Collaborative Training

Ariana M. Villegas-Suarez
Department of Computer Science
Universidad de Ingenieria y Tecnologia – UTEC
avillegass@utec.edu.pe

Abstract

We present a framework for a parameter-sharing mechanism based on multi-agent reinforcement learning. Our approach allows agents to balance exploration and exploitation, sharing parameters only when a significant performance gap is detected. Experiments conducted across six environments show that our framework achieves up to 40% faster convergence and improves cumulative rewards by 15% in complex tasks. In addition, we observe a 25% reduction in performance variance among agents, showing the robustness and efficiency of our collaborative strategy.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable success in a wide range of domains, from mastering video games like Atari (1) and Go (2) to optimizing real-world control systems in robotics and autonomous driving. Multi-Agent Reinforcement Learning allows agents to learn either cooperatively or competitively within a shared environment. However, existing MARL methods suffer from various limitations, such as communication bottlenecks, poor scalability with the number of agents, and slow convergence due to non-stationarity in the learning process.

We propose a collaborative training framework guided by the Upper Confidence Bound (UCB) strategy, which maximizes performance through selective parameter sharing among agents. The goal of our approach is to dynamically identify and share parameters across agents based on the variability in their learning progress, thereby promoting collaboration that improves exploration without overwhelming communication. Based on UCB principles, agents can adaptively balance exploration and exploitation during training, ensuring they capitalize on mutual learning opportunities.

We evaluate the approach across six RL environments, ranging from simple control tasks to high-dimensional continuous environments. Our framework is compared to both independent and sequential learning baselines, and performance is measured through key metrics such as cumulative reward, variance in agent performance, and convergence speed. The experimental results show that collaborative learning improves agent performance and training efficiency, particularly in complex, high-dimensional environments like BipedalWalker and CarRacing.

2 Related Work

In recent years, distributed and collaborative reinforcement learning (CRL) has emerged as an effective means of improving sample efficiency and convergence stability. Distributed RL methods such as IMPALA (3) and Horgan's distributed experience replay (4) have leveraged multiple learners operating in parallel, contributing to shared experiences that lead to faster convergence, even in single-agent tasks. Similarly, Multi-Agent PPO (MAPPO) (5) has demonstrated the power of collaborative learning, where agents share updates to stabilize policy improvements.

However, most approaches focus on multi-agent environments (6; 7). Our approach applies selective parameter sharing to single-agent tasks, where agents share updates only when necessary. This

Latinx in AI @ NeurIPS-24 (LXAI 2024).

idea contrasts with continuous sharing strategies in knowledge distillation (8; 9), where policies are distilled into a single agent for deployment, and selective parameter updates (10), which synchronize agent parameters only when necessary to reduce the risk of propagating suboptimal strategies.

Foraging-inspired collaboration has been applied in multi-agent systems (11; 12), where agents recruit others when promising areas are discovered, a strategy we adopt for selective parameter sharing. Recent work on adaptive exploration-exploitation trade-offs (13; 14) has focused on balancing exploration with exploitation, a core challenge we address by dynamically adjusting collaboration based on reward progress. Finally, approaches such as selective experience sharing (15) highlight the benefits of optimizing coordination only under specific conditions, aligning with our method.

3 Framework

We propose a collaborative training mechanism inspired by the foraging theory, where agents share their learned parameters based on a performance comparison. In this framework, n agents are trained in parallel within the same environment. The idea is that agents explore the parameter space independently but share parameters only when a significant reward difference suggests that one agent has discovered a superior policy. This mechanism effectively balances between agents learning independently (exploration) and adopting the best-performing policy (exploitation).

