基于价值进行演化的进化强化算法

摘要

本文提出了基于价值演化的强化学习（VEB-RL）框架，旨在融合进化算法（EAs）与基于价值的强化学习（RL）。以往研究多将 EAs 与基于策略的 RL 结合，而忽视了基于价值的 RL。VEB-RL 通过维护价值函数种群，而非策略种群，采用负时间差分（TD）误差作为适应度度量，相较于累积奖励，该度量更样本高效且与价值函数准确性密切相关。此外，VEB-RL 引入精英交互机制，仅让精英个体与环境交互，提供高质量样本，避免低质量样本的负面影响。实验在 MinAtar 和 Atari 平台上进行，结果表明 VEB-RL 显著提升了 DQN、Rainbow 和 SPR 等算法的性能，优于其他先进方法。VEB-RL 不仅提高了样本效率，还在性能上取得了突破，为基于价值的 RL 发展提供了新思路，对游戏 AI、机器人控制等领域具有重要应用价值。

关键词：强化学习；进化算法；进化强化学习；价值函数

1 引言

强化学习在众多实际任务中，如游戏人工智能 [13]、机器人控制 [4] 以及自动驾驶等领域取得了显著的成功。尤其是深度强化学习（Deep Reinforcement Learning，DRL），它结合了强化学习和深度神经网络，借助神经网络强大的近似能力，能够高效地利用梯度信息进行学习，从而实现了令人印象深刻的成果。然而，DRL 也存在一些局限性，包括探索能力不足 [8]、收敛性差以及容易陷入次优策略等问题。此外，DRL 对奖励信号的准确性高度依赖，低质量的奖励（例如稀疏、噪声或具有欺骗性的奖励）会严重影响其性能。

进化算法（Evolutionary Algorithms，EAs）作为一种无需梯度的优化方法，可以作为 RL 的替代方案。与 RL 不同，EAs 维护一个种群而非单一个体，并通过对个体引入扰动来产生后代，以寻求更优的解决方案。EAs 在探索能力、收敛性和鲁棒性方面相较于 RL 表现出更高的效率 [6]。尽管 EAs 具有这些优势，但它们也存在样本效率低下的问题，通常需要比 RL 多几个数量级的样本。

为了结合 EAs 和 RL 的优势，实现更高效的策略搜索，已经提出了许多方法。例如，ERL [5] 将遗传算法（Genetic Algorithm，GA）与 DDPG [7] 相结合。在 ERL 中，EA 和 RL 并行优化，EA 在种群评估过程中生成的样本提供给 RL 进行优化，旨在提高样本效率。同时，RL 将优化后的策略注入到种群中参与进化过程。ERL 定义种群中策略的适应度为与环境交互多个回合的平均累积奖励，这一度量方式随后成为 ERL 相关工作的常用适应度度量。尽管 ERL 及其扩展方法取得了成功，但以往将 EAs 与 RL 结合的方法主要关注基于策略的 RL，即通

过策略构建种群，并根据与环境交互获得的累积奖励对个体进行排名。然而，基于价值的 RL 及其特性在 ERL 领域中常常被忽视。

将现有框架应用于基于价值的 RL 时面临两个问题：一是基于价值的算法优化价值函数，直接将价值函数视为策略通过累积奖励进行进化是可行的，但这忽略了价值迭代的原则；二是种群产生的低质量样本可能会导致次优的 RL 优化，并降低样本效率。为了解决这些问题，本文提出了一种新的框架——基于价值演化的强化学习（VEB-RL），将 EAs 与基于价值的 RL 相结合。VEB-RL 维护一个 Q 网络（动作价值网络）及其对应目标网络的种群，不使用累积奖励作为种群评估的适应度度量，而是定制了一个新的适应度度量——负时间差分（TD）误差，该度量量化了 Q 网络估计值与目标值之间的差异。定制的适应度度量与价值迭代的目标一致，且无需与环境交互，从而提高了样本效率。借助定制的适应度度量，我们可以在无需与环境交互的情况下评估个体适应度并进行排名。为了提供高质量样本并避免低质量样本对 RL 优化的潜在负面影响，本文提出了一种新的精英交互机制，仅限制种群中的精英个体与环境进行交互。本文的主要贡献如下：

