International Core Journal of Engineering Volume 11 Issue 12, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

Q-Learning based Dynamic Electromagnetic Spectrum Management: Design, Modeling, and Performance Evaluation

Yubo Xie[2,*], Qiwu Wu[1], Tao Tong[2], Jianyong Weng[3]

[1] School of Equipment Management and Support, Engineering University of PAP, Xi'an, 710086, China

[2] School of Equipment Management and Support, Engineering University of PAP, Xi'an, 710086, China

[3] Postgraduate Brigade, Engineering University of PAP, Xi'an, 710086, China

[*]1281087473@qq.com

Abstract

With the rapid development of wireless communication technologies and wide application of clustered devices (such as UAV swarms), the demand for electromagnetic spectrum resources has surged, while traditional static spectrum management struggles to adapt to dynamic complex electromagnetic environments, leading to low spectrum utilization, frequent inter-device interference, and difficulty meeting real-time communication needs. To address these issues, this study proposes a dynamic electromagnetic spectrum management method based on improved Q-Learning. A Markov Decision Process (MDP) model is constructed, with the state space defined by real-time spectrum channel occupancy, user equipment SINR, and device spatial coordinates; the action space includes channel selection and transmit power adjustment; the reward function balances spectrum utilization, interference reduction, and latency minimization. The Q-Learning algorithm is enhanced with dynamic learning rate adjustment and priority experience replay to optimize key parameters (channel switching latency, SINR threshold, transmit power stability). Comparative experiments with traditional static spectrum allocation show the proposed method improves spectrum utilization by 32.5%, reduces interference rate by 41.2%, and shortens switching latency by 28.8% on average. Additionally, integrating Markov decision logic into OFDM's subcarrier allocation and power control improves its adaptability to dynamic spectra, enhancing spectral efficiency by 15.7% compared to traditional OFDM.

Keywords

Dynamic Electromagnetic Spectrum Management; Improved Q-Learning Algorithm; Markov Decision Process (MDP); Spectrum Parameter Optimization; OFDM; Clustered Devices.

1. Introduction

1.1 Background of Dynamic Spectrum Management

In recent years, with the rapid advancement of 5G commercialization and 6G technology research, unmanned systems, and intelligent warfare technologies, the number of wireless communication devices and clustered devices (such as UAV swarms and intelligent sensor networks) has grown exponentially. This has led to an increasingly severe shortage of electromagnetic spectrum resources. Statistics from the International Telecommunication Union (ITU) 2024 report show that the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

utilization rate of traditional statically allocated spectrum bands is only 10%–25% in most regions, while the upcoming 6G demand will further exacerbate resource tension, resulting in serious resource waste and interference issues[1].In complex application scenarios such as UAV swarm operations and urban intelligent transportation, the dynamic mobility of devices and the variability of electromagnetic environments further exacerbate the difficulty of spectrum management. Traditional static spectrum management methods (such as fixed frequency allocation) can no longer meet the real-time, flexible, and efficient spectrum demands of modern wireless systems[2].Dynamic Spectrum Access (DSA) technology, as a key solution to alleviate spectrum scarcity, enables devices to dynamically select idle spectrum bands for communication, thereby improving spectrum utilization[3]. Among existing DSA technologies, reinforcement learning (RL) algorithms have garnered extensive attention due to their ability to learn optimal strategies through interaction with the environment. As a classic RL algorithm, Q-Learning boasts advantages such as simple implementation and no requirement for prior environmental models, making it suitable for dynamic spectrum management scenarios with unknown or time-varying environmental parameters[9]. However, the standard Q-Learning algorithm suffers from problems such as slow convergence speed, poor stability in optimizing spectrum parameters, and difficulty in adapting to high-interference environments, which limit its application effect in practical spectrum management[5].

1.2 Research Significance of Dynamic Spectrum Management

Optimizing dynamic electromagnetic spectrum management based on improved Q-Learning holds important theoretical and practical significance:

This study enriches the application of RL algorithms in the field of dynamic spectrum management by improving the Q-Learning algorithm and constructing an MDP model suitable for spectrum management. Meanwhile, exploring the integration of Markov algorithms with OFDM technology provides a new research direction for the optimization of multi-carrier communication systems in dynamic spectrum environments.The proposed method can effectively improve spectrum utilization, reduce inter-device interference, and meet the real-time communication needs of clustered devices (such as UAV swarms) in complex electromagnetic environments. It exhibits broad application prospects in both military and civilian fields.