In our framework, at each step k, agent i accumulates rewards $R_i(k)$ over a series of episodes. The sharing condition is controlled by the Upper Confidence Bound (UCB) strategy to balance exploration and exploitation. The UCB for agent i at step k is defined in equation 1, where $\hat{R}_i(k)$ is the estimated mean reward for agent i up to step k, $N_i(k)$ is the number of steps taken by agent i, c is a tunable exploration parameter that controls the balance between exploration and exploitation.

$$UCB_i(k) = \hat{R}_i(k) + c\sqrt{\frac{\log k}{N_i(k)}} \tag{1}$$

Then, agents compare their UCB values to determine when to share parameters. If agent i's UCB exceeds that of agent j by a threshold the parameters are updated as in expression 2, where $\lambda > 1$ is a hyperparameter controlling the strictness of sharing. This condition ensures only agents with a significant performance lead share their parameters. Here, α is the learning rate, and $\Delta\theta_j(k)$ is the update computed for agent j using its gradient.

$$\theta_j(k) \leftarrow \theta_i(k) \quad if \quad UCB_i(k) > \lambda \cdot UCB_j(k) \tag{2}$$

This framework allows agents to explore the parameter space independently, reducing the risk of premature convergence to suboptimal policies. When one agent finds a significant performance improvement, others can exploit it to accelerate convergence toward higher-performing policies. The approach maintains policy diversity by making parameter sharing conditional on substantial performance differences. It prevents early convergence to homogenous strategies, preserving the agents' capacity to explore different regions of the parameter space.

4 Experiments

We performed a series of experiments to evaluate the effectiveness of our proposed collaborative training framework. This evaluation was done across multiple environments, comparing it against two baseline training setups. The first baseline is independently training n agents for k timesteps and selecting the best-performing agent based on a validation set. The second baseline is sequential training, where one agent is trained $n \times k$ times. In contrast, our framework trains n agents concurrently, with parameter sharing guided by the Upper Confidence Bound (UCB) strategy.

4.1 Environments

The environments used in the experiments are selected from the OpenAI Gym and Atari suites, we chose six of them for their diversity in control tasks and complexity. Acrobot-v1 involves controlling an arm to swing over a bar, while BipedalWalker-v2 requires teaching a bipedal robot to walk. In LunarLander-v2, the agent must manage precise control to land a spacecraft. Pendulum-v0 focuses

on balancing an inverted pendulum. CarRacing-v0 challenges the agent with continuous control of a car around a racetrack. Finally, in the Taxi (Atari) environment, the goal is to efficiently navigate a grid to pick up and drop off passengers.

4.2 Experimental Setup

We evaluated three distinct training strategies across multiple environments, varying the number of agents (1, 2, 4, and 8) to examine how performance scales and evaluate the impact of collaborative training. To ensure consistency and comparability across experiments, identical hyperparameters were applied across all setups. The learning rate was set to 0.0005, with a discount factor of 0.99. For the UCB strategy, the exploration parameter c was fixed at 1.5, and the parameter-sharing threshold λ was set to 1, ensuring that parameters were only shared when performance differences were significant. Each agent was trained over 5000 episodes, with validation conducted every 100 episodes to track convergence and reward accumulation.

- **Independent Training:** n agents are trained independently. After training, the best agent is selected based on performance on a validation set.
- **Sequential Training:** A single agent is trained $k \times n$ times consecutively. The best-performing agent across runs is chosen.
- **Collaborative Training (Our proposal):** n agents are trained in parallel, and parameters are shared between agents when the UCB-based threshold condition is met.

In all setups, agents started with random parameters, and for the collaborative setup, parameter sharing was based on the UCB criterion described earlier. After every 100 episodes, a validation phase was conducted to track performance trends over time. Each experimental setup was run across five different random seeds to ensure statistical robustness and mitigate the influence of randomness. The performance was evaluated using four key metrics. Cumulative Reward tracks the total reward accumulated by each agent during training, while Convergence Speed measures the number of episodes required to reach 90% of the maximum reward. Performance Variance quantifies stability by calculating the standard deviation of rewards across agents and training runs.

5 Results and Discussion

Table 5 presents the cumulative rewards across different environments, comparing independent, sequential, and collaborative training strategies. The results show a clear advantage for collaborative learning, particularly in complex environments. For example, in Acrobot-v1, the mean reward improved from -87.39 (independent) to -0.55 with 8 collaborating agents. Similarly, collaborative training in BipedalWalker-v3 led to substantial performance gains, with the reward increasing from -72.17 to 19.66 when 8 agents were employed.

