Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37323

Performance Evaluation of Exploration Strategies on Learning Efficiency and Stability in Q-Learning

Xinhao Zhang

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
u430026278@mail.bnbu.edu.cn

Abstract. A fundamental problem in reinforcement learning is the trade-off between exploration and exploitation. The quality of the final policy is not the only factor that varies with different exploration strategies, so do learning speed and result stability. Based on Q-learning, this study compares three exploration strategies in a custom 5×5 GridWorld environment: fixed ε=0.10, fixed ε=0.30, and linearly decaying ε (1.00→0.05). A baseline experiment is first performed with slip_prob=0.10. The environmental noise is then extended to three levels, 0.00, 0.10, and 0.20, and the number of episodes required for the rolling 50-episode success rate to first reach 50% is used to measure learning speed. The results show that the fixed lower exploration rate learns faster overall under all three noise levels and demonstrates better stability across random seeds. Decaying exploration achieves a higher final success rate under low-noise conditions, but converges much more slowly in the early stage. The fixed higher exploration rate performs relatively poorly overall and has particular difficulty forming a stable policy in high-noise environments. These experiments show that exploration strategies should not be evaluated only by final success rate; environmental randomness, learning speed, and stability across random seeds should also be considered.

Keywords: Reinforcement learning, Q-learning, Exploration-exploitation balance, Learning efficiency, Stability

1. Introduction

The basic goal of reinforcement learning is to allow an agent to gradually develop a behavioral policy that can obtain a high cumulative return through continuous interaction with the environment. Q-learning is a typical model-free temporal-difference algorithm. It approximates the optimal action-value function by continuously updating state-action values and can converge under certain conditions [1]. In this process, the agent must simultaneously manage the trade-off between exploration and exploitation. Exploration can help the agent discover states and actions that have not been sufficiently tried, while exploitation relies on the current value estimates to choose better actions. If exploration is insufficient, the agent may prematurely settle on a locally optimal path; if exploration is excessive, the learned effective strategies may be continuously disrupted by random actions. Therefore, the choice of exploration strategy directly affects how an agent balances the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37323

acquisition of new information with the exploitation of existing knowledge [2]. Its effects on learning speed and training stability are evaluated experimentally in this study.

ε-greedy remains a commonly used basic exploration method in Q-learning and many value-based reinforcement learning experiments due to its simple structure, ease of interpretation, and implementation. The advantage of a fixed ε is that its parameter meaning is straightforward, but the same exploration rate is used throughout the training process, making it difficult to balance early-stage information acquisition with later-stage stable exploitation. Addressing the limitations of fixed random exploration, Fortunato et al. proposed NoisyNet, which generates exploratory behavior through learnable parameter noise [3]. Unlike directly perturbing parameters, Burda et al. used the prediction error of random network distillation to construct intrinsic rewards, turning exploration from 'random attempts' into active visits to novel states [4]. Agent57 further enhances exploratory capability in complex tasks through multiple strategies with varying degrees of exploration and an adaptive selection mechanism. This shows that modern exploration mechanisms have expanded from simple random actions to different approaches such as parameter perturbation, intrinsic motivation, and policy scheduling [5]. Although these methods use different mechanisms, their common goal is to increase effective state coverage while reducing ineffective randomness.

In recent years, research has paid greater attention to the adaptiveness and sample efficiency of the exploration process. Fang et al. used adaptive task generation to alleviate hard-exploration problems, allowing the agent to gradually encounter more challenging training tasks [6]. Zhang et al. proposed MADE, which adjusts the direction of exploration according to how far current behavior deviates from already explored regions [7]. The former mainly promotes exploration by changing the training tasks, while the latter directly adjusts behavior based on explored regions. Although they act at different points in the learning process, both emphasize that exploration should change dynamically with the learning state. Related surveys classify modern exploration methods into approaches that reward novel states, reward diverse behaviors, goal-based methods, probabilistic methods, imitation-based methods, safe exploration, and random-based methods. They also identify sparse rewards, scalability, and the exploration-exploitation dilemma as important challenges in reinforcement learning [8]. Chen et al. used constrained optimization to automatically adjust the weight between intrinsic and extrinsic rewards, further reflecting the idea of adaptive exploration intensity [9]. Hao et al. systematically reviewed exploration methods from single-agent to multi-agent tasks and emphasized that noise and complex state spaces can amplify the effects of exploration strategy selection [10]. Further studies have designed exploration rewards based on state-difference metrics in order to encourage agents to visit states with greater information value [11, 12]. Other work uses structural information to guide exploration and reduce dependence on purely random actions [13]. This will increase the efficiency of exploration and decrease the damage of ineffective exploration to learning efficiency and stability.

