Research on Policy Stability of Reinforcement Learning in Complex Dynamic Decision-Making Environments

Aisheng Zhang*

Guangzhou University of Software, Guangzhou, 510990, China
**Corresponding author: zas219@126.com.*

Abstract: *The policy stability of reinforcement learning is critical to its practical application when addressing sequential decision-making problems in complex dynamic environments. This paper systematically analyzes the impact mechanisms of environmental dynamics, algorithm sensitivity, and state-space complexity on policy stability, and it constructs an enhanced framework that integrates constrained optimization, probabilistic environment modeling, meta-learning, and multi-agent coordination. The research provides systematic theoretical support and methodological pathways for enhancing the robustness and adaptability of reinforcement learning in non-stationary environments.*

Keywords: *reinforcement learning, policy stability, complex dynamic environments, Markov Decision Processes, multi-agent systems, meta-learning*

Introduction

The application of reinforcement learning in complex dynamic environments, such as robotic control and autonomous driving, poses severe challenges to the stability of its policies. High environmental uncertainty, dynamic changes, and complex interactions among agents often lead to performance fluctuations and convergence difficulties during the policy learning process, which directly affects the reliable deployment and practical effectiveness of the systems. Addressing the lack of systematic analysis on the mechanisms of policy instability in existing research, this paper constructs an attribution framework from the three dimensions of environmental dynamics, algorithm sensitivity, and multi-agent non-stationarity. It thoroughly examines the internal mechanisms affecting stability and, based on this analysis, proposes an enhanced methodology that integrates constrained optimization, meta-learning, and environmental modeling. This work aims to provide theoretical support and practical pathways for building highly robust reinforcement learning algorithms.

1. An Overview of Reinforcement Learning and Complex Dynamic Decision-Making Environments

1.1 The Basic Principles and Framework of Reinforcement Learning

Reinforcement learning is a computational method centered on a trial-and-error mechanism, which learns optimal decision sequences through interaction with the environment. Its mathematical foundation is typically modeled as a Markov Decision Process (MDP), a framework that consists of four core elements: state, action, reward, and state transition probability. The agent's objective is not to maximize immediate reward but to seek the optimum expected long-term cumulative discounted return. To achieve this goal, the value function serves as a key tool for evaluating the long-term value of following a specific policy in a given state. A policy defines the agent's behavior in any given state and acts as a bridge connecting perception to decision-making. The learning process is essentially the iterative updating and optimization of the policy or value functions, ultimately leading to intelligent decision-making strategies that can effectively cope with environmental challenges. This learning paradigm gives reinforcement learning unique advantages distinct from other machine learning methods when addressing sequential decision-making problems.

1.2 The Definition and Characteristics of Complex Dynamic Decision-Making Environments

Complex dynamic decision-making environments refer to interactive scenarios characterized by high uncertainty, non-linearity, and temporal evolution. Their complexity is first reflected in the variability of environmental factors; the environmental state changes not only based on the agent's own actions but is also influenced by numerous external, independent factors, making its dynamics difficult to fully predict or model. Second, environmental feedback often exhibits delay, where the consequences of a current action may only manifest after multiple time steps, which poses challenges for attribution analysis. Furthermore, the environment may contain multiple autonomous or semi-autonomous agents, whose competitive and cooperative relationships introduce social complexity, making the overall environmental dynamics even more elusive. Such environments also possess partially observable characteristics, meaning agents cannot directly access complete state information and must infer the true state based on incomplete observation sequences. These characteristics collectively constitute severe challenges that reinforcement learning agents must address, thereby imposing higher demands on the robustness and adaptability of learning policies [1].

1.3 Adaptability Analysis of Reinforcement Learning in Dynamic Environments

The inherent interactive learning mechanism of reinforcement learning endows it with potential adaptive capabilities when facing dynamic environments. Value-based learning methods, by continuously updating the estimation of state-action values, can gradually track slow changes in environmental dynamics. Policy search methods directly optimize policy parameters, and their stochastic policy characteristic helps maintain a balance between exploration and exploitation, thereby discovering and adapting to new patterns in the environment. When the environment undergoes non-stationary changes, meta-learning and online learning mechanisms allow the agent to leverage prior knowledge for rapid adjustment or to continually update its policy based on new data. However, this adaptability is not achieved spontaneously; its effectiveness is heavily constrained by factors such as algorithm design, state representation, and exploration strategy. The process through which an agent adapts to the environment is essentially a process where its internal model or policy seeks consistency with the true dynamics of the environment. Any mismatch during this process may cause fluctuations in policy performance or even lead to its failure. Therefore, a deep understanding of adaptive capability serves as the logical starting point for researching policy stability issues, revealing that stability is not a static adherence but a continuous, robust performance maintained under dynamic equilibrium [2].