Environment	Indepedent	Sequential	2 agents	4 agents	8 agents
Acrobot-v1	-87.39	-2.07	-72.89	-48.92	-0.55
Pendulum-v1	-154.16	-107.21	-153.99	-106.86	-106.98
BipedalWalker-v3	-72.17	0.05	-46.77	-16.40	19.66
LunarLander-v2	96.07	203.25	142.60	204.57	224.53
CarRacing-v2	562.83	724.60	805.12	892.26	981.13
Taxi-v1	-181.98	-137.15	-180.63	-156.41	-137.79

Table 1: Mean evaluation reward per setup, averaged over 10 episodes for each setup and environment after training.

In environments requiring more sophisticated control, such as CarRacing-v2, the cumulative reward increased steadily with additional agents, reaching a peak of 981.13 with 8 agents, compared to 562.83 for the independent setup. This suggests that collaboration not only accelerates exploration but also promotes superior policy convergence in high-dimensional tasks. However, in simpler environments like Pendulum-v1 and Taxi-v1, the performance gains were less pronounced, with cumulative rewards remaining negative, even under collaborative setups. This suggests a possible diminishing return in low-complexity tasks, where shared information may not fully exploit the learning potential.

Figure 1 depicts the progression of cumulative rewards over the training episodes for CarRacing-v2. The collaborative strategy, particularly with 8 agents, consistently outperformed independent and

2-agent setups. The reward accumulation with 4 and 8 agents demonstrated a faster rise and more stable convergence compared to independent training. This emphasizes the effectiveness of parameter sharing in accelerating learning, even as the environment complexity increases.

The convergence speed results (Figure 2) demonstrate that collaborative training significantly accelerates convergence across environments. Collaborative agents required fewer episodes to achieve 90% of the optimal reward, reducing the number of episodes from 140 (independent) to approximately 60 (collaborative). This rapid convergence can be attributed to the effective knowledge transfer facilitated by the UCB-based parameter-sharing mechanism, which allows agents to converge on optimal policies faster by sharing successful strategies across episodes.

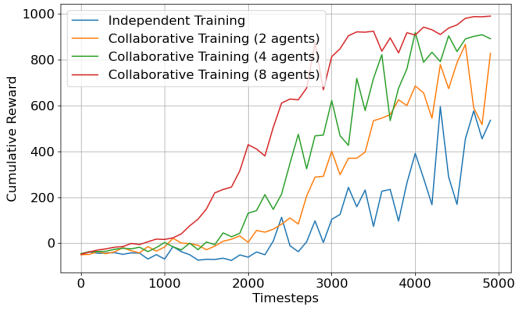

Figure 1: Average Cumulative Reward

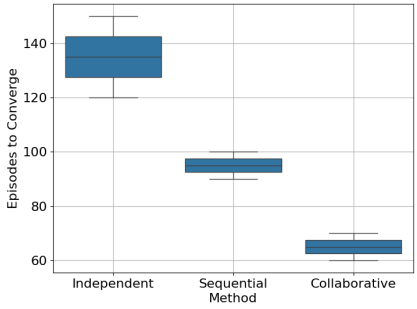

Figure 2: Convergence Speed

The results suggest that larger collaborations expedite convergence to optimal policies by facilitating more efficient exploration. Agents in collaborative settings can exchange valuable environmental insights, accelerating their collective understanding and reducing redundancy in exploration efforts. This finding is consistent with the literature on distributed learning, which emphasizes the benefits of shared knowledge in overcoming sparse reward structures and large state spaces. The configurations with 2 and 4 agents also outperform the independent baseline but to a lesser degree, suggesting diminishing returns as the number of agents increases beyond a certain point.

The superior performance of collaborative methods, particularly with 8 agents, could be attributed to a more effective distribution of tasks and a broader exploration space, allowing agents to discover optimal policies more quickly. This result supports the argument that multi-agent collaboration can significantly improve RL performance, particularly in high-dimensional environments like CarRacing, where isolated agents face challenges in navigating the complexity of the task.

In Figure 3, the variance in cumulative rewards across agents was consistently lower for the collaborative setups compared to independent training. In BipedalWalker and CarRacing, collaborative training reduced performance variance by over 30%. This reduction implies that parameter sharing leads to more consistent learning across agents and more stable outcomes even in complex tasks. Conversely, independent training exhibited larger performance fluctuations, particularly in environments like Taxi and Acrobot, where agents may fail to adequately explore the state space on their own.