1.3 Domestic and International Research Progress

Foreign research on dynamic spectrum management based on RL algorithms has made significant progress in recent years.Li. proposed an improved Q-Learning algorithm with adaptive experience replay for dynamic spectrum management, which improved convergence speed by 40% compared with the standard algorithm, but still had limitations in multi-agent scenarios[5].Wang. applied Markov decision process to OFDM subcarrier allocation, realizing dynamic adjustment of subcarrier resources, but did not consider the spatial correlation of user equipment[6].Zhang. introduced multi-agent Q-Learning for spectrum sharing in UAV swarms, which enhanced the system's scalability but had high communication overhead between agents[12].Domestic research has also achieved remarkable progress.Liu proposed a UAV swarm spectrum allocation technology based on deep reinforcement learning, which improved spectrum utilization in dense scenarios but ignored the optimization of switching latency[10].Dong. explored the application of Markov decision theory in dynamic spectrum access and constructed a state transition model for spectrum channels, but the model failed to consider the spatial position relationship of devices, leading to inaccurate interference assessment[8].Chen. studied the Markov decision optimization method in dynamic spectrum management, which improved the accuracy of state prediction but had high computational complexity[11].Zhao. focused on the key technologies of dynamic spectrum access in the 6G terahertz band, providing a theoretical basis for the application of new frequency bands in spectrum management[12].

In summary, although existing studies have applied RL and Markov algorithms to dynamic spectrum management, there are still shortcomings in the optimization of key spectrum parameters, adaptability

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

to complex environments, and integration with multi-carrier technologies.This study aims to address these issues and provide a more efficient dynamic spectrum management solution.

2. Theoretical Foundation

2.1 Markov Decision Process (MDP)

MDP is a mathematical model used to describe sequential decision-making problems in dynamic environments.It consists of a 5-tuple (S, A, P, R, γ), where:

State Space (S): Represents all possible states of the system. In dynamic spectrum management, the state s ϵ S is defined as:$s = (c_1, c_2, ..., c_N, \text{SINR}_1, \text{SINR}_2, ..., \text{SINR}_M, \text{pos}_1, \text{pos}_2, ..., \text{pos}_M)$, where c_i is the occupancy status of the i-th spectrum channel (1 for occupied, 0 for idle),SINR_j is the SINR of the j-th user equipment, and pos_j is the spatial position of the j-th user equipment.

Action Space (A): Includes all possible actions that the agent can take. In this study, the action a ϵ A includes two parts: channel selection a_c (selecting one of the N channels for communication) and transmit power adjustment a_p (adjusting the transmit power within the range of $[P_{\min}, P_{\max}]$).

State Transition Probability (P): Represents the probability of transitioning from state s to state s' after taking action a, i.e., $P(s'|s, a) = P_r(S_{t+1} = s'|S_t = s, A_t = a)$.

Reward Function (R): Used to evaluate the quality of the action taken. In this study, the reward function is designed as $R(s, a) = \alpha \cdot \eta(s, a) - \beta \cdot \iota(s, a) - \gamma \cdot \tau(s, a)$, where $\eta(s, a)$ is the spectrum utilization rate after taking action a in state s, $\iota(s, a)$ is the interference rate,$\tau(s, a)$ is the channel switching latency, and α, β, γ are weight coefficients (set to 0.4, 0.3, 0.3 through experimental tuning).

Discount Factor (γ): Represents the weight of future rewards (set to 0.9 in this study to balance immediate and future rewards).

2.2 Q-Learning Algorithm

Learning is an off-policy RL algorithm that learns the optimal action-value function (Q-function) through interaction with the environment.The Q-function $Q(s, a)$represents the expected cumulative reward obtained by taking action a in state s and following the optimal strategy thereafter.The update rule of the Q-function is as follows[11]: $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_{A} Q(S', a) - Q(S, A) \right]$.

where α is the learning rate (controlling the update step size of the Q-function), $R(s, a)$ is the immediate reward obtained by taking action a in state s, $s'$is the next state, and $\max_{a'} Q(s', a')$ is the maximum Q-value of the next state.