But the majority of the above studies are related to deep reinforcement learning, sparse rewards, or more complicated high dimensional tasks. Even for small, completely controllable, discrete environments, it is still important to ask and answer a fundamental question in an intuitive way, using experiments: what is the effect of varying ε on learning speed, final performance, and stability across multiple random seeds? When only average success rates at the end of training are compared, it is possible to miss that some strategies have a higher success rate at the end of training, but take longer to get to a usable level; and averages may also mask large fluctuations between different random seeds. Hence, this study compares three strategies: fixed ε=0.10, fixed ε=0.30, and linearly decaying ε (1.00→0.05), and investigates the impact of different levels of environmental randomness (action slip probability) on the conclusions.

DOI: 10.54254/2755-2721/2026.37323

Based on this background, this study mainly addresses three questions. First, compared to a decaying exploration rate, which strategy can learn effective paths faster, and which has a higher final success rate? Second, when environmental randomness increases, does the relative performance among the three exploration strategies change? Third, beyond the final success rate, are there noticeable trade-offs between learning speed and stability across random seeds for different strategies? This study aims to illustrate through a reproducible small GridWorld experiment that the evaluation of exploration strategies should consider final outcomes, the learning process, and result stability simultaneously, rather than making judgments based on a single metric.

2. Research methods

2.1. Data source and experimental environment

This study does not use external public datasets, but instead generates interaction data through a reproducible simulation environment. The experiment uses a custom 5×5 GridWorld environment, which can be regarded as a finite Markov decision process (MDP). The starting point is at the top-left corner, the goal is at the bottom-right corner, and several failure cells are set in between; the state is represented by the agent's position in the grid, so the state space size is 25, and the action space includes four discrete actions: up, right, down, and left. Reaching the goal cell will yield a positive reward and be considered a successful termination; entering a failure cell will yield a negative reward and be considered a failed termination; all other steps will receive a slight negative reward to limit path length and reduce meaningless movements. This environment has a simple structure and clear evaluation criteria, allowing for a clearer observation of the impact of exploration strategies on the learning process while controlling for other variables.

For each random seed, the agent interacts with the environment for 1200 episodes, with a maximum of 80 steps per episode, and with the corresponding ε-greedy strategy. The experiment stores the success or failure of each episode (success), the number of steps in each episode (steps), and the learning curve statistics based on these. To account for the uncertainty in the state transitions, an action slip probability, slip_prob, is added, which means that the action that is intended to be taken has a probability of being replaced by an action in an adjacent direction. The same controlled experiments are repeated at each level of slip_prob (0.00, 0.10, 0.20) for the extended experiment to facilitate comparison.

2.2. Algorithm and parameter settings

The experiment uses Q-learning, and its core update rule is shown in Equation (1) [1].

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_(t+1) + \gamma max_a Q(s_(t+1), a) - Q(s_t, a_t)] \quad (1)$$

where $Q(s_t, a_t)$ denotes the action value of state s_t and action a_t; r_{t+1} is the immediate reward after the transition; α is the learning rate; γ is the discount factor; and $\max_a Q(s_{t+1}, a)$ denotes the maximum estimated action value at the next state.

The learning rate is set to α=0.15, the discount factor is set to γ=0.95, the number of training episodes is 1200, and the maximum number of steps per episode is 80. To reduce the effect of randomness from a single run, each experiment is independently run under 20 random seeds, and the mean and standard deviation across random seeds are used as the main statistics.