2. Factors Influencing Policy Stability of Reinforcement Learning in Complex Dynamic Decision-Making Environments

2.1 The Mechanism of Environmental Dynamics on Policy Stability

2.1.1 Non-Stationary State Transitions and the Model Bias Effect

The non-stationary nature of complex dynamic environments is most directly manifested as a systematic drift in state transition probabilities over time. The environmental model that an agent learns during the initial phase gradually becomes invalid as the dynamics evolve, causing the optimal policy based on the old model to experience significant performance degradation in the new environment. This model bias effect can trigger persistent oscillations in the policy, as the agent must continuously relearn to adapt to the changes, and the learning process itself introduces new estimation errors and instability.

2.1.2 Cognitive Uncertainty under Partial Observability

Within the Partially Observable Markov Decision Process (POMDP) framework, the agent cannot directly access the complete state of the environment and must instead form an estimate of the true state based on a sequence of noisy observations [3]. Increased environmental dynamics directly amplify this cognitive uncertainty. An inherent discrepancy exists between the internal belief state maintained by the agent and the true environmental state, leading to decisions based on this belief state being naturally biased. Dynamic changes make the updating and tracking of the belief state exceptionally difficult, and consequently, policy fluctuations manifest as the external expression of this cognitive uncertainty.

2.1.3 Policy Non-Stationarity in Multi-Agent Environments

In multi-agent environments, the core source of environmental dynamics stems from the concurrent evolution of other agents' policies. From the perspective of an individual agent, the environment it operates in becomes inherently non-stationary due to the presence of other learning agents. Any policy update made by one agent alters the environmental dynamics faced by the remaining agents. This interdependence and continuous gameplay form a dynamically evolving policy ecosystem. Within this ecosystem, converging to a static Nash equilibrium is often exceedingly difficult. Instead, agent policies frequently encounter persistent cyclical fluctuations, policy cycles, or even a complete failure to converge, making global stability challenging to guarantee.

2.2 The Relationship Between Algorithm Design Parameters and Policy Convergence

2.2.1 The Regulation of Learning Rate and Policy Update Magnitude

The learning rate parameter directly determines the step size for each iterative update of the policy or value function. An excessively high learning rate causes parameter updates to exhibit severe oscillations within the neighborhood of the optimal solution, potentially overshooting the optimal region and diverging, thereby causing the policy performance to display high-variance characteristics. Conversely, an excessively low learning rate, while ensuring asymptotic convergence, drastically slows the learning process. This prevents the agent from responding promptly to reasonable changes in the environment, resulting in a policy that lacks effective dynamic stability in practice. Adaptive learning rate mechanisms aim to balance this contradiction, but their own design introduces new hyperparameters and complexity.

2.2.2 The Coupling Between Discount Factor and the Agent's Planning Horizon

The discount factor defines the degree to which the agent discounts future rewards, thereby determining the length of its planning horizon. A discount factor that is too small guides the agent to excessively pursue short-term rewards, often resulting in a myopic and fragile policy that fails to prepare for long-term environmental changes. Conversely, an excessively large discount factor (close to 1) incorporates the uncertainties of the distant future into the current value estimation on a large scale. This makes the estimation of the value function difficult to converge and increases its variance, consequently causing the policy based on this value function to become unstable [4].

2.2.3 The Trade-off Between Exploration Strategy and Behavioral Uncertainty

The exploration-exploitation trade-off is a fundamental dilemma in reinforcement learning, and the choice of exploration strategy directly influences the policy's evolutionary trajectory. An overly aggressive or random exploration strategy, such as an ε -greedy policy with a high parameter, continuously injects behavioral noise into the policy. While this aids in discovering potentially superior regions, it severely disrupts the execution consistency of the current policy, making it difficult to converge into a stable output. Conversely, an overly conservative greedy exploitation strategy easily causes the learning process to become trapped in a local optimum prematurely. Once the environment changes, the inflexible policy, lacking exploration of new regions, fails to adapt effectively, thus exhibiting another form of vulnerability.

2.3 The Constraining Effects of State and Action Space Complexity on Stability

2.3.1 High-Dimensional State Spaces and the Curse of Dimensionality

High-dimensional state spaces cause the size of the state set to grow exponentially with the number of dimensions, preventing the agent from sufficiently traversing the state space within limited experience and interaction time. The direct consequence of this is that the value function exhibits substantial estimation variance across vast regions of states that have not been adequately visited. A policy that performs well in some parts of the state space may produce drastically different behaviors in other, similar but under-learned states, leading the policy to demonstrate severe inconsistency and a lack of generalization capability.