1. 提出了一种简单而高效的 VEB-RL 框架，用于增强基于价值的 RL。

2. 考虑到基于价值的 RL 的特性，提出使用负 TD 误差作为适应度度量。

3. 提出了精英交互机制，以确保数据样本的质量并提高样本效率。

4. 在 MinAtar、Atari 和 Atari 100k 上的实验表明，VEB-RL 显著提升了 RL 算法的性能，并优于其他强大的 ERL 基线。此外，VEB-RL 还可以改进非基于价值的 RL，并在 MUJOCO 任务上证明了其有效性。

2 相关工作

近年来，探索进化算法（EAs）与强化学习（RL）协同作用的研究领域受到了广泛关注。在这一领域，已经提出了一系列显著的研究成果。例如，ERL [5] 首次提出了一种新颖的混合框架，其中同时优化了一个演员-评论家 RL 代理和一个遗传算法（GA）种群。RL 代理和 GA 种群相互受益，种群评估过程中产生的多样化经验被提供给 RL 的回放缓冲区进行梯度优化。同时，通过注入优化后的 RL 演员来加速种群的进化。PDERL [1] 在 ERL 的框架基础上进行了改进，优化了遗传算子以解决灾难性遗忘问题。Supe-RL [9] 则定期在参数空间中围绕当前的 RL 演员进行搜索，并评估生成的策略。随后，RL 演员会向表现最佳的策略进行软更新。CEM-RL [11] 结合了交叉熵方法（CEM）和 TD3 算法，利用 RL 评论家的梯度信息更新种群中的一半个体，从而影响 CEM 的分布。ERL-Re2 [3] 将策略解耦为共享的非线性状态表示和个体线性策略表示，旨在共享知识并减少探索空间。这些算法利用与环境交互获得的累积奖励作为适应度度量来评估种群，忽略了价值函数的近似。此外，这些算法要求种群中的所有个体都与环境交互并生成经验。然而，低质量的经验可能会对 RL 的优化产生负面影响 [10]。这些局限性使得进一步提升基于价值的 RL 变得困难，而基于价值的 RL 正是这篇工作的主要关注点。

3 本文方法

本节介绍了基于价值演化的强化学习（VEB-RL），这是一种利用进化算法（EAs）增强基于价值的强化学习（RL）的混合框架。VEB-RL 包含两个核心组成部分：一是为与基于价

值的 RL 整合而定制的适应度度量；二是用于提升种群生成样本质量的精英交互机制。以下将详细介绍这些组成部分，并介绍基于遗传算法（GA）和交叉熵方法（CEM）的 VEB-RL 的两种不同变体。

3.1 定制的适应度度量

以往将 EAs 与基于策略的 RL 结合的方法具有两个主要特点。首先，种群由策略网络（即演员）组成；其次，使用与环境交互获得的累积奖励作为适应度度量来评估种群。然而，当将 EAs 与基于价值的 RL 结合时，进化的个体变成了价值网络。在这种情况下，使用基于累积奖励的适应度度量来评估种群进化可能并非最佳选择，因为基于价值的 RL 涉及到对期望值的显式建模。使用累积奖励作为适应度度量忽略了价值函数的近似，可能导致不准确的价值近似。

以深度 Q 网络（DQN）为例，其优化目标是通过 Bellman 方程的价值迭代获得最优 Q 网络 Q_θ^*，以准确估计状态-动作值。Q 的损失函数定义为：

$$L(\theta) = \mathbb{E}_{s,s',a\sim D}\left[\left(r + \max_{a'} Q_{\theta'}(s',a') - Q_\theta(s,a)\right)^2\right] \tag{1}$$

其中，s, s', a 是从回放缓冲区 D 中采样的，$Q_{\theta'}$ 是目标 Q 网络。最优策略是选择一组动作 a，以最大化 $Q^*(s,a)$ 的期望值。显然，基于价值的 RL 的优化目标是获得准确的价值函数，而基于累积奖励的适应度函数的优化目标是增强从价值函数衍生的策略，这与价值函数的优化目标不一致。这种不一致性可能使得获得足够准确的价值函数变得困难。