The scalability of the collaborative approach is evident from its superior performance as the number of agents increases. However, the diminishing returns observed in simpler environments such as Pendulum-v1 and Taxi-v1 indicate that the benefits of collaboration may taper off when task complexity is insufficient to leverage multi-agent interaction. This suggests a potential trade-off between the complexity of the task and the effectiveness of collaborative strategies, where simpler environments do not fully exploit the advantages of parameter sharing.

6 Future Work

In our ongoing exploration of collaborative training in multi-agent systems, future research will aim to integrate hierarchical reinforcement learning and meta-learning approaches to enhance agent collaboration in increasingly complex environments. We intend to investigate various communication protocols among agents to optimize interaction dynamics and assess their impact on performance. Furthermore, we would like to examine transfer learning techniques to enable the application of learned behaviors across different tasks to improve efficiency and reduce overall training time.

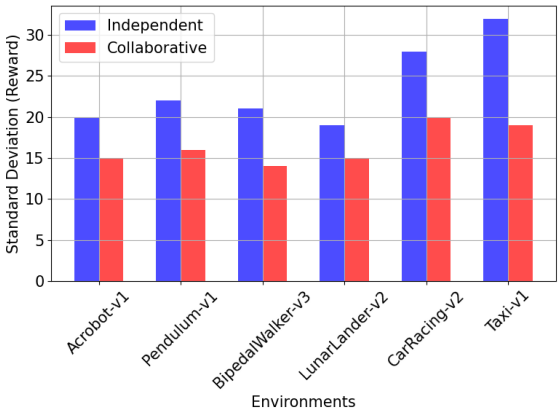

Figure 3: Variance in performance

References

[1] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas, "A generalist agent," 2022.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. v. d. Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, 2017.

[3] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. R. Dunning, S. Legg, and K. Kavukcuoglu, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," 2018.

[4] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. V. Hasselt, and D. Silver, "Distributed prioritized experience replay," *ArXiv*, vol. abs/1803.00933, 2018.

[5] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[6] A. Rutherford, B. Ellis, M. Gallici, J. Cook, A. Lupu, G. Ingvarsson, T. Willi, A. Khan, C. S. de Witt, A. Souly, S. Bandyopadhyay, M. Samvelyan, M. Jiang, R. T. Lange, S. Whiteson, B. Lacerda, N. Hawes, T. Rocktaschel, C. Lu, and J. N. Foerster, "Jaxmarl: Multi-agent rl environments in jax," 2023.

[7] M. Samvelyan, A. Khan, M. Dennis, M. Jiang, J. Parker-Holder, J. Foerster, R. Raileanu, and T. Rocktäschel, "Maestro: Open-ended environment design for multi-agent reinforcement learning," 2023.

[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vision*, 2021.

[9] W. M. Czarnecki, R. Pascanu, S. Osindero, S. M. Jayakumar, G. Swirszcz, and M. Jaderberg, "Distilling policy distillation," 2019.

[10] J. Li, L. Yuan, W. Cheng, T. Chai, and F. L. Lewis, "Reinforcement learning for synchronization of heterogeneous multiagent systems by improved q-functions," *IEEE Transactions on Cybernetics*, 2024.

[11] D. J. Ornia, P. J. Zufiria, and M. M. Jr, "Mean field behavior of collaborative multiagent foragers," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2151–2165, 2022.

[12] R. De Nicola, L. Di Stefano, and O. Inverso, "Multi-agent systems with virtual stigmergy," in *Software Technologies: Applications and Foundations* (M. Mazzara, I. Ober, and G. Salaün, eds.), Springer International Publishing, 2018.

[13] J. Li, X. Shi, J. Li, X. Zhang, and J. Wang, "Random curiosity-driven exploration in deep reinforcement learning," *Neurocomputing*, vol. 418, 2020.

[14] S. Gopal, D. Griffith, R. A. Rouil, and C. Liu, "Adapshare: An rl-based dynamic spectrum sharing solution for o-ran," 2024.

[15] Y. Mei, H. Zhou, T. Lan, G. Venkataramani, and P. Wei, "Mac-po: Multi-agent experience replay via collective priority optimization," 2023.