2.3.2 Generalization Error and Error Propagation in Function Approximators

Using parameterized function approximators, such as neural networks, has become necessary for handling high-dimensional or continuous state spaces. However, any function approximation inevitably introduces generalization error. In temporal-difference learning algorithms that employ bootstrapping,

this approximation error is not independent; instead, it is iteratively used as part of the target value during the update process, leading to propagation and accumulation effects. This error propagation can ultimately cause the value function estimates to diverge, a phenomenon known as the deadly triad. A policy built upon a diverging value function inherently cannot remain stable.

2.3.3 Policy Gradient Variance under Complex Action Spaces

The complexity of the action space, whether stemming from a large-scale discrete action set or a high-dimensional continuous action space, significantly increases the difficulty of policy optimization. In discrete spaces, the policy must select from a vast set of actions, making it extremely challenging to evaluate the expected return of each action. In continuous spaces, the policy output is a multi-dimensional vector where minor parameter perturbations can lead to significant differences in the output action, thereby causing drastic variations in environmental feedback. Both scenarios result in a sharp increase in the variance of policy gradient estimates. High-variance gradient estimates cause the policy to undergo an inefficient and unstable random walk within the parameter space, making it exceedingly challenging to converge to a stable and high-performing policy.

3. Methods and Strategies for Enhancing Reinforcement Learning Policy Stability

3.1 Stability Enhancement Techniques for Policy Optimization Algorithms

3.1.1 Mathematical Framework of Constrained Policy Optimization

Constrained policy optimization methods introduce explicit mathematical constraints to limit the magnitude of each policy update, thereby avoiding drastic policy shifts. Trust Region Policy Optimization (TRPO) and its approximate algorithm, Proximal Policy Optimization (PPO), are typical representatives of this approach. By constructing surrogate objective functions and constraining the divergence between the old and new policies, they ensure that each iterative update occurs within a trusted region [5]. This mechanism transforms the policy improvement process into a stable, monotonically improving trajectory, effectively preventing policy collapse caused by a single poor update.

3.1.2 Ensemble Methods and Smooth Policy Generation

Ensemble methods reduce estimation variance and enhance decision-making robustness by constructing multiple value function or policy models and fusing their outputs. In value function learning, ensembles of multiple Q-value estimators can mitigate overestimation bias and provide more robust value assessments. At the policy level, generating smoother and more conservative policy behavior can be achieved by averaging the parameters of multiple policy networks or ensembling their action distributions. This approach essentially incorporates uncertainty explicitly into the decision-making process, enabling the output of more stable and reliable actions when encountering unfamiliar states or noise interference.

3.1.3 Coordinated Regulation of Regularization and Policy Entropy

Introducing regularization terms into the policy optimization objective function is an effective technique for enhancing stability. Among these, policy entropy regularization is widely used; it encourages a degree of stochasticity in the policy to maintain exploration capability and prevent premature convergence to fragile deterministic behaviors. Moderate entropy regularization acts like a "buffer" for the policy optimization process, allowing the policy to optimize long-term return while simultaneously considering the smoothness of the action distribution, thereby enhancing its tolerance to minor environmental changes. Additionally, parameter regularization constrains the weights of the policy network or value function network to control model complexity, mitigate overfitting, and improve generalization performance.

3.2 Methods for Environmental Modeling and Uncertainty Handling

3.2.1 Probabilistic Inference-Based Environmental Dynamics Models

Constructing probabilistic environment models, which involves learning the probability distribution of the state transition function and reward function rather than a deterministic mapping, allows the agent to explicitly quantify predictive uncertainty. During planning, the agent can use this model for prospective reasoning, such as selecting the most robust policy from multiple possible future

trajectories via Monte Carlo Tree Search or Model Predictive Control. When model predictions exhibit high uncertainty, the agent can adopt more cautious actions or proactively initiate exploration to reduce this uncertainty. This decision-making logic, grounded in uncertainty, inherently enhances the policy's adaptability.

3.2.2 Meta-Learning and Rapid Adaptation Mechanisms

The meta-learning framework aims to train agents to acquire a higher-order learning ability, enabling them to leverage prior experience accumulated from a set of related tasks to quickly adapt to new or slowly evolving dynamic environments. Its core idea is to model changes in environmental dynamics as changes in the task distribution. Through algorithms such as Model-Agnostic Meta-Learning, the agent is trained to develop policy parameters with favorable initial properties. These parameters can achieve rapid performance improvement after just a few gradient updates using samples from the new environment. This mechanism transforms the adaptation process from learning from scratch into efficient fine-tuning, significantly shortening the policy's period of instability in new environments [6].

