境	ARISE	近端策略优化	深度Q网络	优势演员-评论家
Hopper-v4	279.44	229.10	–	69.53
Walker2d-v4	333.27	337.54	–	51.75
Ant-v4	-108.40	-143.49	–	-1594.45
MountainCarCont-v0	-1.06	-1.61	–	-0.33

5.3 非平稳任务

为评估奖励偏移下的鲁棒性，我们修改了 CartPole-v1 和 LunarLander-v3 环境。与近端策略优化和优势演员-评论家相比，ARISE 展现出更快的适应速度和更高的回报（表3），证实了自适应探索的优势。

表 3. 非平稳任务上的表现（平均回合回报）

环境（奖励偏移）	ARISE	近端策略优化	优势演员-评论家
CartPole-v1	460.94	385.56	313.59
LunarLander-v3	248.38	221.12	70.69

5.4 消融研究

为单独评估 ARISE 关键组件的贡献，我们进行了消融实验：移除自适应参数调优（ARISE NO ADAPTIVE）并禁用 PSO粒子群 -（ARISE NO SWARM）。表4中的结果表明，这两个要素都对性能有显著贡献；完整的 ARISE 配置始终优于所有消融变体和近端策略优化基线。

表4. 消融研究结果（平均回合回报）

环境	ARISE	ARISE NO ADAPTIVE	ARISE NO SWARM	近端策略优化
CartPole-v1	333.25	323.25	309.87	330.96
Hopper-v4	279.44	259.44	258.23	229.10

6 讨论

实验结果表明，在多种经典控制、连续控制和非平稳任务中，ARISE 的表现始终优于标准强化学习基线（近端策略优化、深度Q网络和优势演员-评论家）。在 CartPole-v1 等经典控制环境中，ARISE 的平均每回合回报达到 333.25，略高于近端策略优化（330.96），并显著高于深度Q网络（130.53）和优势演员-评论家（64.39）。这表明，将受粒子群启发的动作空间探索与逐智能体 PPO 策略相结合，即使在维度相对较低的设置中，也能有效稳定学习并加快收敛。

在连续控制任务中，ARISE 展现出尤为显著的优势。在 Hopper-v4 中，ARISE 的表现比近端策略优化高出约 50 分，而优势演员-评论家则明显落后，这凸显了朴素演员-评论家方法在高维连续空间中面临的困难。在 Walker2d-v4 中，ARISE 的表现略逊于近端策略优化（ARISE：333.27，近端策略优化：337.54），这表明高维运动任务中存在固有的随机波动；然而，ARISE 的表现仍具竞争力，同时对奖励扰动具有更强的鲁棒性，非平稳变体中的优异结果便证明了这一点。在 Ant-v4 中，ARISE 缓解了优势演员-评论家所出现的灾难性失败（-1594.45），取得了 -108.40 的成绩，相比之下近端策略优化为 -143.49，展现了其在高度不稳定的高维环境中防止灾难性发散的能力。

非平稳任务清晰地展现了自适应群体机制的优势。即使奖励突然变化，ARISE 的表现仍显著优于近端策略优化和优势演员-评论家。例如，在带有奖励扰动的 CartPole-v1 中，ARISE 的成绩为 460.94，而近端策略优化和优势演员-评论家的成绩分别为 385.56 和 313.59。这验证了这一假设：基于新颖性的奖励增强、最优智能体广播和自适应参数调优相结合，使 ARISE 能够针对环境非平稳性动态调整探索与利用之间的平衡。

消融研究进一步凸显了各架构组件的贡献。移除自适应参数调优（ARISE NO ADAPTIVE）会导致所有环境中的性能持续下降，这凸显了动态惯性和社会/认知权重调整在引导基于群体的探索方面的作用。同样，取消群体集成（ARISE NO SWARM）会显著降低性能，证实混合强化学习-粒子群优化机制是 ARISE 提升学习效率的关键。与原始近端策略优化相比，这些消融结果表明，混合群体组件和自适应参数化协同提升了收敛速度和鲁棒性。