DOI: 10.54254/2755-2721/2026.37323

2.3. Exploration strategies and control groups

Three exploration strategies were used in this experiment. The first had a fixed ε=0.10, meaning that it explored less and relied more on its current value estimates; the second had a fixed ε=0.30, meaning that it explored more and relied less on its current value estimates; the third had a linearly decreasing ε, meaning that it started with ε=1.00, and then gradually decreased to 0.05, reflecting the idea of exploring more at the beginning of training and then exploiting more. The three groups of experiments kept the environment structure, number of training episodes, reward settings, and Q-learning parameters unchanged, with only the exploration strategy adjusted, thereby increasing the interpretability of the comparison.

2.4. Evaluation metrics

The experiment evaluates performance from three dimensions: final performance, learning speed, and stability. Final performance is measured by the average success rate over the last 100 episodes to represent the overall performance of the policy in the late stage of training. A 100-episode statistical window is used because it accounts for about 8.3% of the total training budget. Compared with the last 50 episodes, it better smooths random fluctuations across individual episodes; compared with the last 200 episodes, it is less likely to include too much of the middle-to-late training stage that may still be changing. It therefore provides a more suitable representation of overall performance near the end of training. Learning speed is measured by the number of episodes required for the rolling 50-episode success rate to first reach 50%; a smaller value indicates that a usable policy is formed earlier. Stability is mainly evaluated using the standard deviation of final success rate across seeds and the proportion of random seeds that reach the 50% threshold, so that differences among random runs are not ignored by relying only on averages.

3. Experimental results and analysis

3.1. Baseline experiment

In the baseline experiment, the environmental noise is fixed at slip_prob=0.10 to compare the basic differences among the three exploration strategies under the same level of randomness. Table 1 summarizes the final success rates over the last 100 episodes. The average success rate of fixed ε=0.10 is 0.581, slightly higher than 0.560 for decaying ε, while fixed ε=0.30 reaches only 0.325. The fixed higher exploration rate continues to introduce more random actions in the later stage of training and therefore fails to form a policy with the same level of stability in the current environment.

Table 1. Baseline experimental results (mean±std, across random seeds)

Exploration strategy	Mean final success rate	Standard deviation	Description
Fixed ε=0.10	0.581	0.084	Exploitation-oriented
Fixed ε=0.30	0.325	0.091	Exploration-oriented
Decaying ε 1.00→0.05	0.560	0.101	Explore first, exploit later

As shown in Figure 1, fixed ε=0.10 rapidly improves its success rate early in training and enters a relatively stable plateau sooner. The learning curve of fixed ε=0.30 also rises in the early stage, but

DOI: 10.54254/2755-2721/2026.37323

remains at a lower overall level. The decaying ε clearly exhibits a 'slow start' pattern: exploration is very active in the early stages, but the success rate remains low for a long time, and then gradually increases as ε decreases. The shaded areas and mean curves combined suggest that the differences between the three strategies are not only at the end of training, but throughout the entire learning process.

