Hybridizing Evolutionary Algorithms and Reinforcement Learning for Superior Decision-Making in Complex Systems

Denis-Claudiu ASAN
Bucharest University of Economic Studies
Faculty of Cybernetics, Statistics and Economic Informatics
Bucharest, Romania
asandenis21@stud.ase.ro

This study examines the effectiveness of hybridizing Evolutionary Algorithms (EA) and Reinforcement Learning (RL) for decision-making in sparse-reward environments. Three models are implemented and compared, namely: (1) a population-based Evolutionary Algorithm with elitism, (2) an actor-critic Reinforcement-Learning method regularized with Behavior Cloning, and (3) a hybrid Lamarckian Memetic framework that combines evolutionary search with gradient-based fine-tuning. All models are trained offline using the antmaze-umaze-v2 dataset from the D4RL benchmark and evaluated both through critic-based offline estimates and online MuJoCo simulations. While all methods achieve comparable offline critic values, the hybrid approach demonstrates superior online normalized performance scores. The results suggest that combining population-based exploration with gradient-based exploitation mitigates local optima in sparse-reward decision-making tasks.
Keywords: *Evolutionary Algorithms, Reinforcement Learning, Offline RL, Memetic Algorithms, Sparse Reward, Hybrid Optimization*

1 Introduction

Learning effective policies under conditions of limited or delayed feedback remains a significant challenge, particularly for data-driven control methods. This difficulty is further compounded when agents must rely on incomplete or fixed datasets, where no additional interaction with the environment is possible. The issue is especially evident in offline reinforcement learning, where models are trained exclusively on previously collected data. While this setting is valuable in scenarios where safety constraints or limited data availability are important, it also introduces challenges such as poor generalization and convergence to suboptimal solutions.

To address these limitations, a range of approaches has been explored. Reinforcement Learning (RL), particularly actor-critic methods, has shown strong potential for policy improvement; however, its effectiveness often depends on the availability of informative reward signals and careful hyperparameter tuning. In contrast, Evolutionary Algorithms (EA) prioritize exploration through population-based search, making them more robust in environments with sparse rewards. Nevertheless, this advantage often comes at the cost of lower sample efficiency and reduced precision.

In this work, I investigated whether combining these complementary approaches can lead to improved performance in complex decision-making tasks. Specifically, I implemented and compared three models, all trained offline on the D4RL antmaze-umaze-v2 dataset: a population-based evolutionary method, a behavior-regularized actor-critic model, and a hybrid Lamarckian memetic approach that integrates evolutionary search with gradient-based fine-tuning. The models are evaluated using both offline critic estimates and online performance in MuJoCo simulations. The objective is to assess whether this hybrid approach can more ef-

fectively balance exploration and exploitation in sparse-reward environments.

2 Literature review

The challenge of effective decision-making in environments characterized by sparse rewards remains a significant hurdle in artificial intelligence [1]. Traditional reinforcement learning agents often struggle when feedback is infrequent, as they must perform extensive exploration before encountering any meaningful signal to guide their learning [2]. In order to address these limitations hybridizing Evolutionary Algorithms and Reinforcement Learning can be a viable solution, focusing on how population-based exploration can complement gradient-based exploitation in complex tasks like the D4RL AntMaze benchmark [3][4].

Evolutionary Algorithms, such as Genetic Algorithms (GA) and Evolution Strategies (ES), offer a gradient-free approach to optimization that mimics natural selection [1][5]. These methods are particularly resilient in the face of local optima and sparse rewards because they evaluate fitness at an episodic level rather than requiring step-by-step feedback [2].

Significant advancements have shown that GAs can be a competitive alternative for training deep neural networks [1][6]. For instance, Such et al. (2017) demonstrated that a simple GA could evolve networks with millions of parameters, often outperforming gradient-based methods in "deceptive" environments where reward signals are misleading [6]. Similarly, Salimans et al. (2017) positioned ES as a scalable alternative to RL, highlighting its invariance to action frequency and delayed rewards while achieving expert-level performance on MuJoCo tasks [7]. These findings suggest that the global search capabilities of EAs provide a robust foundation for exploration that single-agent RL often lacks.

In contrast to EAs, gradient-based RL—particularly actor-critic methods—excels at fine-grained exploitation of state-action pairs [8]. However, these methods are notoriously sensitive to hyperparameters and often exhibit brittle convergence properties. In sparse-reward settings, such as those found in the MuJoCo simulator, the difficulty of temporal credit assignment often causes standard algorithms like DDPG or PPO to fail entirely [2].

