AI-Based Q-Learning Approach for Performance Optimization in MIMO-NOMA Wireless Communication Systems

Original Scientific Paper

Ammar A. Majeed
Middle Technical University, Kut Technical Institute
Baghdad, Iraq
Ammar.alaa@mtu.edu.iq

Douaa Ali Saed
University of Wasit, Electrical engineering department
Wasit, Iraq
dsaed@uowasit.edu.iq

Ismail Hburi
University of Wasit, Electrical engineering department
Wasit, Iraq
isharhan@uowasit.edu.iq

Abstract *– In this paper, we investigate the performance enhancement of Multiple Input, Multiple Output, and Non-Orthogonal Multiple Access (MIMO-NOMA) wireless communication systems using an Artificial Intelligence (AI) based Q-Learning reinforcement learning approach. The primary challenge addressed is the optimization of power allocation in a MIMO-NOMA system, a complex task given the non-convex nature of the problem. Our proposed Q-Learning approach adaptively adjusts power allocation strategy for proximal and distant users, optimizing the trade-off between various conflicting metrics and significantly improving the system's performance. Compared to traditional power allocation strategies, our approach showed superior performance across three principal parameters: spectral efficiency, achievable sum rate, and energy efficiency. Specifically, our methodology achieved approximately a 140% increase in the achievable sum rate and about 93% improvement in energy efficiency at a transmitted power of 20 dB while also enhancing spectral efficiency by approximately 88.6% at 30 dB transmitted Power. These results underscore the potential of reinforcement learning techniques, particularly Q-Learning, as practical solutions for complex optimization problems in wireless communication systems. Future research may investigate the inclusion of enhanced channel simulations and network limitations into the machine learning framework to assess the feasibility and resilience of such intelligent approaches.*

Keywords*: MIMO-NOMA Networks, Power Allocation Strategies, Optimization of Wireless Communication Systems, Reinforcement Learning Techniques, Q-Learning Approach*

1. INTRODUCTION

The exponential proliferation of wireless devices, accompanied by a commensurate increase in data generation, has imposed unparalleled demands on contemporary wireless networks. This evolving landscape mandates a fundamental paradigmatic shift in the design and optimization strategies for forthcoming wireless communication systems, emphasizing maximizing Spectral efficiency (SE) and Energy Efficiency (EE) [1, 2]. NOMA has emerged as a seminal technology to address these burgeoning challenges. Unlike conventional Orthogonal Multiple Access (OMA) schemes, NOMA enables the concurrent utilization of identical time-frequency resources by multiple users, thereby substantially augmenting both SE and system connectivity [3, 4].

MIMO-NOMA systems, which integrate MIMO technology into NOMA, further enhance system capacity and efficiencies, providing an advanced solution for next-gen wireless networks [5]. Utilizing Spatial Data Multiplexing and Signal Diversity techniques, MIMO-NOMA systems are capable of realizing considerable improvements in data rate and EE. Nevertheless, the augmented complexity of MIMO-NOMA, particularly concerning power allocation and user grouping strategies, introduces formidable challenges that must be meticulously addressed. [6, 7]. NOMA offers key advantages like enhanced connectiv-

ity, reduced latency, and improved QoS, contributing to a more efficient wireless network [8, 9]. A salient constraint in the field of wireless communications is the stipulation that the quantity of user entities must not surpass the count of Radio Frequency (RF) Chains. However, NOMA overcomes this drawback by allowing various users to utilize the equivalent frequency and time resource, which results in higher SE and improved quality of service [6]. Fig. 1 demonstrates the NOMA system scheme [10].