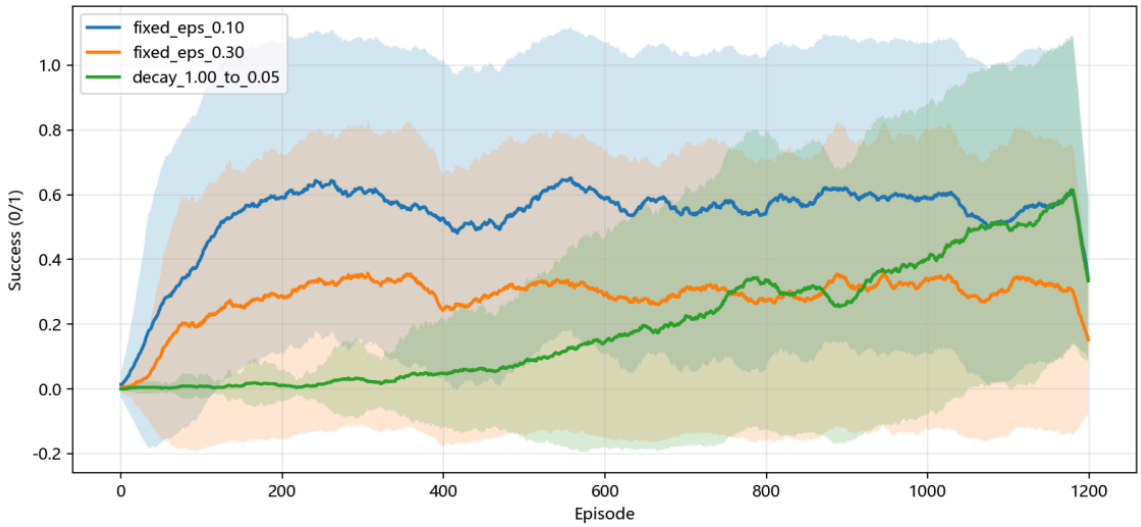

Figure 1. Baseline learning curve: success rate (mean±std, across random seeds) (picture credit: original)

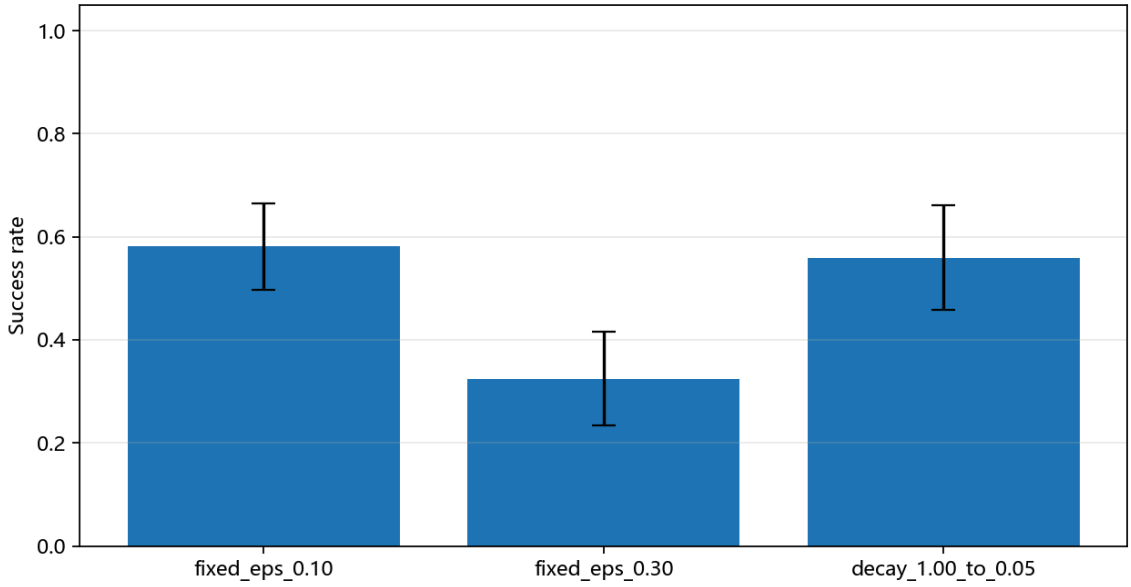

Figure 2. Comparison of final success rates (last 100 episodes, mean±std) (picture credit: original)

The average success rates over the last 100 episodes are also compared in Figure 2. The fixed ε=0.10 and decaying ε strategies show similar final performance, but follow different learning trajectories. With a fixed ε of 0.10, the agent enters the exploitation phase earlier and can achieve

DOI: 10.54254/2755-2721/2026.37323

usable performance more quickly, while a decaying ε allows the agent to explore more in the beginning, leaving more room for strategy improvements later on. This outcome demonstrates that a comparison of the final success rate alone can obscure clear differences in training efficiency.

3.2. Extended experiments under different environmental noise levels

The action slip probability was then explored at three levels: 0.00, 0.10, and 0.20, based on the baseline experiment. The final success rate, number of episodes needed to achieve a 50% rolling success rate and proportion of seeds reaching the threshold were also noted. As seen in Table 2, the final success rates of all three strategies generally decrease with an increase in environmental noise, suggesting that the uncertainty in action execution systematically increases the difficulty of learning. Meanwhile, the extent of performance drop varies across the different exploration strategies.

