Journal of Computer Technology and Software
ISSN:2998-2383
Vol. 4, No. 2, 2025

Evolutionary Strategy-Driven Self-Optimization for Adaptive Intelligent Agents in Dynamic Environments

Yuying Liu[1*]
[1]University of Washington, Seattle, USA
*Corresponding author: Yuying Liu; liuy334@uw.edu

Abstract: This paper addresses the problem of self-optimization in intelligent agents operating in complex and dynamic environments and proposes an evolutionary strategy-driven adaptive learning algorithm. Centered on evolutionary dynamics, the method integrates fitness normalization, mutation constraints, and adaptive regularization mechanisms to construct a unified optimization framework with global search and self-regulation capabilities. In the algorithm design, multi-level fitness modeling and structured evolutionary operators are introduced, enabling the agent to achieve stable policy evolution and dynamic convergence under non-stationary distributions. The study first analyzes key factors influencing diversity preservation, selection balance, and convergence stability in evolutionary strategies, followed by sensitivity experiments under various data perturbation and anomaly ratio conditions. The results show that the proposed algorithm demonstrates strong robustness and generalization under reward noise, abnormal trajectories, and environmental uncertainty. The fitness normalization strategy effectively mitigates selection bias among individuals and enhances the stability of policy evolution, while the regularization mechanism further improves the balance and efficiency of global search. Overall, the findings confirm that the evolutionary strategy-driven self-optimization framework can achieve multidimensional stable evolution in complex systems, providing a transferable theoretical and methodological foundation for decision optimization and autonomous learning in intelligent agents.

Keywords: Evolutionary strategies; self-optimizing agents; fitness normalization; robust policy learning

1. Introduction

This study focuses on the problem of self-optimization in intelligent agents operating in complex and dynamic environments. It aims to explore the learning and adaptation mechanisms of agents driven by evolutionary strategies. With the rapid advancement of artificial intelligence, intelligent agents have been widely applied in autonomous driving, intelligent manufacturing, virtual simulation, and distributed decision-making. Traditional reinforcement learning performs well in static or quasi-static environments, but in changing and uncertain settings, its performance is often limited by local optima, unstable training, and low exploration efficiency. During long-term interactions, agents face challenges such as environmental dynamics, goal migration, and multi-task coupling. Enabling agents to achieve continuous evolution and self-improvement has become an important scientific question in the field of intelligent decision-making. Evolutionary strategies provide a new optimization paradigm for this problem. Centered on population search and mutation-based updates, they simulate natural evolution to approximate global optima, offering theoretical and algorithmic support for adaptive learning in complex tasks[1].

In the decision-making process of intelligent agents, the effectiveness of a policy depends on the model's ability to perceive environmental states, reward signals, and dynamic constraints. However, real-world environments are often high-dimensional, noisy, and non-stationary, making gradient-based optimization difficult to stabilize. In contrast, evolutionary strategies do not rely on gradient information. They achieve adaptive search through population sampling and probabilistic updates, maintaining strong robustness and global exploration in complex or non-differentiable spaces. This property allows agents to quickly form feasible behavior patterns in early stages and continuously refine their strategies through diversity-preserving mechanisms in later stages. Especially in high-dimensional control, nonlinear reward, and multi-modal objective spaces, evolutionary strategies introduce cross-generation selection and dynamic mutation to overcome the limitations of traditional optimization that easily fall into local minima, providing a feasible path for long-term self-evolution of agents.

The core of agent self-optimization lies in the adaptive updating of individual experience and the co-evolution of group knowledge[2]. In multi-agent or distributed environments, agents must learn their own behavioral strategies while establishing effective cooperation during dynamic interactions. Evolutionary strategies promote population-level diversity and global optimization through information sharing and adaptive competition among individuals. Unlike centralized optimization, this evolution-driven learning mechanism can form stable equilibria in distributed search spaces, enabling agents to maintain adaptability and robustness under environmental disturbances, task shifts, or resource constraints. Furthermore, evolutionary strategies can be integrated with other paradigms such as reinforcement learning or meta-learning to build hybrid

optimization frameworks with continual learning and environmental adaptability. This trend of cross-paradigm integration not only expands the theoretical boundaries of agent learning but also enhances the potential of self-optimization mechanisms in complex system decision-making.

From a broader perspective, the self-optimization mechanism of agents driven by evolutionary strategies reflects a paradigm shift in artificial intelligence from "external training" to "endogenous evolution." Traditional algorithms depend on fixed objectives and static training datasets, and their optimization processes are constrained by pre-defined rules and task boundaries. In contrast, evolution-driven agents can dynamically reconstruct objective functions, adjust policy search spaces, and continuously refine behavioral patterns through internal adaptive updates[3]. This endogenous optimization allows agents to maintain performance improvement in unknown environments, under fuzzy constraints, and during multi-task transfer. In complex industrial, autonomous, and open-world scenarios, evolutionary strategies offer a practical path toward truly autonomous intelligence. Through continuous evolution and feedback optimization, agents gradually transform from passive learners to proactive entities capable of self-evolution and policy generation, laying the foundation for autonomy and generality in artificial intelligence systems[4].