3.2.3 Latent Representation and Abstract State Space Learning

Learning low-dimensional, decision-relevant latent state representations from high-dimensional raw observations through representation learning can effectively reduce the complexity of the state space. The learned abstract state space should ideally satisfy the Markov property and filter out task-irrelevant noise. Conducting policy learning within this compact latent space can alleviate the curse of dimensionality, improve the generalization capability of the value function, and reduce oversensitivity to irrelevant environmental details. Techniques such as adversarial training can further encourage the learned representations to be invariant to specific nuisance sources, thereby enhancing the policy's robustness under distributional shifts.

3.3 Policy Coordination and Stabilization Mechanisms in Multi-Agent Systems

3.3.1 Centralized Training with Decentralized Execution Architecture

The Centralized Training with Decentralized Execution (CTDE) paradigm provides an effective framework for addressing multi-agent credit assignment and environmental non-stationarity. During the training phase, algorithms utilize global system information, such as the actions and states of all agents, to learn a powerful centralized critic or value function, which guides the policy updates of individual agents. During the execution phase, each agent makes decisions based solely on its own local observations. This architecture enables agents to account for the influence of other agents' behaviors during training, learning coordinated policies, while maintaining independence and scalability during execution, thereby preventing policies from getting stuck in cycles within complex interactions.

3.3.2 Consensus-Seeking and Equilibrium Selection Algorithms

In multi-agent environments, guiding all agents' policies towards a common, efficient equilibrium point, such as a Nash equilibrium, is crucial for ensuring system stability. Some algorithms promote consensus formation by modifying the learning objective, for instance, incorporating assumptions about opponent policies during policy updates or directly learning equilibrium strategies. Methods such as role assignment and curriculum learning structure the interaction relationships among agents, reducing learning complexity and gradually steering the system towards a desired stable state. These methods aim to provide an implicit coordination signal for the decentralized decision-making process, preventing policies from endlessly switching between multiple potential equilibria.

3.3.3 Multi-Agent Experience Replay and Communication Protocols

Specially designed multi-agent experience replay buffers help stabilize value function learning by storing and mixing joint experience tuples from all agents, which breaks the temporal correlations in the training data. Furthermore, designing constrained and structured inter-agent communication protocols allows agents to exchange limited, task-relevant information, effectively reducing misunderstandings and misjudgments between policies. By communicating intentions or future plans, agents can better predict each other's actions, leading to more foresighted and coordinated decisions. This lays the foundation for establishing and maintaining a stable policy landscape in cooperative or mixed-motive scenarios.

Conclusion

This paper systematically investigates the issue of policy stability in reinforcement learning within complex dynamic environments. The research reveals the internal mechanisms of policy instability from three dimensions: environmental dynamics, algorithm parameter sensitivity, and state-space complexity, identifying key challenges such as non-stationary environments, partial observability, and high-dimensional space representation. To address the aforementioned problems, this paper proposes a comprehensive system of stability enhancement techniques, including constrained policy optimization, probabilistic environment modeling, meta-learning mechanisms, and multi-agent coordination methods, which effectively improve the robustness and adaptability of policies.

Future research still needs to make breakthroughs in several important directions: establishing more comprehensive formal metrics for policy stability, developing adaptive algorithms for non-stationary environments that possess both theoretical guarantees and high efficiency, and resolving the challenge of scalable coordination in large-scale multi-agent systems. Breakthroughs in these directions will promote the reliable application of reinforcement learning in complex real-world scenarios.

References

[1] Sang Jimao, et al. "A Meta-Reinforcement Learning Approach for Multi-Interceptor Cooperative Decision-Making in Adversarial Environments." Acta Armamentarii, 1-16.
[2] Li Mingye, et al. "Development of a Dynamic Nuclear Emergency Evacuation Optimization Decision-Model Based on Deep Reinforcement Learning." Radiation Protection, vol. 45, no. 05, 2025, pp. 517-529.
[3] Luo Mingyue. Research on Socially Compliant Motion Decision-Making and Control for Robots Based on Reinforcement Learning in Dynamic Pedestrian Environments. 2025. Changchun University of Technology, PhD dissertation.
[4] Yue Qiqiang, et al. "A Dynamic Routing and Scheduling Algorithm for Computing and Network Collaboration Based on Deep Reinforcement Learning." Telecommunications Science, vol. 41, no. 08, 2025, pp. 33-50.
[5] Yang Zhipeng. Evolution of Cooperation in Dynamic Decision-Making and Group Social Dilemma Games under Reinforcement Learning. 2024. Wuhan University of Science and Technology, PhD dissertation.
[6] Shen Le. Research on Dynamic Offloading and Resource Allocation in Multi-UAV Assisted Mobile Edge Computing Networks Based on Deep Reinforcement Learning. 2023. Nanjing University of Posts and Telecommunications, MA thesis.