Table 2. Summary of extended experimental results (mean±std, across random seeds)

Environmental noise	Exploration strategy	Final success rate	Episodes to reach 50%	Proportion of seeds reaching threshold
slip=0.00	Fixed ε=0.10	0.825 ± 0.047	27.6 ± 3.3	1.00
slip=0.00	Fixed ε=0.30	0.511 ± 0.072	75.3 ± 34.2	1.00
slip=0.00	Decaying ε 1.00→0.05	0.889 ± 0.041	743.2 ± 92.8	1.00
slip=0.10	Fixed ε=0.10	0.581 ± 0.084	116.8 ± 46.0	1.00
slip=0.10	Fixed ε=0.30	0.325 ± 0.091	326.4 ± 213.5	1.00
slip=0.10	Decaying ε 1.00→0.05	0.560 ± 0.101	951.1 ± 106.6	1.00
slip=0.20	Fixed ε=0.10	0.373 ± 0.059	397.3 ± 209.5	1.00
slip=0.20	Fixed ε=0.30	0.197 ± 0.063	557.5 ± 7.5	0.10
slip=0.20	Decaying ε 1.00→0.05	0.414 ± 0.065	1109.7 ± 43.6	0.70

In terms of final success rate, as shown in Figure 3, decaying ε achieves the highest value, 0.889, under the no-noise condition, followed by 0.825 for fixed ε=0.10. As slip_prob rises to 0.10, the difference between the two strategies decreases. At slip_prob=0.20, decaying ε remains slightly higher than fixed ε=0.10, while both are clearly lower than under the no-noise condition. The fixed ε=0.30 strategy has the lowest final success rate under all three noise levels and performs worse as the noise level increases. This indicates that combining a higher level of internal random exploration with external action noise can reduce the consistency of policy execution.

DOI: 10.54254/2755-2721/2026.37323

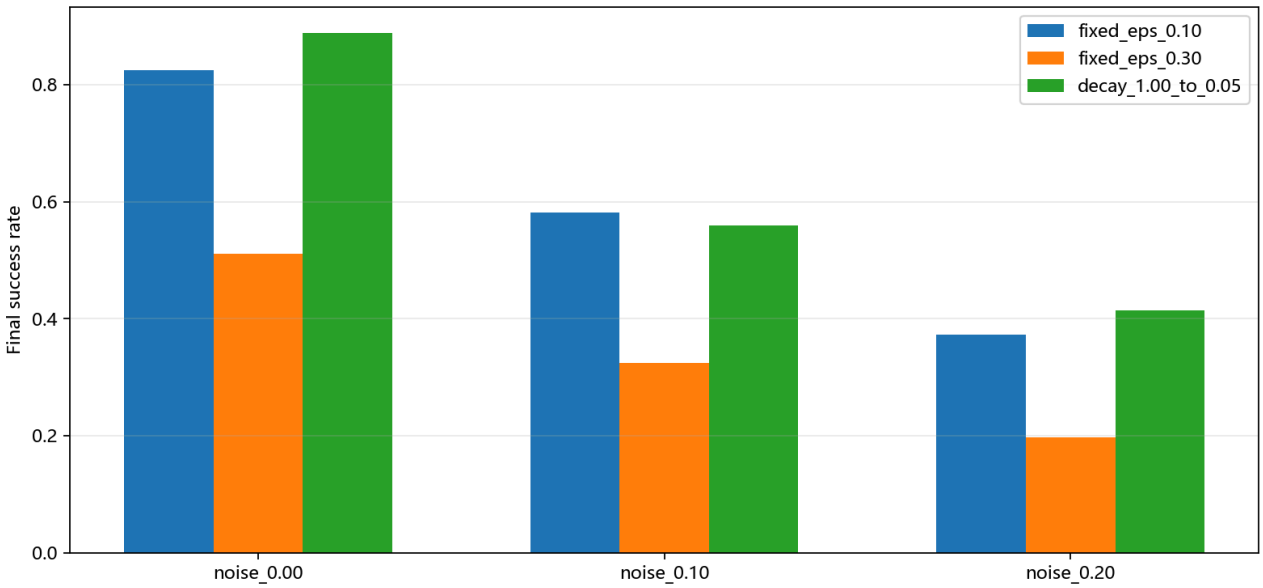

Figure 3. Comparison of final success rates under different environmental noise levels (picture credit: original)