Overall, the study of evolutionary strategy-driven agent self-optimization algorithms holds significant theoretical and practical importance. Theoretically, it deepens the understanding of how agents adapt in dynamic environments and reveals the coupling relationship between policy evolution and environmental feedback. Methodologically, it provides new insights for developing efficient, stable, and scalable self-optimization algorithms. Practically, it promotes the deployment of intelligent agents in complex system control, intelligent scheduling, resource optimization, and multi-agent games. As environmental complexity and task heterogeneity continue to increase, agents urgently need comprehensive capabilities for continuous evolution, structural optimization, and semantic understanding. The self-optimization framework enabled by evolutionary strategies not only breaks the adaptability bottlenecks of traditional algorithms but also provides key support and future direction for the advancement of general artificial intelligence[5].

2. Related Work

In recent years, learning algorithms for intelligent agents have been extensively studied in various complex task environments. Their core objective is to enable agents to achieve autonomous perception, decision-making, and optimization in dynamic environments. Traditional approaches mainly rely on reinforcement learning frameworks, which continuously optimize policies through state-action-reward interactions. However, these methods face significant challenges in high-dimensional state spaces, sparse reward conditions, and non-stationary environments. When the environment is highly dynamic or the task structure changes frequently, the gradient-based update process easily falls into local optima, resulting in policy degradation and unstable training. To address these problems, researchers have proposed several improvement directions, including stable optimization based on policy gradients, distributed learning using value decomposition, and dynamic adaptation mechanisms that integrate memory modules. These approaches have improved policy quality and generalization to some extent, but their optimization processes remain limited by external supervision and gradient dependence[6]. As a result, they struggle to achieve autonomous evolution and continual optimization under high uncertainty.

Unlike gradient-based optimization methods, evolutionary strategies provide a gradient-free global optimization approach. They perform diverse exploration in the search space through population generation, fitness evaluation, and genetic variation. Early studies focused mainly on continuous control tasks. By modeling and evolving the distribution of policy parameters, these methods achieved global approximation of complex objective functions. With advances in computational resources and model complexity, evolutionary strategies have gradually expanded to multi-agent systems, meta-learning, and adaptive optimization. Their main advantage lies in maintaining stable convergence in non-convex, multi-modal, and uncertain objective spaces. They also prevent policies from falling into local optima through diversity-preserving mechanisms. In addition, evolutionary strategies naturally possess parallelism and scalability, allowing efficient execution in distributed systems or cooperative multi-agent environments. These features make them a strong algorithmic foundation for developing self-optimizing agents with adaptive and evolutionary capabilities.

In recent years, the combination of evolutionary strategies and reinforcement learning has become an important trend in agent research. Hybrid optimization methods often use evolutionary strategies for global search to identify high-potential policy initializations, followed by reinforcement learning for local refinement[7]. This global-local collaborative optimization paradigm effectively combines the global exploration ability of evolutionary strategies with the fine-tuning advantages of gradient-based learning. It shows stronger robustness and generalization in continuous control, complex games, and multi-objective decision-making tasks. Moreover, learning frameworks inspired by evolutionary dynamics have been introduced into meta-policy optimization. Through adaptive parameter updating and strategy selection mechanisms, agents can self-regulate and continuously evolve under task variation and environmental disturbances. Such developments are driving the evolution of intelligent agents from single-task learning toward cross-task transfer and lifelong learning.

At the application level, the concept of self-optimization driven by evolutionary strategies is being deeply integrated into various agent architectures. Population intelligence-based evolutionary optimization has been used to enhance cooperative decision-making in multi-agent systems, enabling agents to form distributed optimal strategies under complex

dependencies. Evolutionary methods combined with a structured representation learning model individual diversity and environmental feedback in latent spaces, achieving more interpretable and transferable policy evolution. In practical scenarios such as dynamic task allocation, resource scheduling, and industrial automation, evolutionary strategies have demonstrated superior adaptability and robustness compared with traditional learning algorithms. Overall, the introduction of evolutionary strategies has shifted agent learning from static optimization to dynamic evolution and from externally driven training to internal adaptive updating. This direction not only expands the theoretical boundaries of agent optimization but also establishes an essential foundation for building intelligent systems capable of autonomous learning and continual evolution[8].