To mitigate these issues, researchers have turned to regularization and demonstration-based guidance. For example, the LOGO algorithm utilizes offline demonstration data to provide a "policy guidance" step, helping the agent navigate sparse rewards without being strictly limited to imitating a suboptimal policy [9].

The field of Offline RL (or batch RL) presents a unique challenge: learning purely from a fixed dataset without environment interaction [10]. This paradigm is critical for real-world applications where data collection is expensive or dangerous. However, agents often suffer from extrapolation errors when they take actions not covered in the training data [11].

Recent literature has proposed several strategies to constrain the learned policy to the behavior distribution. Conservative Q-Learning (CQL) addresses this by regularizing the value function to assign low values to out-of-distribution actions, which is essential for "stitching" suboptimal trajectories in AntMaze tasks [12]. Other minimalist approaches, such as TD3+BC, simply add a Behavior Cloning (BC) term to the policy update, matching state-of-the-art performance with significantly less complexity [10]. More advanced methods like Implicit Q-Learning (IQL) and AWAC further refine this by avoiding unseen actions during value training or using advantage-weighted updates to facilitate online fine-tuning [13]. Despite these successes, maintaining the balance between following the dataset and discovering improved policies remains a central tension in the field.

The hybrid framework of Evolutionary Reinforcement Learning (ERL) seeks to combine the best of both worlds: the diverse exploration of EAs and the sample efficiency of gradient-based RL. In ERL, a population of actors generates diverse experiences for a shared replay buffer, while an RL agent is periodically injected back into the population to provide gradient information.

These models have demonstrated improved performance and robustness, particularly in environments where hyperparameters are sensitive [1]. For instance, ERL-Re2 improves efficiency by sharing state representations across the population, addressing the knowledge-transfer limitations of older frameworks [8].

Memetic Algorithms (MA) represent a specific subset of EvoRL that combines global population-based search with local refinement techniques [14]. While Darwinian evolution only passes on genetic traits, Lamarckian learning allows an individual's acquired knowledge, such as policy weights fine-tuned via RL, to be inherited by the next generation [14].

Contemporary applications of memetic structures include production scheduling and routing problems, where RL agents are injected into a GA to improve sequencing decisions. For example, Zou et al. (2024) utilized a Q-learning guided local search within an EA to solve the Latency Location Routing Problem, outperforming existing metaheuristics [15]. Furthermore, Qu et al. (2019) proposed an evaluation-free local search that uses Q-learning to exploit frames collected during evolutionary exploration, significantly accelerating convergence [14].

While the reviewed literature highlights the strengths of hybrid approaches, a critical gap remains in the systematic comparison of offline-trained population-based methods versus regularized actor-critic models in high-dimensional, sparse-reward tasks like AntMaze.

The present study contributes to this field by implementing a hybrid Lamarckian Memetic framework trained offline on the D4RL dataset. By combining population-based exploration with gradient-based exploitation, this research aims to show that the hybrid approach can more effectively mitigate local optima in sparse-reward tasks than either individual model, providing a path toward more robust decision-making in complex environments.

3 Data

This study utilizes the antmaze-umaze-v2 dataset from the D4RL (Datasets for Deep Data-Driven Reinforcement Learning) benchmark, a standard environment for evaluating offline reinforcement learning methods under sparse-reward conditions. The dataset consists of trajectories collected from a MuJoCo-based simulated environment, where an agent must navigate a quadruped robot (Ant) through a maze to reach a target location.

The dataset includes tuples of the form (s, a, r, s', d), where s denotes the state (29-dimensional observation space), a represents the action (8-dimensional continuous control), r is the reward, s' is the next state, and d indicates episode termination. Rewards in this environment are sparse, typically binary (0 or 1), reflecting whether the goal has been reached or not. This characteristic makes the task particularly challenging, as informative feedback is limited and delayed.

To support efficient training, the dataset is loaded into an offline replay buffer, which enables random mini-batch sampling during optimization. In cases where next-state observations are not explicitly provided, they are reconstructed through sequential shifting. All data are converted into tensor format and processed.

The use of a fixed dataset ensures that all models operate under identical constraints, preventing additional environment interaction and emphasizing the challenges of ex-

trapolation and generalization inherent to offline reinforcement learning. These challenges are reflected in the discrepancy between offline estimates and actual performance observed later in the online results.