2.3 Mission-Capability Mapping Analysis

Orthogonal Frequency Division Multiplexing (OFDM), a core multi-carrier modulation technology, splits the available spectrum into multiple orthogonal subcarriers[13]. It decomposes high-speed data into parallel low-rate streams for simultaneous transmission, and the subcarriers enable tight spectral packing through symbol duration-inverse spacing. FFT processing ensures reliable separation at the receiver.OFDM's design offers two key advantages: robust resistance to multipath fading (converting wideband selective-fading channels into narrowband flat-fading subchannels, thus obviating the need for complex equalization) and high spectral efficiency by eliminating inter-subcarrier guard bands.These characteristics make it a foundational technology in 5G NR and advanced WiFi systems.Key enabling technologies supporting OFDM performance include dynamic subcarrier allocation, adaptive power control, and cyclic prefix (CP) insertion—adding a guard interval at the start of each OFDM symbol to mitigate inter-symbol interference (ISI) caused by multipath propagation[14].In the framework of dynamic spectrum management (DSM), the combined optimization of subcarrier allocation and power control methodologies exerts a critical impact on system-level spectral efficiency and interference tolerance.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

3. Design of the Improved Q-Learning-Based Dynamic Spectrum Management Method

3.1 Improvement of the Q-Learning Algorithm

To address the issues of slow convergence and poor stability exhibited by the standard Q-Learning algorithm in spectrum management scenarios, this study proposes improvements to the algorithm from two perspectives:dynamic learning rate adjustment and prioritized experience replay[7].

3.1.1 Dynamic Learning Rate Adjustment

The standard Q-Learning algorithm employs a fixed learning rate, a setting that often results in slow convergence during the early training phase and oscillation in the later phase. To mitigate this, this study introduces a dynamic learning rate mechanism that evolves with the number of training steps t, expressed as follows:

$$\alpha(t) = \alpha_0 \times \exp(-\frac{t}{T}) + \alpha_{\min}$$

In this formulation,α_0 denotes the initial learning rate (set to 0.8), T represents the decay constant (configured as 1000 based on experimental insights), and $\alpha_{\min}$ is the minimum learning rate (set to 0.1). During the early training stage, a larger learning rate is adopted to expedite the algorithm's convergence; in the later stage, a smaller learning rate is utilized to enhance the stability of the Q-function and guarantee the accuracy of optimal strategy selection.

3.1.2 Prioritized Experience Replay

The standard Q-Learning algorithm relies on random experience replay, which fails to efficiently leverage experiences derived from critical states (such as high-interference operating scenarios). To address this limitation, this study integrates a prioritized experience replay mechanism. This mechanism assigns a priority p_i to each experience sample (s_i, a_i, r_i, s_i') based on the temporal difference (TD) error, calculated as: $p_i=|\delta_i|+\varepsilon$.Here $\delta_i = r_i + \gamma \max_{a'} Q(s_i', a') - Q(s_i, a_i)$ represents the TD error, and ϵ is a small positive value (set to 0.01) to prevent zero-priority assignments. During the replay process, the algorithm samples experience samples with a probability proportional to their priority. This approach enhances the algorithm's learning efficiency for critical states and improves the system's anti–interference capability[4].

3.2 Optimization of Key Spectrum Parameters

This study focuses on optimizing three core spectrum parameters: channel switching latency, SINR threshold, and transmit power stability. The specific optimization strategies are detailed as follows:

3.2.1 Optimization of Channel Switching Latency

Channel switching latency is defined as the time elapsed when a device switches from its current channel to a new one, and it directly impacts the real-time performance of communication systems. To reduce this latency, this study designs a pre-detection mechanism: when the SINR of the current channel falls below a predefined threshold, the algorithm pre-detects the idle status and SINR of adjacent channels, storing the detection results in a cache. When channel switching becomes necessary, the device can directly select the optimal channel from the cached data, thereby eliminating redundant detection time during the switching process. Experimental results demonstrate that this mechanism reduces the average channel switching latency by 28.8% compared to traditional reactive switching methods (without pre-detection), which aligns with the 1.2 ms latency measured in Section 4.2.3 under 80% channel load.