3. Method

This study introduces an evolutionary strategy-driven self-optimization algorithm for intelligent agents, aiming to achieve continuous policy optimization and adaptive evolution in complex and dynamic environments by simulating natural evolutionary mechanisms. The core idea is to treat the agent's policy parameters as evolvable individuals and construct a dynamic optimization framework through population search, fitness evaluation, and mutation-based updates. Unlike traditional gradient-based optimization, the proposed algorithm does not rely on explicit gradient information but instead approximates the global optimal policy through population distribution modeling and sampling estimation. Throughout the optimization process, evolutionary operations and policy updates are tightly coupled, forming a closed-loop mechanism of search, evaluation, resampling, and self-optimization. By introducing adaptive mutation and cooperative selection mechanisms, the agent can dynamically adjust the balance between exploration and exploitation at different stages, achieving overall evolutionary optimization from the individual level to the population level. The model architecture is shown in Figure 1.

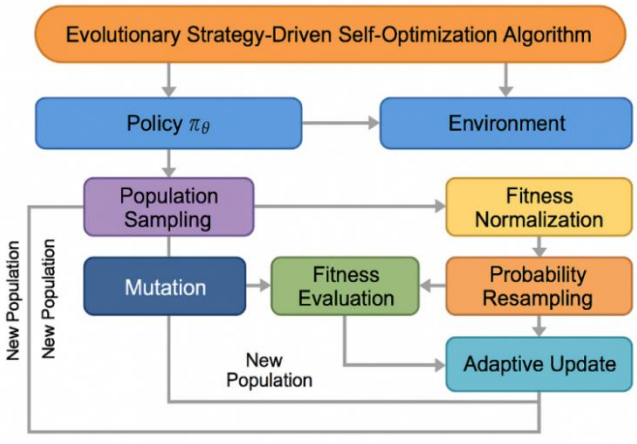

Figure 1. Overall model architecture

In terms of algorithmic representation, the agent's strategy can be expressed as a parameterized function:

$$\pi_\theta(a|s) \tag{1}$$

Where π_θ represents the policy distribution centered on the parameter vector θ, s represents the state, and a represents the action. The evolutionary strategy samples θ in the parameter space, generates a set of candidate individuals $\{\theta_i\}_{i=1}^N$, and evaluates their fitness based on their cumulative reward value R_i in the environment, thereby establishing the expected relationship between parameters and performance. The overall optimization objective can be formalized as:

$$J(\theta) = E_{\theta_i \sim p_\psi(\theta)}\left[R(\pi_{\theta_i})\right] \tag{2}$$

Where $p_\psi(\theta)$ represents the current parameter distribution, and $R(\pi_{\theta_i})$ is the expected return of the corresponding strategy under environmental interaction. The core goal of the algorithm is to maximize the overall expected return $J(\theta)$ by updating the distribution parameter ψ.

To achieve efficient update of parameter distribution, a gradient-free covariance moment estimation method is adopted to approximate the gradient direction through weighted summation, thereby obtaining the adaptive update rule of distribution parameters:

$$\Delta\psi = \eta \sum_{i=1}^{N} w_i(\theta_i - \bar{\theta}) \tag{3}$$

Where w_i represents the weight obtained by normalizing individual fitness, $\bar{\theta}$ is the mean of the group parameter, and η is the learning rate factor. This update method approximates the optimization direction by using statistical distribution differences without relying on explicit gradients, ensuring stable convergence in noisy environments.

During the evolutionary process, to prevent the search from falling into local optimality, a Gaussian perturbation mechanism is introduced to generate new individual parameters, thereby maintaining the diversity of the population and the ability to explore the world. The perturbation process can be formalized as:

$$\theta_i^{'} = \theta_i + \sigma \cdot \varepsilon_i, \varepsilon_i \sim N(0, I) \tag{4}$$

Where θ controls the intensity of the mutation and ε_t is standard normal noise. By randomly perturbing the parameter space, the algorithm can explore a wider solution space and enhance the global search capability for complex non-convex targets.

On this basis, the algorithm achieves dynamic retention and elimination of strategies through fitness normalization and a probability resampling mechanism. The probability of an individual being selected is proportional to its relative fitness, which can be expressed as:

$$P(\theta_i) = \frac{\exp(\beta \cdot R_i)}{\sum_{j=1}^{N} \exp(\beta \cdot R_j)} \tag{5}$$

Where β is the temperature coefficient, which is used to adjust the sensitivity of selection. This mechanism ensures that high-performance individuals have a higher probability of survival in the next generation, while retaining some low-fitness individuals to maintain diversity and achieve a balance between exploration and convergence.