4 Methodology

To investigate the effectiveness of hybridizing Evolutionary Algorithms and Reinforcement Learning, three distinct approaches were implemented and evaluated under a unified experimental framework: (1) an EA-only model, (2) an RL-only model, and (3) a hybrid EA+RL model based on a Lamarckian memetic paradigm.

4.1. Model Architecture

All approaches share a common function approximator. The actor network is a feed-forward neural network with two hidden layers (256 units each) and ReLU activations, mapping states to continuous actions via a final tanh activation. The critic network consists of two parallel Q-functions, each estimating the expected return of a state–action pair. The minimum of the two Q-values is used to mitigate overestimation bias.

4.2. Critic Pretraining

Before training the policies, the critic is pretrained using the offline dataset. The objective is to learn a stable approximation of the Q-function through temporal difference learning. Targets are computed using a bootstrapped estimate with a discount factor $\gamma = 0.99$, and optimization is performed via mean squared error minimization. This step is relevant, as all subsequent methods rely on the critic for evaluating policy quality in the absence of online interaction.

4.3. Reinforcement Learning Approach

The RL only approach follows an actor-critic framework regularized with Behavior Cloning (BC). The actor is trained to maximize the critic's Q-values while remaining close to the dataset distribution. The objective function is defined as:

$$\mathcal{L} = -\Sigma[Q(s, \pi(s))] + \alpha \cdot ||\pi(s) - adata||^2$$

where a controls the trade-off between exploitation and adherence to the dataset. Training proceeds in two stages: an initial BC-heavy phase for stable initialization, followed by a longer optimization phase with reduced regularization.

4.4 Evolutionary Algorithm Approach

The EA only approach maintains a population of candidate policies, each represented by an actor network. Policies are evaluated using a regularized fitness function that combines critic estimates with a BC penalty:

$$Fitness = Q(s, \pi(s)) - \lambda \cdot BC\ Loss$$

This formulation discourages out-of-distribution actions that may exploit inaccuracies in the critic. At each generation, the top-performing individuals, also named the elites, are selected, and new policies are generated through parameter averaging and stochastic mutation. This population-based search encourages exploration across diverse regions of the policy space.

4.5 Hybrid Lamarckian Memetic Approach

The hybrid model integrates EA and RL in a Lamarckian memetic framework, where individuals undergo both evolutionary selection and gradient-based refinement. Each generation consists of two phases:

1. The evolutionary phase, during which the population evolves using the same selection and mutation mechanisms as in the EA only method;
2. The RL fine-tuning phase, when the elites are further optimized using RL updates. Importantly, the updated parameters are retained and passed to subsequent generations, enabling learned improvements to propagate through the population.

To ensure a fair comparison, the total computational budget in terms of gradient updates is approximately matched across methods. The hybrid configuration allocates fewer generations but significantly more RL updates per elite, enabling deeper exploitation of promising solutions.

4.6 Evaluation

Performance is assessed using two complementary metrics:

1. Offline Evaluation: Average critic Q-values over sampled states, reflecting the model's estimated future reward;

2. Online Evaluation: Normalized scores and success rates obtained through interaction with the MuJoCo environment.

The final comparison across all methods is presented in *Table 1*, while the relationship between offline and online performance is visualized in *Fig. 1*.

Table 1. Evaluation scores for all three approaches

Actor	Online Norm Score	Offline Q-Score
RL	55.00	0.1587
EA	30.00	0.1630
Hybrid	60.00	0.1611

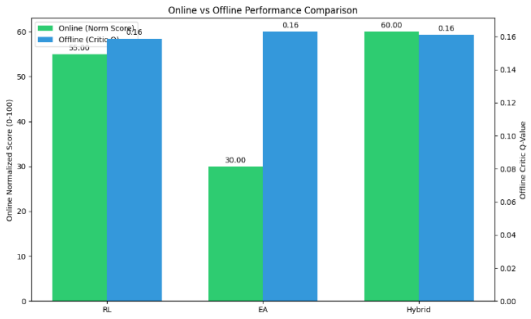

Fig. 1. Online vs Offline Performance Comparison

The training dynamics of the RL based components are illustrated in *Fig. 2*, which presents the evolution of critic Q-values and BC loss over training steps. The hybrid model demonstrates highly stable behavior that closely tracks the RL only baseline. Both models show nearly identical fluctuations in Q-values, centered around a mean of 0.16, and a synchronized, consistent decrease in BC loss from roughly 0.22 to 0.17. This suggests that during the RL fine-tuning phases, the hybrid model maintains strong adherence to the dataset distribution and does not suffer from significant distributional shift or instability typically introduced by evolutionary exploration. The overlapping nature of these curves indicates that the gradient-based updates effectively regularize the policies, ensuring that the refinement phase remains as grounded as the pure RL approach.