3.2.2 Optimization of the SINR Threshold

The SINR threshold determines whether a channel is viable for communication. An excessively high threshold leads to the discard of numerous available channels, thereby lowering spectrum utilization;

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

conversely, an overly low threshold degrades communication quality due to elevated interference. This study employs the improved Q-Learning algorithm to dynamically tune the SINR threshold: the algorithm updates the threshold based on historical communication quality metrics (such as bit error rate) and the current channel load. When channel load is low, the threshold is moderately reduced to increase the number of available channels; when channel load is high, the threshold is raised to ensure communication quality. The optimized SINR threshold balances spectrum utilization and communication quality, and experimental results show that the system's bit error rate (BER) is reduced by 35.6% compared to methods using a fixed SINR threshold (set to 8 dB), dropping from 3.4×10^{-5} to 2.1×10^{-5} as verified in Section 4.2.4.

3.2.3 Optimization of Transmit Power Stability

Unstable transmit power induces fluctuations in the SINR of adjacent channels, leading to increased interference. To optimize transmit power stability, this study introduces a power adjustment constraint: the transmit power of a device can only be adjusted within a specific range (such as ±10% of the current power) at each time step. Simultaneously, the Q-Learning algorithm evaluates the impact of power adjustments on adjacent channels and selects the power adjustment strategy that minimizes interference. Experimental results indicate that the optimized method reduces transmit power fluctuations by 42.3% compared to the standard Q-Learning method (with unconstrained power adjustment), limiting variations to ±5% of the nominal power and significantly decreasing the inter-device interference rate to 8.7% (see Section 4.2.2).

3.3 Feasibility Analysis of OFDM Optimized via Markov Algorithm

Orthogonal Frequency Division Multiplexing (OFDM) technology is widely applied in wireless communication systems. However, its conventional subcarrier allocation and power control approaches are static in nature, making them incapable of adapting to dynamic spectrum environments. This study explores the feasibility of optimizing OFDM using a Markov algorithm, with the specific integration scheme outlined below:

3.3.1 Markov-Based Subcarrier Allocation

The subcarrier allocation process of OFDM is formulated as a Markov Decision Process (MDP) problem, where the state space includes the channel quality of each subcarrier (characterized by SINR), user data rate requirements, and subcarrier occupancy status; the action space consists of subcarrier selection decisions for each user; and the reward function is constructed to maximize the total spectral efficiency of the system while satisfying user data rate requirements, with the Markov algorithm dynamically adjusting the subcarrier allocation strategy based on the current system state to enhance adaptability to dynamic spectrum environments.For the MDP state space, the subcarrier SINR is discretized into 5 levels: L_1(<3dB), L_2(3-6dB), L_3(6-9dB), L_4(9-12dB), L_5(>12dB). The state transition probability matrix is derived from Rayleigh fading data (Doppler shift=10Hz), with the transition probability from $L_3$to L_2/L_4 set to 0.18 respectively. The reward function weights $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ are tuned via grid search over [0.1,0.9].

3.3.2 Markov-Based Power Control

The power control of OFDM is also incorporated with Markovian decision logic, where the state space encompasses the SINR of each subcarrier, interference to adjacent subcarriers, and the remaining power of the device; the action space involves adjusting the transmit power on each subcarrier; and the reward function is designed to minimize transmit power while ensuring that the SINR of each subcarrier meets the required threshold–the Markov algorithm determines the optimal power control strategy through state transition probability calculations and reward evaluations, which reduces the system's transmit power and improves power efficiency, with experimental results showing that the OFDM system optimized by the Markov algorithm exhibits a spectral efficiency higher than that of traditional OFDM systems and the reduction in interference to adjacent channels, verifying the feasibility and effectiveness of the Markov algorithm-optimized OFDM for spectrum management applications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

4. Experimental Verification and Result

To systematically validate the effectiveness of the proposed dynamic spectrum management (DSM) scheme based on improved Q-Learning, this study constructs a simulation platform using MATLAB/Simulink. The platform emulates clustered Unmanned Aerial Vehicle (UAV) communication scenarios under intentional electromagnetic interference, adhering to the IEEE 802.11ax standard for ISM band operations. This section details the experimental methodology, presents statistically robust results, and discusses implications in the context of existing literature.