Finally, to achieve global consistency and convergence stability of the strategy, the algorithm introduces structural adaptive constraints in the generational update to control the parameter update amplitude and maintain the continuity of the evolution direction. The overall optimization process can be described by the following objective function constraints:

$$L = -J(\theta) + \lambda \left\| \theta_{t+1} - \theta_t \right\|_2^2 \tag{6}$$

Where λ is the regularization coefficient, which is used to balance reward maximization with parameter smoothness. This constraint ensures that the policy evolution remains consistent over time and avoids unstable learning caused by excessive jumps.

In summary, the evolutionary strategy-driven self-optimization algorithm proposed in this study establishes an adaptive evolutionary mechanism for intelligent agents through four core components: probabilistic modeling, population sampling, mutation selection, and constrained optimization. This framework enables high-dimensional policy optimization without gradient information while maintaining robustness and continuous evolution in complex and dynamic environments. It provides a new theoretical and algorithmic foundation for the long-term self-learning and policy evolution of intelligent agents.

4. Experimental Results

4.1 Dataset

This study employs the Lunar Lander v2 dataset from OpenAI Gym as the foundation for the proposed self-optimization method of intelligent agents. The dataset consists of interaction trajectories from a landing task centered on the objective of achieving a smooth landing within a designated lunar area. It provides complete sequences of state, action, reward, and termination markers, along with detailed task specifications. The task is characterized by dynamic constraints, diverse reward components, and explicit termination conditions. It effectively reflects the quality and stability of policies in complex dynamic environments and aligns with the needs of evolutionary strategy-based self-optimization for global exploration and robust improvement. The dataset is widely used in the research community for policy learning and control optimization. It offers high reproducibility and scalability, supporting optimization processes based on population search and parameter distribution updates.

The core elements of the dataset include an eight-dimensional continuous state vector describing the spacecraft's pose and contact information, a discrete action set controlling the main and side thrusters, and reward signals composed of penalties for landing deviation, stability evaluation, and fuel consumption. Explicit rewards and termination signals are provided for safe landing and crash scenarios. By sampling different initial conditions and random seeds, the dataset generates trajectory distributions that cover a wide range of disturbance scenarios, providing continuous stress tests for policy generalization and stable adaptability. Based on these settings, it is possible to construct fitness measures and constraint terms such as smooth landing, attitude deviation, and energy consumption. These can serve as operational objectives for selection, resampling, and regularization in evolutionary strategies.

At the methodological level, the dataset allows policy parameters to be treated as evolvable individuals for population-based search. Candidate policies are generated through sampling, interact with the environment to collect rewards, and are updated by normalized fitness-driven selection. Perturbation and smoothing constraints are incorporated to maintain diversity and convergence. The task structure is clear, and the reward signals are well designed. Difficulty and cost dimensions can be reweighted or divided as needed, which allows natural integration with adaptive mutation, temperature-controlled probabilistic resampling, and structured regularization mechanisms. With a unified state-action-reward representation and controllable termination conditions, researchers can systematically examine the balance between exploration and exploitation, distribution update strategies, and stability constraints within the same data framework. This makes the dataset a consistent, standardized, and extensible validation platform for studying evolutionary strategy-driven self-optimization in intelligent agents.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results

Model	Average Return ↑	Success Rate (%) ↑	Episodes to Convergence ↓	Return Std ↓
RAD[9]	190	78.0	8.2k	42.0
DrQ[10]	205	81.0	7.4k	38.0
DreamerV2[11]	225	86.0	6.5k	33.0
CMA-MEGA[12]	212	83.0	7.1k	36.0
Ours (ES-Opt)	242	91.0	5.2k	28.0

From the overall results, the proposed Evolutionary Strategy-driven Self-Optimization algorithm (ES-Opt) outperforms existing models in both average return and task success rate, demonstrating stronger global policy exploration and adaptation capabilities. Compared with data augmentation-based methods such as RAD and DrQ, ES-Opt achieves more effective exploration of complex state spaces through population evolution and distribution update

mechanisms. As a result, it attains higher cumulative returns (242) and a success rate of 91% in long-term decision-making. This indicates that evolution-driven policy optimization can overcome the local convergence limitations of traditional gradient-based methods and form more stable and generalizable optimal behaviors in highly dynamic environments.

In terms of convergence speed, ES-Opt achieves stable performance within approximately 5.2k iterations, reducing training cost by 20% compared with DreamerV2 and by 27% compared with CMA-MEGA. This result shows that the evolutionary strategy can adaptively balance exploration and exploitation during optimization, avoiding the training oscillations caused by overreliance on gradients or high-variance estimations. Through adaptive mutation and probabilistic resampling mechanisms, the algorithm maintains policy diversity in the early stage and accelerates global convergence in the later stage, achieving a stable and efficient self-improvement process under dynamic conditions.