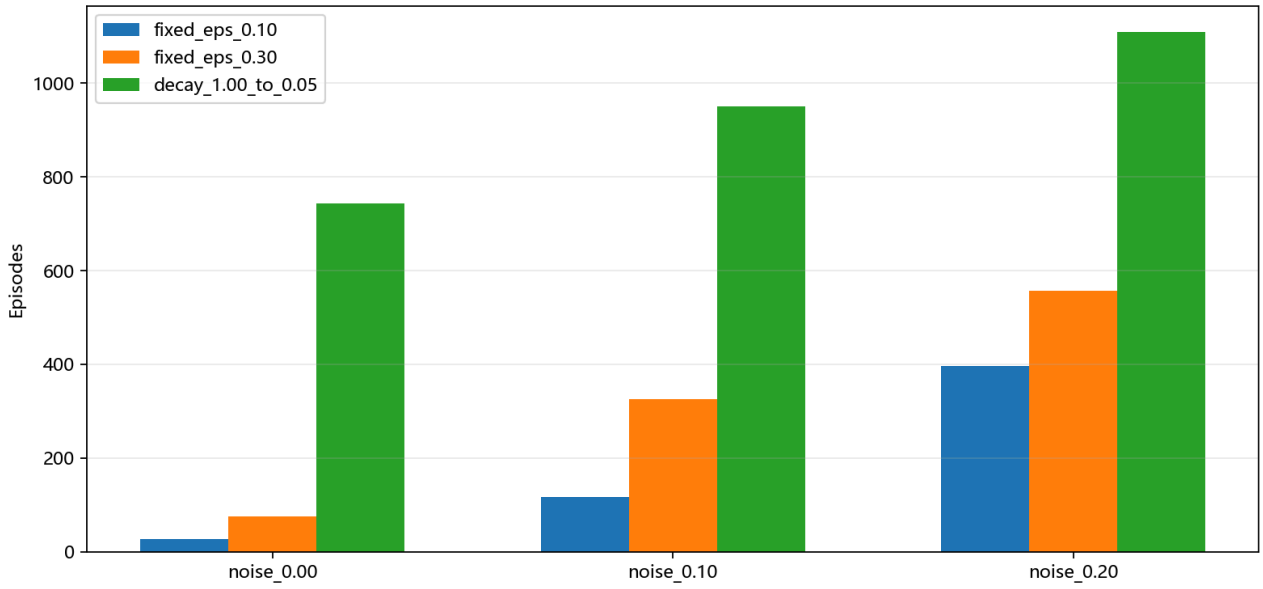

Figure 4. Episodes required to reach a 50% rolling success rate (lower is faster) (picture credit: original)

The differences in learning speed are even more obvious. Figure 4 shows that fixed ε=0.10 requires the fewest episodes to reach a 50% rolling success rate under all three noise levels: approximately 27.6, 116.8, and 397.3 episodes, respectively. Although decaying ε has a higher final success rate under low-noise conditions, it requires approximately 743.2, 951.1, and 1109.7 episodes, respectively, to reach the threshold. This means that the linear decay strategy uses more of

DOI: 10.54254/2755-2721/2026.37323

the training budget for early exploration and can gain some room for better later-stage performance, but at the cost of clearly slower early convergence.

Again, the learning curves are shown for the medium-noise condition slip_prob=0.10, in Figure 5. With a fixed ε=0.10, the success rate is quickly brought to a relatively high level around the first 200 episodes and then stays in a relatively stable range. The curve for fixed ε=0.30 is at a lower level for most of the training process, and decaying ε mainly catches up during the second half of training. This is in line with the learning-speed measures in Table 2 and indicates that similar final success rates do not imply that the learning processes are similar.