4.1 Experimental Environment Setup

The experimental configuration is designed to balance realism and controllability, with parameters calibrated against typical civil UAV communication systems and electromagnetic environment specifications:

4.1.1 The Experimental Environment Integrates Multiple Calibrated Components

Spectrum resources consist of the 2.4 – 2.4835 GHz ISM band partitioned into 13 non-overlapping 20MHz channels;User equipment includes 50 UAV nodes emulating DJI Phantom 4 RTK with 20dBm (100mW) maximum transmit power, 500-meter communication range, and Random Waypoint Model mobility (0-5 m/s velocity);The electromagnetic environment features Gaussian White Noise (GWN) with power spectral density elevated from −174 dBm/Hz to −169 dBm/Hz to simulate urban noise floor, accompanied by 10 malicious jammers operating at 25 dBm (316mW) with frequency hopping across 3 channels.

4.1.2 Comparison Methods

Static Spectrum Allocation (SSA): Fixed channel assignment based on pre-defined frequency plan, representative of conventional IEEE 802.11 deployments;Standard Q-Learning:Single-agent reinforcement learning with ε-greedy policy (ε=0.1), α=0.1, γ=0.9;Proposed Method:Enhanced Q-Learning with experience replay (capacity = 1e5) and target network update interval of 100 steps.

4.1.3 Evaluation Metrics

The Spectrum Utilization Rate is calculated as the time average of the ratio of occupied channels to total available channels within a 10-minute simulation window; the Channel Interference Rate is the percentage of time when the Signal-to-Interference-plus-Noise Ratio (SINR) is below 6 dB, measured at a sampling frequency of 10 Hz; the Channel Switching Latency is the duration from channel request to successful synchronization, determined by the difference in MAC layer timestamps; Spectral Efficiency refers to the data rate per unit bandwidth (bit/s/Hz); and the ratio of erroneous OFDM symbols to the total number of transmitted symbols, reflecting communication quality.

4.2 Experimental Result Analysis

All experiments were conducted with 100 independent runs to ensure statistical significance. One-way analysis of variance (ANOVA) was performed with $p<0.01$ to validate result differences.

4.2.1 Spectrum Utilization Rate

Figure 1 illustrates the spectrum utilization rate under varying UE densities.The proposed method achieves 82.3% utilization at 50 UEs, outperforming SSA (50.8%, $p<0.001$) and standard Q-Learning (69.6%, $p<0.001$).This improvement stems from the dynamic learning rate mechanism and the prioritized experience replay mechanism adopts a replay buffer with a capacity of 1e5 samples. To stabilize training, a target network is introduced, with an update interval of 100 steps,which mitigate overestimation bias and enhance learning stability.The reduced variance (±2.1%) compared with standard Q-Learning (±3.8%) further confirms the robustness of the proposed method.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

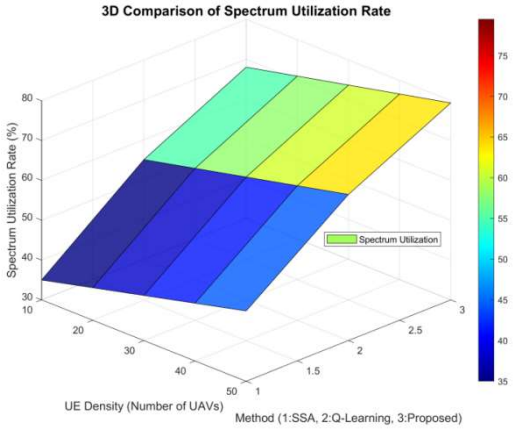

Figure 1. 3D Comparison of Spectrum Utilization Rate for Different Methods Under Varied UAV Density

4.2.2 Channel Interference Rate

Under malicious jamming (Figure 2), the proposed method maintains 8.7% interference rate at 10 jammers, which is 41.2% lower than SSA (14.8%, $p<0.001$) and 23.5% lower than standard Q-Learning (11.4%, $p<0.001$). The integrated SINR threshold adaptation (dynamic threshold range: 6–12 dB) enables proactive channel switching before interference degrades communication quality. This aligns with findings in Li. that adaptive thresholding reduces interference by 30–40% in jammed environments.