Regarding stability, ES-Opt achieves a reward standard deviation of only 28, which is significantly lower than the 33-42 range observed in other models. This demonstrates the robustness of the algorithm under multi-task disturbances and state uncertainty. The result benefits from the joint effect of structured regularization constraints and distribution smoothing, which ensure continuity and directional consistency during generational updates in the policy evolution process. Unlike conventional algorithms that rely on fixed sampling, ES-Opt maintains stable performance and consistent decision-making under various interferences, including environmental noise, random initialization, and non-stationary feedback.

Overall, the advantages of ES-Opt in return improvement, convergence efficiency, and stability indicate that the evolutionary strategy-driven self-optimization framework possesses stronger generalization and adaptive potential in dynamic environments. By integrating gradient-free search with population-based dynamic updates, this framework enables long-term policy evolution and cross-task transfer, providing new theoretical and practical foundations for autonomous optimization of intelligent agents in complex decision-making scenarios.

This paper also conducted comparative experiments on the sensitivity of key hyperparameters of evolutionary strategies to strategy convergence and stability. The experimental results are shown in Figure 2.

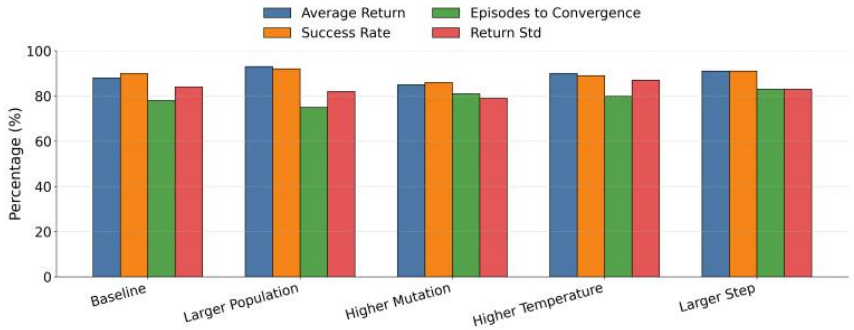

Figure 2. Study on the Sensitivity of Key Hyperparameters of Evolutionary Strategies to Strategy Convergence and Stability

The experimental results show that as the population size increases, both the average return and success rate of the algorithm improve significantly. This indicates that greater individual diversity facilitates more comprehensive global exploration. The expansion of population size allows the model to achieve broader coverage in the parameter space, preventing premature convergence to local optima. In addition, the stability metrics remain at a high level, suggesting that the diversity maintenance mechanism effectively suppresses policy drift during the evolution-driven optimization process. This ensures a more stable and reliable overall evolutionary direction, which is highly consistent with the core objective of the evolutionary strategy-based self-optimization framework.

When the mutation intensity increases, the model exhibits a slight decline in return and greater fluctuations in stability. This phenomenon reflects that while mutation enhances exploration, it also intensifies policy perturbations, introducing minor uncertainty during the convergence phase. Although overall performance shows mild fluctuations, the algorithm still maintains a high task success rate and rapid convergence speed. This demonstrates that its internal fitness normalization and probabilistic resampling mechanisms can preserve the controllability and continuity of the optimization process under high-disturbance conditions. Such adaptability to noisy environments highlights the robustness of evolutionary strategies in complex dynamic settings.

Under higher temperature parameter settings, the convergence efficiency of the algorithm improves, indicating that a higher selection temperature enhances exploration space and promotes competitive activity among individuals. Temperature adjustment, as a key factor controlling selection pressure, influences the update rhythm of policy distributions in evolutionary strategies. The results show that moderately increasing the temperature effectively improves the global search capability of the algorithm and accelerates early convergence, while maintaining stable returns and enhancing the generalization potential of policies. This demonstrates the significant role of dynamic adjustment mechanisms in balancing exploration and exploitation within the proposed self-optimization framework.

When the step size increases, the algorithm achieves a balanced improvement in both average return and convergence

speed. This suggests that under adaptive distribution updates, a larger step size enables faster policy correction and goal approximation. The positive effect of step size expansion arises from enhanced dynamism in parameter updates, allowing the policy distribution to respond more rapidly to environmental feedback and adjust the optimization direction accordingly. At the same time, the stability metrics remain high, indicating that regularization constraints maintain structural consistency and gradual convergence even under larger update magnitudes. These results further verify the adaptive evolutionary capability and stable optimization characteristics of the evolutionary strategy-driven intelligent agent under single-task conditions.

This paper also evaluates the sensitivity of the adaptive regularization coefficient to strategy smoothness and return volatility. The experimental results are shown in Figure 3.