为了解决这个问题，这篇论文提出了新的种群组成和为种群评估定制的适应度度量。具体来说，VEB-RL 维护一个 Q 网络及其对应目标网络的种群，记为 $P = \{Q_{\theta_i}, Q_{\theta_i'}\}_{i=1}^n$，为适应度计算提供了基础。为了引导种群进化向更准确的价值函数发展，定义了用于评估种群内个体的适应度度量为负时间差分（TD）误差：

$$f(\theta_i, \theta_i') = -\mathbb{E}\left[\left(r + \max_{a'} Q_{\theta_i'}(s',a') - Q_{\theta_i}(s,a)\right)^2\right] \tag{2}$$

其中，s, a, s' 是从回放缓冲区 D 中采样的，$Q_{\theta_i}, Q_{\theta_i'}$ 是从种群 P 中采样的。当样本覆盖所有可能的转换时，TD 误差越小，价值函数近似就越准确。由于我们无法访问所有状态转换，因此从回放缓冲区中采样足够大的实例数 M 来近似真实误差。随着 $f(\theta_i, \theta_i')$ 的增加，Q 值近似变得越来越准确。具体来说，种群中的状态-动作值函数 $\{Q_{\theta_i}\}_{i=1}^n$ 通过 EAs（即 GA 和 CEM）进行优化。

目标网络 $Q_{\theta_i'}$ 每 H 代由其对应的 Q 网络 Q_{θ_i} 进行硬更新。此外，我们还保持了 RL 注入机制：由 RL 优化的 $Q_{\theta_{\text{rl}}}$ 及其目标网络 $Q_{\theta_{\text{rl}}'}$ 每代都注入到种群中。如果 RL 个体获得更高的 $f(\theta_{\text{rl}}, \theta_{\text{rl}}')$ 值并被选为精英，则它将引导种群并促进种群的进化；否则，$Q_{\theta_{\text{rl}}}$ 和 $Q_{\theta_{\text{rl}}'}$ 将被消除。

为了证明这一定制适应度度量的有效性，文章在 Space Invader 和 Freeway 任务上进行了 Spearman 等级相关性分析 [12]，比较了基于负 TD 误差适应度的个体排名和通过与环境交互 5 个回合获得的累积总奖励的排名。Spearman 等级相关系数忽略具体值，仅关注值的相对顺序。它的范围从-1 到 1，值越接近 1 表示排名相似度越高，而值越接近-1 表示排名差异越大。图 1所示的结果表明，在训练过程中，基于两种适应度度量的种群排名之间的相关性显