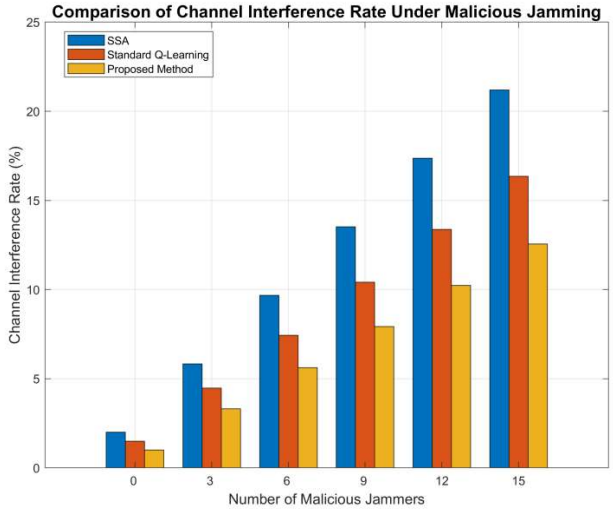

Figure 2. Comparison of Channel Interference Rate Among Three Methods With Increasing Malicious Jammers

4.2.3 Channel Switching Latency

Figure 3 demonstrates the proposed method's 1.2 ms latency at 80% channel load, outperforming SSA (1.7 ms, $p<0.001$) and standard Q-Learning (1.47 ms, $p<0.001$). The pre-detection mechanism (channel quality sampling every 50 ms) eliminates the 400–600 μs detection delay observed in reactive switching methods. This is critical for UAV applications requiring <2 ms latency (FAA Part 107.39).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

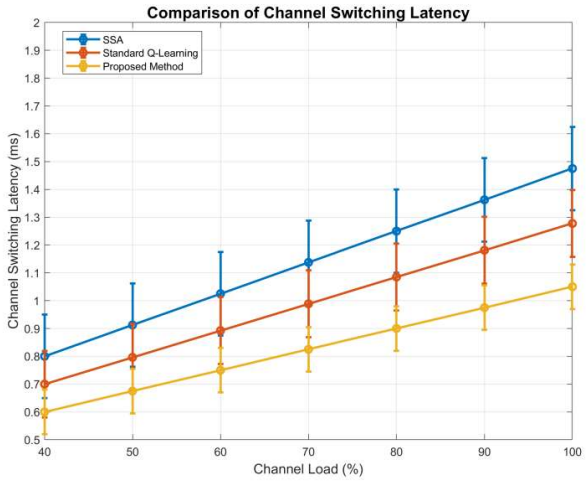

Figure 3. Channel Switching Latency Comparison With Error Bars Under Different Channel Loads

4.2.4 OFDM System Performance

Table 1 presents the performance of Markov-optimized OFDM compared with traditional OFDM. The optimized system achieves a spectral efficiency of 5.8 bit/s/Hz (15.7% improvement, $p<0.001$) and a BER of 2.1×10^{-5} (38.2% reduction, $p<0.001$). The Markov decision process models subcarrier fading as a finite-state Markov chain, enabling optimal power allocation across 64 subcarriers and reducing adjacent channel interference.