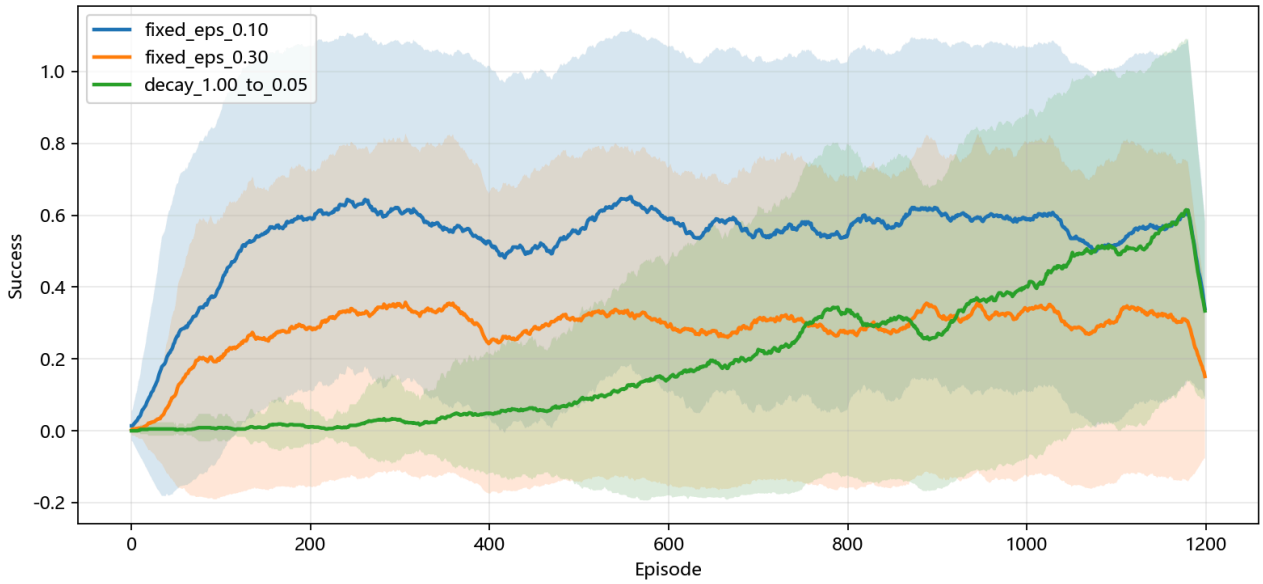

Figure 5. Learning curve under medium noise: success rate (mean±std) (picture credit: original)

3.3. Speed-performance trade-off and reachability under high noise

Overall, the results first show a clear "speed-performance" trade-off. Fixed ε=0.10 reaches a usable level relatively quickly under all three noise conditions and therefore has an advantage in small tasks that require shorter training time. Decaying ε can achieve a higher final success rate under low-noise conditions, but the longer high-exploration stage significantly delays convergence. In other words, more extensive exploration may improve policy quality in the later stage of training, but it does not necessarily improve learning efficiency under a limited training budget.

Second, the proportion of seeds reaching the threshold under high-noise conditions reveals stability differences beyond the average success rate. At slip_prob=0.20, the proportion for fixed ε=0.10 remains 1.00, while fixed ε=0.30 is only 0.10 and decaying ε is 0.70. The mean final success rate of fixed ε=0.30 is already low, and the low threshold-reaching proportion further shows that its results are sensitive to random runs. Therefore, in an environment with external randomness, evaluation of exploration strategies should consider the mean, standard deviation, learning speed, and threshold-reaching proportion together, so as to avoid treating occasional success or unstable results as a general advantage.

Therefore, within the parameter range of this study, a fixed lower exploration rate is more suitable for tasks that require a stable path to be formed quickly. Linearly decaying exploration is more suitable when the training period is longer, slower early learning is acceptable, and higher

DOI: 10.54254/2755-2721/2026.37323

later-stage performance is preferred. With the current GridWorld and noise, there is no overall benefit from a fixed higher exploration rate. This conclusion does not imply that fixed ε=0.10 is the best value for all tasks, but rather that the intensity of exploration should be adjusted to the randomness of the environment, the budget of training, and the goal of evaluation.

4. Research limitations and future work

4.1. Research limitations

There are a number of limitations to this study. First, the experimental environment is a small discrete GridWorld, with both the state space and action space being relatively small. The conclusions are mainly intended to illustrate basic exploration mechanisms and cannot be directly generalized to high-dimensional continuous control tasks or large-scale deep reinforcement learning tasks. Second, this study only tested three settings: a fixed low exploration rate, a fixed high exploration rate, and a linearly decaying exploration rate, and did not include more sophisticated methods such as exponential decay, adaptive ε, or approaches based on uncertainty or intrinsic rewards. Third, the experiments primarily focus on the final success rate, threshold learning speed, and statistics over random seeds, with no additional analysis of cumulative return, path length, state coverage, or the influence of different reward designs.