著且不断增加。这些结果表明，使用负 TD 误差作为适应度度量可以相对准确地反映个体的质量，且无需样本成本，从而提高了样本效率。

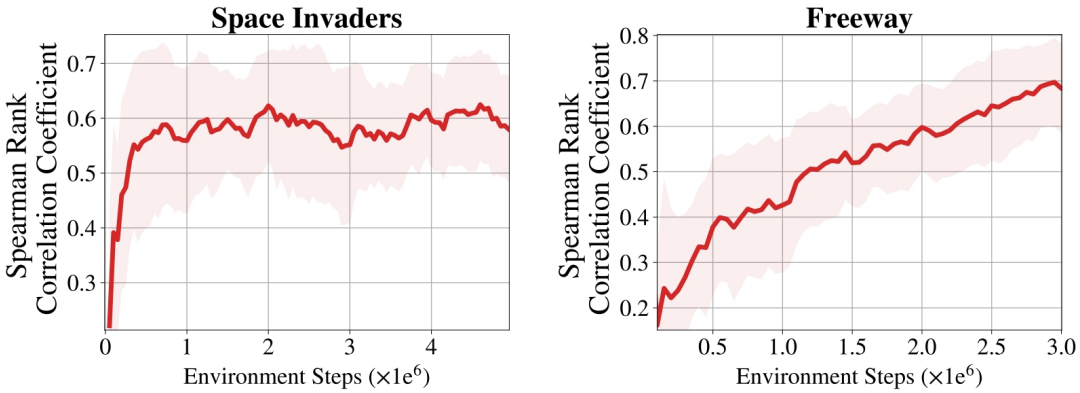

图 1. Spearman 等级相关性分析图

3.2 精英交互

以往的 ERL 工作要求种群中的所有个体都与环境交互以获得适应度 [6]，我们称这一过程为均匀交互。在这些工作中，增强 RL 的一个关键机制是将通过均匀交互生成的样本纳入共享的回放缓冲区进行 RL 优化。这一机制缓解了 RL 探索能力差的问题，并避免了评估样本的浪费。然而，这也引入了新问题：只有高质量个体的样本对 RL 策略优化有贡献，而低适应度个体生成的样本可能是无效的，并可能导致 RL 陷入次优解 [14]。

得益于新的适应度度量，种群中的个体不再需要直接与环境交互以获得适应度。相反，它们可以根据回放缓冲区中的样本来计算适应度。因此，我们可以主动选择哪些个体与环境交互以生成样本。文章称这种方式为选择性交互。图 2展示了均匀交互与选择性交互的对比。为了解决均匀交互所提到的问题，我们提出了一种新的交互方式，称为精英交互，它仅允许精英个体与环境交互以生成高质量样本。具体来说，我们首先根据公式 2 获得种群中个体的适应度 $\{f(\theta_i, \theta_i')\}_{i=1}^{n}$。随后，我们选择适应度最高的前 N 个个体与环境交互。直观上，精英交互使得表现良好的个体能够为回放缓冲区 D 贡献高质量样本，而表现不佳的个体则不会浪费资源进行交互，并避免了低质量样本对 RL 优化过程的潜在负面影响。

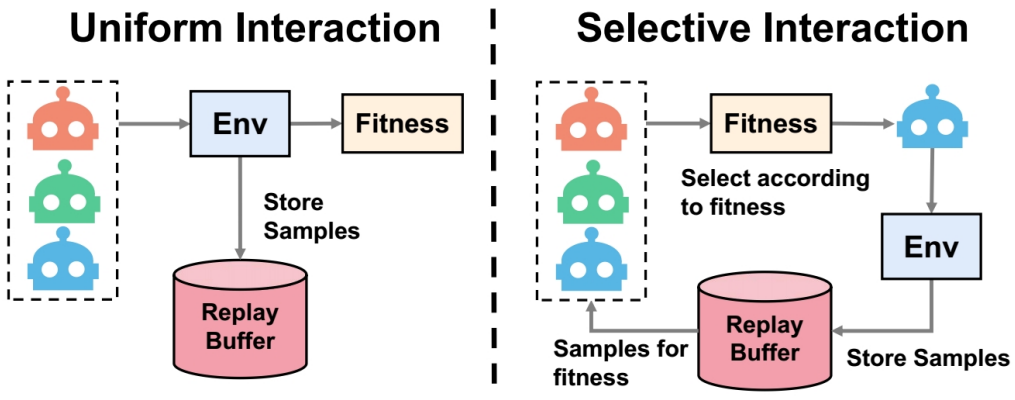

图 2. 均匀交互与选择性交互的架构对比图

3.3 VEB-RL 的算法框架

得益于新的适应度度量和精英交互，EA 和 RL 在 VEB-RL 中更好地融合，以高样本效率进行价值函数搜索。图 1 展示了 VEB-RL 的架构。对于进化过程，种群基于回放缓冲区 D 中的样本来获得适应度（公式 2）。适应度最高的前 N 个个体被选为精英个体，并与环境交互，为回放缓冲区 D 提供高质量样本以供 RL 优化。随后，我们使用 EA 进化种群，以寻找更好的 Q 网络并构建新的种群。对于强化过程，RL 个体与环境交互并将经验存储在回放缓冲区 D 中。然后，RL 基于共享的回放缓冲区通过公式 1 优化其 Q 网络。RL 目标网络定期进行硬更新。为了促进种群进化，每代都将优化后的 $Q_{\theta_{\text{rl}}}$ 及其目标网络 $Q_{\theta'_{\text{rl}}}$ 注入到种群中。种群中的目标网络每 H 代由其对应的 Q 网络更新一次，以实现稳定的改进，这与传统 RL 的做法类似。