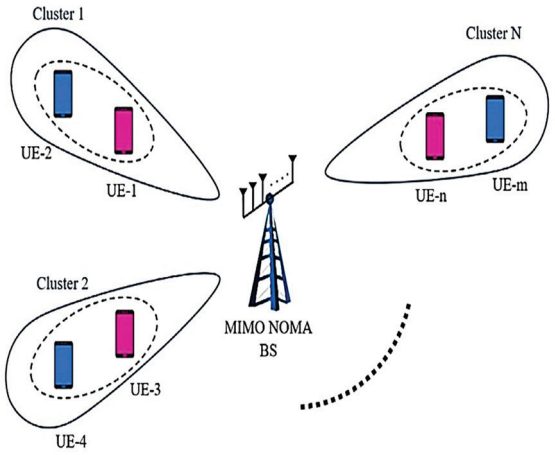

Fig. 1. MIMO-NOMA wireless system scheme [4]

Reinforcement Learning (RL) offers a promising avenue for addressing the complexities of optimizing MIMO-NOMA systems. RL enables an agent to learn optimal strategies through trial-and-error interactions with the environment, receiving feedback as rewards or punishments. Based on this feedback, the algorithm adapts its actions to maximize long-term rewards [11]. Q-learning has garnered considerable focus within the domain of RL algorithms owing to its straightforward implementation and robust performance in addressing intricate challenges. [12].

This study contributes significantly to the field of wireless communication systems, particularly in the optimization of MIMO-NOMA systems through these points:

1. We propose an affordable methodology, employing Q-Learning methodologies to refine power management strategies within the MIMO-NOMA wireless communication systems framework. By applying the Q-learning approach, Our goal is to identify the best power allocation policy to maximize the total data rate while ensuring sufficient levels of EE and SE.
2. We present a detailed system model that articulates the complexities of power allocation within the context of MIMO-NOMA, Which provides a solid foundation for understanding the intricacies of the problem and the motivation behind employing Q-learning as a solution.
3. We delve into the Q-learning approach particularities and demonstrate its application to the power allocation challenge in MIMO-NOMA configurations. Our research highlights the capacity of Q-learning to address complex optimization challenges in diverse network environments.
4. We validate our proposed approach through extensive simulations, providing illustrative results demonstrating significant improvements in the system's performance compared to traditional methods. For example, our methodology achieves an approximate 140% increase in the achievable sum rate compared to conventional NOMA with identical transmitted Power and traditional methods such as that proposed in [13].

The literature review highlights various studies on MIMO-NOMA using RL, such as Q-learning and other approaches, pinpoints areas of incomplete understanding, assesses methodological approaches, and situates our study in the broader field landscape.

This study [14] proposes a novel resource allocation (RA) scheme for massive systems of MIMO-NOMA, leveraging a deep Q-learning network (DQN) and a neural network, which utilizes backpropagation to optimize power management, user grouping, and beamforming. The authors address the significant challenge of downlink RA, aiming to enhance the system's SE while guaranteeing the constraint on the performance of the least efficient user. Their simulated tests indicate that the suggested approach can attain elevated SE for the system, closely mirroring the outcomes of comprehensive searches. This potentially impactful approach could lead to more efficient and reliable wireless communication networks, although the authors acknowledge that further real-world testing is necessary to validate these findings.

The study [15] introduces a deep learning methodology, SARSA λ, for optimizing uplink random access in NOMA-assisted URLLC networks. The algorithm is designed to mitigate decoding inaccuracies in dynamic communication setups and tackles issues related to user grouping, RA optimization, and instantaneous feedback mechanisms. The method reaches convergence within a span of 200 episodes and attenuates the extended average error rate to an order of 10-2. Compared to conventional OMA systems, NOMA-URLLC significantly outperforms error probability and mean error performance over temporal intervals, exhibiting a superiority margin of 70%. The application of Deep RL (DRL) yields superior outcomes compared to both classical and SARSA Q-Learning, manifesting in enhancements of 37% and 38% in average error execution, respectively.