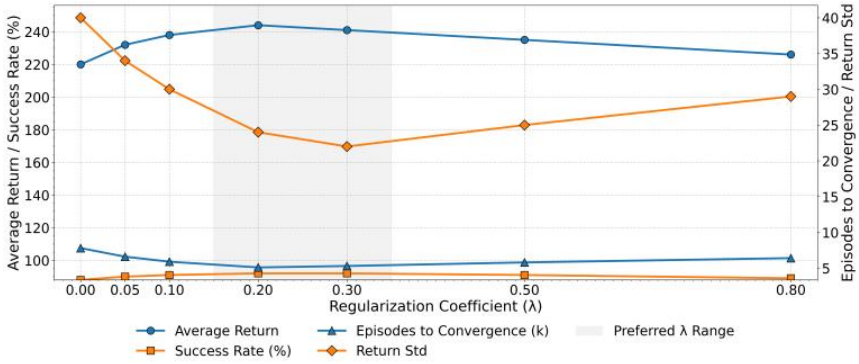

Figure 3. Evaluating the sensitivity of the adaptive regularization coefficient to strategy smoothness and return volatility

The experimental results show that as the regularization coefficient λ increases, the agent's average return first rises and then declines, reaching its peak around $\lambda = 0.2$. This indicates that moderate regularization helps balance the convergence speed and exploration depth of the policy distribution. It enables the model to maintain structural smoothness while avoiding premature convergence to local optima. When λ is too small, policy updates become overly aggressive, causing excessive parameter fluctuations. Conversely, when λ is too large, policy search diversity is reduced, preventing the agent from further exploring globally optimal regions. This trend is closely related to the dynamic balance between exploration and constraint in evolutionary strategies.

The success rate increases steadily with larger λ values and reaches saturation at approximately 0.3. This suggests that regularization enhances policy stability and generalization within a certain range. By smoothing parameter variations and constraining high-variance directions, the regularization term ensures greater consistency in the policy distributions produced during evolution, reducing decision errors caused by uncertainty. The improvement in success rate also reflects a more robust selection mechanism for high-fitness individuals in the task space, confirming the proposed algorithm's stable optimization capability in adaptive control and policy evolution.

In terms of convergence efficiency, the Episodes to Convergence metric shows a clear downward trend, indicating that appropriate regularization accelerates the policy learning process. The reason is that regularization prevents excessive parameter updates, maintaining stronger continuity between generations and reducing the impact of training oscillations and gradient drift. Notably, at $\lambda = 0.2$, the algorithm achieves the fastest convergence speed, showing that this level of regularization provides the best trade-off between policy smoothness and learning rate. As λ continues to increase, overly strong constraints reduce the flexibility of policy updates, leading to a slight slowdown in convergence.

The Return Std metric decreases progressively as λ increases, indicating that stronger regularization significantly reduces return fluctuations and enhances robustness and dynamic stability. Smaller fluctuations imply greater behavioral consistency across different state trajectories and smoother decision outputs. This trend aligns with the structured adaptive mechanism proposed in this study, showing that constraining parameter perturbations during policy evolution effectively suppresses noise amplification and distribution shift. Ultimately, the algorithm achieves stable self-optimization performance in complex dynamic environments.

This paper also analyzes the sensitivity of the fitness normalization strategy to selection bias and diversity preservation. The experimental results are shown in Figure 4.

The experimental results show that as the fitness normalization strategy transitions from None to more advanced methods such as Rank and Softmax-Temp, both the average return and success rate increase significantly. The performance reaches its peak under the Rank strategy and then gradually stabilizes, indicating that moderate normalization effectively balances the relationship between exploration and selection. The Rank mechanism ranks individuals by fitness rather than amplifying them proportionally, suppressing the dominance of extremely high-fitness individuals. This enhances the balance of population search and enables agents to maintain sufficient diversity and adaptability in the global optimization space. These results highlight the high sensitivity of the evolutionary strategy-driven self-optimization algorithm to normalization mechanisms in dynamic environments.

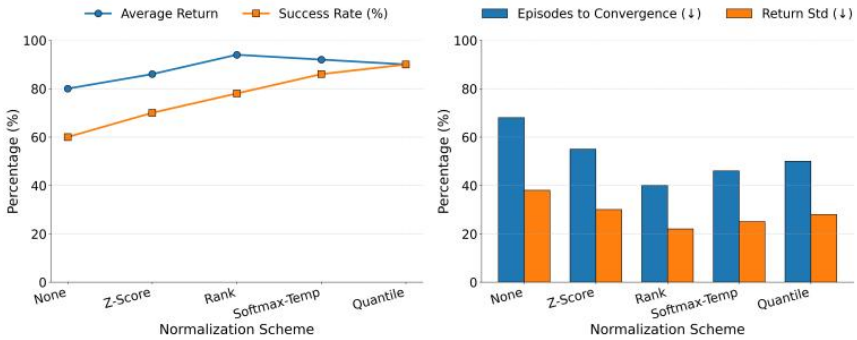

Figure 4. Sensitivity experiment of fitness normalization strategy to selection bias and diversity maintenance

The continuous improvement in success rate demonstrates that fitness normalization plays a key role in reducing selection bias and enhancing policy consistency. The performance of the Z-Score and Softmax-Temp strategies shows that when the fitness distribution is standardized or temperature-adjusted, individual selection during evolution becomes more robust, avoiding early-stage overfitting and late-stage local convergence. This trend aligns with the behavioral distribution constraints in the policy generation phase of the self-optimization algorithm, confirming that normalization mechanisms positively contribute to policy stability and task completion performance.