原则上，VEB-RL 是一个通用框架，可以与任意 EAs 结合。我们使用两种不同的 EAs 实现进化过程：GA 和 CEM。VEB-RL 的伪代码如算法 1 所示。在基于 GA 的 VEB-RL 中，进化涉及选择、交叉和变异操作。适应度最高的前 N 个个体被选为精英。接下来，我们使用锦标赛选择来选择胜者，这涉及到在每次迭代中从随机选择的三个个体中挑选最佳个体，并重复这一过程一定次数。未被选为精英或胜者的个体被记录为弃用者。对于交叉操作，我们从精英和胜者中随机选择两个个体作为父母，通过 k 点交叉产生后代以替换弃用者，直到所有弃用者都被替换。对于变异操作，所有非精英个体根据一定概率添加高斯噪声。种群通过交叉和变异进行更新。对于基于 CEM 的 VEB-RL，进化主要涉及分布更新。种群的前半部分根据公式 1 更新 π_μ 和 Σ。此外，每代开始时，新种群从 $\mathcal{N}(\mu, \Sigma)$ 中抽取。

4 复现细节

4.1 与已有开源代码对比

本篇论文是开源代码的，源代码可以在 https://github.com/yeshenpy/VEB-RL 上找到。本次复现的工作主要如下：1. 复现了原论文在 MUJOCO 上 Humanoid 、Ant 和 Walker 的环境的结果。2. 修改了代码里面一些小 bug，如一些运行效率问题

4.2 实验环境搭建

本次复现主要在 MUJOCO 的连续控制环境上进行复现。MUJOCO 是一个用于物理仿真的物理引擎和运动动力学库，提供了一个高效的物理引擎，用于模拟多关节动力学系统的运动。这些系统可以包括机器人、生物学模型以及其他复杂的多体动力学系统。本次复现用的是 MUJOCO210 版本，环境仿真的效果如图 3