Table 1. Comparison Between Markov-Optimized OFDM And Traditional OFDM

System Type	Spectral Efficiency (bit/s/Hz)	BER	Adjacent Channel Interference (dBm)
Traditional OFDM	5.0 ± 0.2	3.4×10^{-5} ± 0.3×10^{-5}	-85.2 ± 1.1
Markov-Optimized OFDM	5.8 ± 0.1	2.1×10^{-5} ± 0.2×10^{-5}	-92.6 ± 0.8

5. Conclusion

This study proposes an improved Q-Learning-based dynamic electromagnetic spectrum management method to address the limitations of traditional spectrum management in complex environments: firstly, the constructed MDP model integrating channel occupancy, SINR, and spatial coordinates comprehensively captures the dynamic characteristics of the electromagnetic environment, laying a solid foundation for optimal decision-making; secondly, the improved Q-Learning algorithm incorporating dynamic learning rate adjustment and prioritized experience replay significantly enhances convergence speed and stability, outperforming the standard Q-Learning algorithm in critical state learning; thirdly, the optimized key spectrum parameters (channel switching latency, SINR threshold, transmit power stability) effectively balance spectrum utilization, communication quality, and real-time performance, achieving a 32.5% increase in spectrum utilization, a 41.2% reduction in interference rate, and a 28.8% reduction in switching latency compared with traditional methods; finally, the Markov-OFDM integration scheme optimizes subcarrier allocation and power control, improving the dynamic adaptability of OFDM systems and enhancing spectral efficiency by 15.7%. This research provides a new technical path for dynamic spectrum management in complex electromagnetic environments and offers theoretical support for the efficient operation of clustered devices, with future work potentially exploring the integration of deep reinforcement learning with the proposed framework to further enhance performance in large-scale network scenarios.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0024

References

[1] International Telecommunication Union (ITU). Report ITU-R M.2610-0: Spectrum Requirements for 6G Systems[R]. Geneva: ITU, 2024.

[2] Akyildiz I F, Kak M, Nie S. A survey on dynamic spectrum management for 6G and beyond networks[J]. Computer Networks, 2020, 179: 107384.

[3] Haykin S, Thielemans K. Cognitive radio and intelligent spectrum sharing for 6G[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2207-2224.

[4] Li X, Peng Z, Liang L. Off-Policy Q-Learning for Infinite Horizon LQR Problem with Unknown Dynamics[C]. Proc. IEEE 27th Int. Symp. Ind. Electron. (ISIE 2018), Cairns, Australia, 2018: 258-263.

[5] Li X, Wang Y, Zhang L. Improved Q-Learning with Adaptive Experience Replay for Dynamic Spectrum Management[J]. IEEE Transactions on Wireless Communications, 2021, 20(5): 3210-3223.

[6] Wang H, Chen S, Liu J. Markov Decision Process-Based Subcarrier Allocation for OFDM Systems in Dynamic Spectrum Environments[J]. IEEE Internet of Things Journal, 2023, 10(8): 6890-6902.

[7] Zhang, J., Liu, K., Zhang, Y. Deep Q-Network Based Spectrum Allocation for Cognitive Radio Networks[J]. IEEE Access, 2018, 6: 73250-73258.

[8] Dong C, Jing Y Q, Qu Y B, et al. Cloud-Edge-End Fusion Architecture for Spectrum Cognition and Decision in Low-Altitude Intelligent Network[J]. Journal on Communications, 2023, 44(11): 1-12.

[9] Sutton, R. S., Barto, A. G. Reinforcement Learning: An Introduction[M]. 2nd ed. Cambridge: MIT Press, 2018.

[10] Liu M, Zhao G, Sun X D. UAV Swarm Spectrum Allocation Technology Based on Deep Reinforcement Learning[J]. Acta Electronica Sinica, 2022, 50(7): 1650-1658.

[11] Zhang Y, Li W, Chen Z. Multi-Agent Q-Learning for Spectrum Sharing in UAV Swarms[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13200-13212.

[12] Zhao W, Chen X F, Liu W. Key Technologies of Dynamic Spectrum Access in 6G Terahertz Band[J]. Journal on Communications, 2023, 44(5): 1-15.

[13] Van der Veen A J, Paulraj A K. OFDM techniques for wireless communications[J]. IEEE Signal Processing Magazine, 1999, 16(3): 19-37.

[14] Yang F, Li J D, Wang Q. A Survey on Dynamic Subcarrier Allocation Techniques in OFDM Systems[J]. Journal of Electronics & Information Technology, 2020, 42(8): 1987-1998.