总体而言，结果表明 ARISE 在平稳和非平稳环境中都实现了稳定性与鲁棒性的独特结合。将逐智能体近端策略优化与受粒子群启发的动作空间探索相结合，引入了一种有原则的机制，以促进策略搜索的多样性；而新颖性奖励和广播策略则确保高性能策略能够高效传播。这些发现表明，当高维连续控制和非平稳性给传统强化学习算法带来巨大挑战时，ARISE 尤其适用于此类领域。

7 结论

在本研究中，我们提出了 ARISE（自适应强化学习与群体探索集成），这是一种新颖的混合强化学习框架，将逐智能体 PPO 策略与受粒子群启发的动作空间探索机制相结合。通过整合基于新颖性的奖励增强、最优智能体广播和自适应参数调优，ARISE 同时提升了探索效率、加快了收敛速度，并增强了对非平稳环境的鲁棒性。

在标准控制、高维连续控制和非平稳任务上进行的实验评估证实，ARISE 在所有情况下的表现始终优于常用的强化学习方法（近端策略优化、深度Q网络和优势演员-评论家）。实验还表明，群体集成和自适应参数调制对于提升学习性能都不可或缺。尤其是，ARISE 抵御奖励波动的能力令人印象深刻，这凸显了该方法适用于充满动态变化和不确定性的现实场景。

总而言之，ARISE 在融合强化学习技术方面实现了突破，提出了一种有原则的方案，将个体策略优化与集体群体动力学结合起来。此外，未来研究可进一步拓展至多智能体系统、观测受限的环境以及现实世界机器人控制任务，并有望通过混合探索-利用范式进一步提升性能。

参考文献

1. Bellemare, M.G., Srinivasan, S., Ostrovski, G., 等：统一基于计数的探索与内在动机。载于：神经信息处理系统进展（2016） 2. Cully, A., Clune, J., Tarapore, D., Mouret, J.B.：能够像动物一样适应的机器人。《自然》**521**，503–507（2015） 3. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.：软演员-评论家：结合随机演员的离策略最大熵深度强化学习。载于：国际机器学习会议（ICML）（2018） 4. Houthooft, R., Chen, X., Duan, Y., 等：VIME：变分信息最大化探索。载于：神经信息处理系统进展（NeurIPS）（2016） 5. Jaderberg, M., Mnih, V., Czarnecki, W., 等：神经网络的基于种群的训练。arXiv预印本 arXiv:1711.09846（2017） 6. Kennedy, J., Eberhart, R.：粒子群优化。载于：IEEE国际神经网络会议（ICNN）论文集（1995） 7. Khadka, S., Tumer, K.：进化强化学习。载于：神经信息处理系统进展（NeurIPS）（2018） 8. Lowe, R., Wu, Y., Tamar, A., 等：用于混合合作竞争环境的多智能体演员-评论家。载于：神经信息处理系统进展（NeurIPS）（2017） 9. Mnih, V., Kavukcuoglu, K., Silver, D., 等：通过深度强化学习实现人类水平控制。《自然》**518**(7540)，529–533（2015） 10. Mouret, J.B., Clune, J.：通过精英映射探索搜索空间。arXiv预印本 arXiv:1504.04909（2015） 11. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.：通过自监督预测实现好奇心驱动的探索。载于：国际机器学习会议（ICML）（2017） 12. Pourchot, A., Sigaud, O.：CEM-RL：结合进化方法与基于梯度的方法进行策略搜索。载于：ICML游戏强化学习研讨会（2018） 13. Rashid, T., Samvelyan, M., de Witt, C.S., 等：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（ICML）（2018） 14. Salimans, T., Ho, J., Chen, X., Sutskever, I.：进化策略：一种可扩展的强化学习替代方案。arXiv预印本 arXiv:1703.03864（2017） 15. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.：近端策略优化算法。arXiv预印本 arXiv:1707.06347（2017） 16. Strehl, A.L., Littman, M.L.：基于模型的马尔可夫决策过程区间估计分析。《计算机与系统科学杂志》（2008） 17. Sutton, R.S., Barto, A.G.：《强化学习：导论》。麻省理工学院出版社（2018）