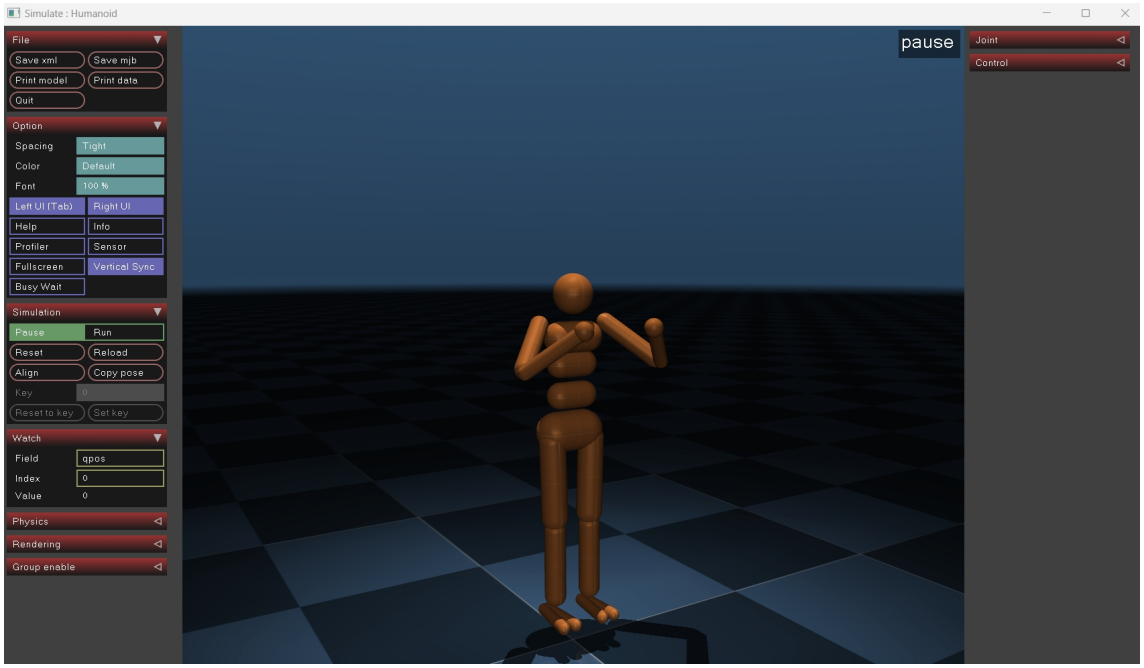

图 3. MUJOCO210

5 实验结果分析

参照论文的实验，在 ERL 相关研究中常用的 Ant、Walker 和 Humanoid 任务上评估了 VEB-TD3、TD3 和其他 ERL 相关方法。每个实验重复五次运行和 300 万个环境步骤。表 1中的结果表明 VEB-RL 的表现优于 TD3 和 CEM-RL, ERL 和 PDERL, 进一步证明了 VEB-RL 的效率和泛化能力。

	TD3	ERL	CEM-RL	PDERL	VEB
Ant	6225	6245	5333	5334	**7370**
Humanoid	6366	6260	209	6738	**6955**
Walker	4800	4770	5017	500	**5487**

表 1. VEB-TD3 与其他基线在 MUJOCO 任务上的平均性能比较

6 总结与展望

这篇文章提出了一种创新的强化学习框架——基于价值演化的强化学习（VEB-RL），它巧妙地融合了进化算法（EAs）和基于价值的强化学习（RL）。VEB-RL 的创新之处在于它采用了负时间差分（TD）误差作为适应度度量，这一度量不仅提高了样本效率，而且与价值函数近似的准确性紧密相关。相当于用更好的方法解决了 TD3 算法主要解决的价值高估问题 [2]，因为 TD3 算法仍然存在价值高估的问题。此外，该框架引入了精英交互机制，确保只有种群中的精英个体与环境交互，从而为 RL 优化提供高质量样本，避免低质量样本的负面影响。在 Ant、Walker 和 Humanoid 任务中，VEB-RL 显著提升了性能，显示出其在样本效

率和最终性能上的优越性。VEB-RL 的提出，为强化进化学习领域带来了新的理论贡献，并在实际应用中展现出广泛的应用前景。

参考文献

[1] Cristian Bodnar, Ben Day, and Pietro Lió. Proximal distilled evolutionary reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3283–3290, 2020.

[2] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[3] HAO Jianye, Pengyi Li, Hongyao Tang, Yan Zheng, Xian Fu, and Zhaopeng Meng. Erl-re ^2: Efficient evolutionary reinforcement learning with shared state representation and individual policy representation. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pages 6023–6029. IEEE, 2019.

[5] Shauharda Khadka and Kagan Tumer. Evolution-guided policy gradient in reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[6] Pengyi Li, Jianye Hao, Hongyao Tang, Xian Fu, Yan Zheng, and Ke Tang. Bridging evolutionary algorithms and reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2401.11963*, 2024.

[7] TP Lillicrap. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[8] Jinyi Liu, Zhi Wang, Yan Zheng, Jianye Hao, Chenjia Bai, Junjie Ye, Zhen Wang, Haiyin Piao, and Yang Sun. Ovd-explorer: Optimism should not be the sole pursuit of exploration in noisy environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13954–13962, 2024.

[9] Enrico Marchesini, Davide Corsi, and Alessandro Farinelli. Genetic soft updates for policy evolution in deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

[10] Evgenii Nikishin, Max Schwarzer, Pierluca D ' Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pages 16828–16847. PMLR, 2022.

[11] Aloïs Pourchot and Olivier Sigaud. Cem-rl: Combining evolutionary and gradient-based methods for policy search. *arXiv preprint arXiv:1810.01222*, 2018.

[12] Philip Sedgwick. Spearman' s rank correlation coefficient. *Bmj*, 349, 2014.

[13] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

[14] Hongchang Zhang, Jianzhun Shao, Yuhang Jiang, Shuncheng He, Guanwen Zhang, and Xiangyang Ji. State deviation correction for offline reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9022–9030, 2022.