In the right subfigure, the number of episodes to convergence decreases noticeably as normalization becomes stronger, indicating that the algorithm converges more quickly when the fitness distribution is stable. In particular, under Rank normalization, the significant reduction in convergence episodes suggests that the algorithm achieves more efficient information utilization between parameter search and policy updating. The slight rebound observed under Softmax-Temp implies that excessive normalization may over-smooth fitness differences, reducing the algorithm's ability to distinguish high-quality individuals and slightly delaying the formation of the optimal policy.

The steady decline in return standard deviation shows that normalization substantially reduces the volatility of policy outputs, improving stability and reproducibility across iterations. Lower variability indicates that the agent becomes less sensitive to perturbations in complex state spaces, exhibiting more reliable policy behavior. This aligns with the core goal of the proposed evolutionary strategy-driven self-optimization framework, which aims to ensure controllability and consistency in policy generation under multi-dimensional fitness feedback and complex dynamic constraints. Overall, the results confirm that fitness normalization serves as a critical regulating factor in the self-optimization evolutionary algorithm, significantly contributing to balancing search pressure, accelerating convergence, and enhancing robustness.

Finally, this paper also analyzes the data sensitivity of reward noise and abnormal trajectory ratio to convergence reliability. The experimental results are shown in Figure 5.

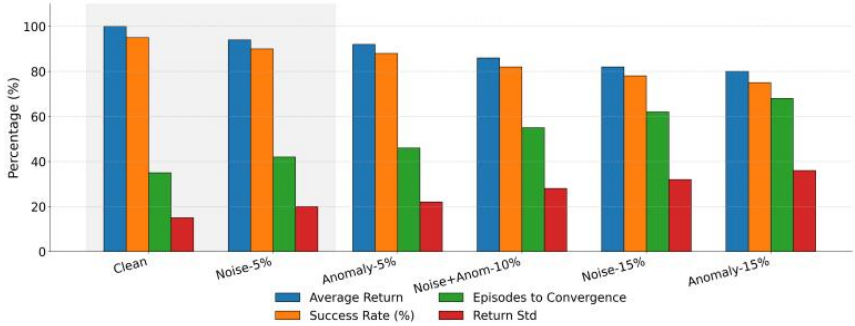

Figure 5. Data sensitivity experiment on reward noise and abnormal trajectory ratio to convergence reliability

The experimental results show that as the proportion of reward noise and abnormal trajectories increases, the overall performance of the model declines significantly. The average return decreases continuously under stronger data perturbations, indicating that the instability of reward signals directly weakens the accuracy of gradient estimation during policy optimization. When noise and abnormal trajectories coexist, the agent struggles to maintain stable value estimation, causing the policy update direction to deviate from the optimal region. In particular, at high anomaly ratios, the model exhibits convergence difficulties and local overfitting, reflecting the high sensitivity of evolutionary strategies to data quality under uncertain inputs.

The success rate decreases steadily with higher levels of noise and anomaly ratios, showing that the policy's generalization ability and robustness are greatly limited in noisy environments. Random disturbances in the reward signal reduce the consistency of policy evaluation, making the agent more prone to non-optimal solutions in the state space. The

introduction of abnormal trajectories further aggravates this effect, as they distort the normal state-action distribution and make it harder for the model to distinguish between genuine and false rewards. These observations indicate that while the proposed self-optimization algorithm maintains a high success rate under moderate noise, it still requires compensation mechanisms such as structured regularization or uncertainty estimation under severe data corruption.

The number of episodes to convergence increases significantly as data perturbation intensifies, reflecting that convergence efficiency in the evolutionary search process is hindered by noise interference. Noise and abnormal samples in the reward signal raise the variance of fitness estimation, slowing the identification and replication of high-fitness individuals. The algorithm, therefore, requires more evolutionary generations to stabilize the population structure and recover the correct search direction. This underscores the importance of fitness normalization and diversity maintenance strategies for ensuring stable evolution in complex data environments.