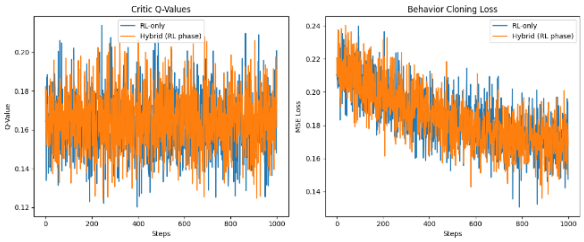

Fig. 2. Critic Q-Values and Behavior Cloning Loss

Further insight into the evolutionary component is provided in *Fig. 3*, which compares the average and maximum fitness across generations for both the EA only and hybrid approaches. Here, the hybrid approach demonstrates a clear and significant performance advantage from the onset. While the EA only method suffers from extreme volatility with average fitness dropping as low as -0.08 in early generations, the hybrid approach starts at a much higher baseline and exhibits a rapid upward trajectory.

In the "Max Fitness" plot, the hybrid model reaches a peak fitness of approximately 0.08 within just five generations, a level the EA-only model struggles to consistently reach even after fifty generations. This indicates

that RL fine-tuning acts as a powerful accelerator, quickly refining the population's policies and insulating the evolutionary process from the high-variance failures seen in the purely stochastic EA baseline. The irregularity previously noted is characteristic of the EA only baseline, whereas the hybrid model provides a more robust and efficient path to high-quality solutions.

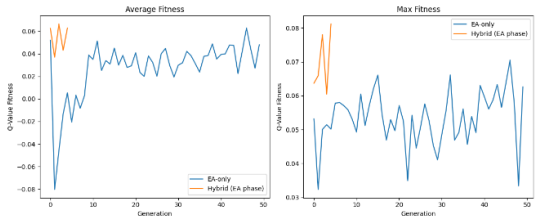

Fig. 3. Average and Maximum Fitness Across Generations

5 Results and Discussion

The experimental results reveal a clear distinction between offline estimates and actual policy performance. While all three methods achieved similar critic Q-values of approximately 0.16, their online behaviors differed substantially.

The hybrid model achieved the highest normalized score, outperforming both the RL only and EA only approaches. This confirms that the hybrid framework is more effective at translating learned value estimates into successful task execution.

Interestingly, the EA only model achieved the highest offline Q-value (0.1630), yet exhibited the weakest online performance. This discrepancy highlights the limitation of offline RL in which critic estimates can be overly optimistic, particularly when policies exploit out-of-distribution actions. Despite the inclusion of a BC penalty, the EA's reliance on critic-based evaluation appears insufficient to guarantee real-world effectiveness.

The hybrid model's superior performance, reaching a normalized online score of 60.00, underscores the efficacy of combining evolutionary exploration with gradient-based refinement. While the RL only model delivered a strong performance 55.00, it lacked the global search capabilities inherent in the EA phase, which allowed the hybrid model to escape local optima that traditional actor-critic methods often encounter in sparse-reward tasks.

Furthermore, the data suggests a critical decoupling between perceived and actual performance in offline settings. The EA only model's highest offline Q-score (0.1630) contrasted sharply with its lowest online performance (30.00). This gap suggests that while the EA successfully optimizes against the critic's value surface, that surface itself may be flawed due to the overestimation of actions not well-represented in the antmaze dataset.

In contrast, the Hybrid model achieves a more balanced profile. It maintains a competitive offline Q-score (0.1611) while securing the highest online utility. This implies that the Lamarckian memetic framework acts as a natural regularizer. By utilizing RL to fine-tune evolutionary candidates, the framework ensures that policies are not just theoretically high-value according to an optimistic critic, but are also robust enough to navigate the physical constraints of the MuJoCo simulation. Ultimately, these results demonstrate that hybridization mitigates the over-optimization trap of offline RL, offering a more reliable pathway for policy deployment in complex, sparse-reward environments.

6 Conclusions

This study set out to evaluate whether hybridizing Evolutionary Algorithms with Reinforcement Learning can improve decision-making in sparse-reward, offline settings. The results support this hypothesis. While all three approaches achieved comparable offline critic estimates, the hybrid framework consistently delivered the strongest online performance, achieving the highest normal-

ized score and success rate under similar computational constraints.