4.2. Future work

There are three avenues for future research. First, sensitivity experiments could be performed using varying initial ε, minimum ε, decay rates, and decay forms to determine if the conclusions reached here are sensitive to specific parameter choices. Second, the same evaluation framework can be used to evaluate larger-scale GridWorlds, environments with random obstacles, and continuous control tasks, and ε-greedy can be compared with intrinsic-reward-based and adaptive exploration methods to see if the trade-offs between learning speed, final performance, and stability remain. Finally, metrics like state visitation coverage and cumulative rewards can be added along with statistical significance tests to extend the assessment of exploration efficiency beyond task success to information acquisition quality and policy reliability.

5. Conclusion

This study is about the exploration-exploitation trade-off in Q-learning and builds a small controllable 5×5 GridWorld experiment. The results of fixed ε=0.10, fixed ε=0.30, and linearly decaying ε (1.00→0.05) are compared in terms of final success rate, learning speed, and stability across random seeds. The experimental results show that the fixed lower exploration rate generally reaches a usable success rate faster and maintains relatively stable threshold-reaching performance under different noise levels. Decaying exploration achieves a higher final success rate in low-noise environments, but learning is clearly slower in the early stage. The fixed higher exploration rate performs relatively poorly overall and has particular difficulty forming a stable policy under high-noise conditions. Therefore, exploration strategies should not be evaluated only by the average success rate at the end of training. Environmental randomness, the training budget required to reach the target performance, and variation across random runs should also be considered. In practical experimental design, the exploration rate should serve the task objective. If quickly forming a usable policy is more important, a lower fixed exploration rate may be more suitable; if the training budget

DOI: 10.54254/2755-2721/2026.37323

is sufficient and later-stage performance is more important, a gradually decaying exploration strategy can be considered.

References

[1] Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3-4), 279-292.

[2] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press, Cambridge, MA.

[3] Fortunato, M., Azar, M. G., Piot, B., et al. (2018). Noisy networks for exploration. *International Conference on Learning Representations*.

[4] Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2019). Exploration by random network distillation. *International Conference on Learning Representations*.

[5] Badia, A. P., Piot, B., Kapturowski, S., et al. (2020). Agent57: Outperforming the Atari human benchmark. *Proceedings of the 37th International Conference on Machine Learning, 119*, 507-517.

[6] Fang, K., Zhu, Y., Savarese, S., & Fei-Fei, L. (2021). Adaptive procedural task generation for hard-exploration problems. *International Conference on Learning Representations*.

[7] Zhang, T., Rashidinejad, P., Jiao, J., et al. (2021). MADE: Exploration via maximizing deviation from explored regions. *Advances in Neural Information Processing Systems, 34*.

[8] Ladosz, P., Weng, L., Kim, M., & Oh, H. (2022). Exploration in deep reinforcement learning: A survey. *Information Fusion, 85*: 1-22.

[9] Chen, E., Hong, Z. W., Pajarinen, J. K., & Agrawal, P. (2022). Redeeming intrinsic rewards via constrained optimization. *Advances in Neural Information Processing Systems, 35*.

[10] Hao, J., Yang, T., Tang, H., et al. (2024). Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems, 35*(7): 8762-8782.

[11] Wang, Y., Yang, M., Dong, R., et al. (2023). Efficient potential-based exploration in reinforcement learning using inverse dynamic bisimulation metric. *Advances in Neural Information Processing Systems, 36*.

[12] Wang, Y., Zhao, K., Liu, F., U, L. H. (2024). Rethinking exploration in reinforcement learning with effective metric-based exploration bonus. *Advances in Neural Information Processing Systems, 37*.

[13] Zeng, X., Peng, H., Li, A. (2024). Effective exploration based on the structural information principles. *Advances in Neural Information Processing Systems, 37*.