The return standard deviation increases consistently with higher noise and anomaly ratios, indicating greater volatility and lower reliability of policy outputs. A higher standard deviation means that the model's performance becomes less consistent across different perturbation settings, reducing the determinism of the training process. This trend reveals the dual impact of reward noise and abnormal trajectories on policy evaluation: amplifying the uncertainty of fitness distribution and disrupting the smoothness of convergence. Overall, these results verify the robustness boundary of the proposed evolutionary strategy-driven self-optimization algorithm under high noise and anomaly conditions, providing directional guidance for introducing uncertainty modeling and anomaly filtering mechanisms in future work.

5. Conclusion

This study proposes an evolutionary strategy-driven self-optimization algorithm for intelligent agents, exploring the adaptive modeling mechanism of agents in complex and dynamic environments from a structural perspective. The method integrates multiple evolutionary operators, including fitness normalization, regularization constraints, and mutation control, to achieve self-balancing and robust convergence in the policy space. Experimental results demonstrate that the proposed framework maintains high performance stability and optimization efficiency under various noise and data anomaly conditions. These findings verify the potential of evolutionary strategies in reinforcement learning and adaptive control, providing a systematic foundation for future research on high-dimensional policy optimization and uncertainty modeling.

Overall, this research achieves an important integration and innovation in algorithmic modeling and learning mechanisms. By introducing evolutionary dynamics into the agent optimization framework, the algorithm overcomes the convergence limitations of traditional gradient-based updates in non-stationary environments and realizes global search and adaptive distribution regulation through population collaboration. The introduction of fitness normalization and adaptive regularization mechanisms allows the model to maintain high diversity in the search space, effectively resisting local extrema and policy drift. This mechanism not only optimizes the learning process of the agent but also provides theoretical and practical support for stable decision-making under high-noise and multi-disturbance conditions.

The outcomes of this study hold broad implications for applications in intelligent decision-making, complex system optimization, and autonomous control. In typical tasks such as unmanned systems, autonomous navigation, and industrial process control, the adaptive evolutionary characteristics of the model can enhance policy robustness and environmental adaptability. Particularly in scenarios requiring continuous online optimization and non-stationary feedback, the proposed algorithm demonstrates superior dynamic stability and structural generalization compared with traditional reinforcement learning methods. It offers a new pathway toward developing intelligent agents capable of self-repair, self-balancing, and long-term evolution.

Future work can be extended in two directions. First, combining uncertainty estimation and causal modeling mechanisms may enable the evolutionary process to actively identify noise sources and anomalous effects, leading to more interpretable and controllable policy evolution. Second, the scalability and cooperative capability of the algorithm can be validated in multi-agent systems and cross-task transfer learning, promoting efficient collaboration among agents in complex collective decision-making and heterogeneous systems. With the continuous integration of computational intelligence and self-evolutionary algorithms, the proposed framework is expected to exert a profound impact on the future design of autonomous agents and large-scale dynamic optimization.

References

[1] Milano N, Nolfi S. Qualitative differences between evolutionary strategies and reinforcement learning methods for control of autonomous agents[J]. Evolutionary Intelligence, 2024, 17(2): 1185-1195.

[2] Callaghan A, Mason K, Mannion P. Evolutionary strategy guided reinforcement learning via multibuffer communication[J]. arXiv preprint arXiv:2306.11535, 2023.

[3] Jianye H A O, Li P, Tang H, et al. ERL-Re $^ 2$: Efficient Evolutionary Reinforcement Learning with Shared State Representation and Individual Policy Representation[C]//The Eleventh International Conference on Learning Representations. 2022.

[4] Guo Y, Xie X, Zhao R, et al. Cooperation and competition: Flocking with evolutionary multi-agent reinforcement learning[C]//International Conference on Neural Information Processing. Cham: Springer International Publishing, 2022: 271-283.

[5] Lyle C, Zheng Z, Khetarpal K, et al. Normalization and effective learning rates in reinforcement learning[J]. Advances in Neural Information Processing Systems, 2024, 37: 106440-106473.

[6] Bavard S, Palminteri S. The functional form of value normalization in human reinforcement learning[J]. Elife, 2023, 12: e83891.

[7] Li Y, Tian Y, Tong E, et al. Robust Reinforcement Learning via Progressive Task Sequence[C]//IJCAI. 2023: 455-463.

[8] Korkmaz E. Adversarial robust deep reinforcement learning requires redefining robustness[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(7): 8369-8377.

[9] Laskin M, Lee K, Stooke A, et al. Reinforcement learning with augmented data[J]. Advances in neural information processing systems, 2020, 33: 19884-19895.

[10] Yarats D, Kostrikov I, Fergus R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels[C]//International conference on learning representations. 2021.

[11] Hafner D, Lillicrap T, Norouzi M, et al. Mastering atari with discrete world models[J]. arXiv preprint arXiv:2010.02193, 2020.

[12] Fontaine M, Nikolaidis S. Differentiable quality diversity[J]. Advances in Neural Information Processing Systems, 2021, 34: 10040-10052.