These findings highlight the complementary strengths of the two paradigms: evolutionary search promotes robust exploration and avoids premature convergence, while gradient-based updates enable efficient fine-tuning of promising solutions. The observed training dynamics further illustrate how the hybrid approach effectively balances these forces, despite increased variability during optimization.

Overall, the hybrid memetic framework demonstrates a clear advantage in overcoming local optima and sparse feedback, suggesting that such integrations are a promising direction for complex offline reinforcement learning problems.

References

[1] Bai, Hui / Cheng, Ran / Jin, Yaochu: *Evolutionary Reinforcement Learning: A Survey*. Intelligent Computing, 2023.

[2] Khadka, Shauharda / Tumer, Kagan: *Evolution-Guided Policy Gradient in Reinforcement Learning*. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 1196-1208.

[3] Grumbach, Felix / Badr, Nour Eldin Alaa / Reusch., Pascal / Trojahn, Sebastian: *A Memetic Algorithm with Reinforcement Learning for Sociotechnical Production Scheduling*. IEEE Access, 2016.

[4] Fu, Justin / Kumar, Aviral / Nachum, Ofir / Tucker, George / Levine, Sergey: *D4RL: Datasets for Deep Data-Driven Reinforcement Learning*. Berkeley Artificial Intelligence Research, 2020.

[5] Lin, Yuanguo / Lin, Fan / Cai, Guorong / Chen, Hong / Zou, Lixin / Wu, Pengcheng: *Evolutionary Reinforcement Learning: A Systematic Review and Future Directions*. Journal of Latex Class Files, 2021.

[6] Such, Felipe Petroski / Madhavan, Vashisht / Contri, Edoardo / Lehman, Joel / Stanely, Kenneth O. / Clune, Jeff: *Deep Neuroevolution: Genetic Algorithms are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning*. arXiv, 2018.

[7] Salimans, Tim / Ho, Jonathan / Chen, Xi / Sidor, Szymon / Sutskever, Ilya: *Evolution Strategies as a Scalable Alternative to Reinforcement Learning*. OpenAI, 2017.

[8] Li, Pengyi / Hao, Jianye / Tang, Hongyao / Fu, Xian / Zheng, Yan / Tang, Ke: *Bridging Evolutionary Algorithms and Reinforcement Learning: A Comprehensive Survey on Hybrid Algorithms*. IEEE Xplore, 2017.

[9] Rengarajan, Desik / Vaidya, Gargi / Sarvesh, Akshay / Kalathil, Dileep / Shakkottai, Srinivas: *Reinforcement Learning with Sparse Rewards using Guidance from Offline Demonstration*. ICLR, 2022.

[10] Fujimoto, Scott / Gu, Shixiang Shane: *A Minimalist Approach to Offline Reinforcement Learning*. arXiv, 2021.

[11] Dadashi, Robert / Rezaeifar, Shideh / Vieillard, Nino / Hussenot, Léonard / Pietquin, Olivier / Geist, Matthieu: *Offline Reinforcement Learning with Pseudometric Learning*. Proceedings of the 38th International Conference on Machine Learning, 2021.

[12] Kumar, Aviral / Zhou, Aurick / Tucker, George / Levine, Sergey: *Conservative Q-Learning for Offline Reinforcement Learning*. Neural Information Processing Systems, 2020.

[13] Kostrikov, Ilya / Nair, Ashvin / Levine, Sergey: *Offline Reinforcement Learning with Implicit Q-Learning*. ICLR, 2022.

[14] Qu, Xinghua / Ong, Yew-Soon / Hou, Yaqing / Shen, Xiaobo: *Memetic Evolution Strategy for Reinforcement Learning*. IEEE Xplore, 2019.

[15] Zou, Yuji / Hao, Jin-Kao / Wu, Qinghua: *A reinforcement learning guided hybrid evolutionary algorithm for the latency location routing problem*. ScienceDirect, 2024.

Denis-Claudiu ASAN is a graduate of the Faculty of Cybernetics, Statistics and Economic Informatics within the Bucharest Academy of Economic Studies, where he obtained his Bachelor's degree in Economic Informatics in 2024. He is currently enrolled in the Master's program in Databases – Business Support, with expected completion in 2026. His academic and research interests encompass data science, machine learning, bioinformatics, and computational neuroscience.