This survey paper [16] explores the role of deep learning (DL) methods in overcoming the limitations of NOMA, a technology pivotal to 5G and beyond 5 G (B5G) development. Despite NOMA's potential for enhancing user connectivity and system efficiency, its practical deployment is constrained by an inflexible design scheme and disparate signal-processing strategies. However, DL-based NOMA can improve key per-

formance indicators such as bit-error rate, throughput, latency, and RA. This analysis underscores DL's transformative capacity to address complex communication challenges and the benefits of its integration with emerging technologies. Future research directions point towards refining DL algorithms for optimized performance, lower latency, and more efficient RA, sparking interest in academic and industrial circles.

Reference [17] the use of RL in managing resources within wireless communication networks, specifically in a single-cell MIMO-NOMA network. The researchers address the challenge of optimizing the total sum rate, a problem complicated by its non-convex nature. They propose an innovative solution that integrates joint beamforming and power allocation by the use of deep DRL. The proposed methodology entails partitioning users into two distinct clusters and formulating an algorithm designed to augment the cumulative data rate for one cluster, while concurrently preserving a minimum threshold rate for the alternate cluster. The authors employ DQN and Double DQN-based algorithms to address this optimization challenge. Empirical findings validate the efficacy of the proposed algorithmic framework, resulting in marked improvements in the cumulative data throughput and rapid stabilization to a steady-state equilibrium.

Authors in [12] suggest a new RA scheme for massive antenna MIMO-NOMA systems using a multi-agent deep Q-network (DQN) algorithm. This approach addresses the slow convergence and suboptimal optimization of traditional algorithms. The researchers create an integrated optimization framework for beamforming, power allocation, and user grouping. Various RL networks are used to allocate Power smartly and group users to improve the system's overall rate. The RA results are fed back into each DQN for iterative optimization. Simulations show that multi-agent DQN improves SE. The study focuses on optimizing both user grouping and power management in massive MIMO-NOMA systems, achieving a balance between power allocation and user clustering while maintaining good performance for weaker users.

In [18], the article explores the susceptibility of NOMA systems to intelligent interference-based attacks using a zero-sum game framework. The base station (BS) determines the transmission power across several antennas as the leader. Conversely, in the follower role, the jammer A Stackelberg equilibrium is attained within the context of the game, considering variables such as the impact of numerous antennas and the states of the radio channels. An RL-based power control strategy is introduced to enhance communication efficiency against intelligent jammers. The hotbooting technique and Dyna architecture are used to speed up the Q-learning-based power management, maximizing NOMA transmission efficiency. Simulations show significant increases in total sum rates of data and user utilities compared to the standard approach. Future work will extend this research to practical situations involving smart interference with multiple jamming policies.

This paper [19] introduces a unique solution for Random Access (RA) optimization in ultra-dense Machine Type Communications (MTC), a central use case for 5G and beyond. The method employs Q-Learning and NOMA, facilitating dynamic RA slot allocation to MTC devices and enhancing network throughput significantly compared to existing techniques. The proposed solution necessitates a minimal increase in complexity on the device end and limited feedback from the BS. Simulations show that a larger discount factor results in better performance, particularly in many-device scenarios, offering faster convergence and improved throughput. This method outperforms existing solutions, promising substantial gains in network performance.

The paper [20], introduces a deep Q-learning (DQL) framework to boost the efficiency of an internal NOMA of Visible Light Communication (NOMA-VLC) downlink network, a crucial component in future wireless communication networks. The focus is on joint power management and Light Emitting Diode (LED) transmission angle tuning, improving challenge, aiming to optimize the average sum rate and EE. Findings suggest that the recommended method substantially enhances the efficiency of NOMA-VLC systems, notably for increased user counts, requiring less computational complexity compared to the "Genetic Algorithm" (GA) and "Differential Evolution" (DE) techniques. Moreover, the combined optimization of power distribution and LED transmission angle gains more effectiveness with growing users, surpassing the traditional solely optimal power allocation method.

This paper [21] proposes a deep Q-learning-based RA approach for uplink NOMA in a cognitive radio network (CRN) to maximize long-term throughput. This work focuses on secondary users (SUs) with limited battery capacity, which can extend their operations using energy harvested from solar sources. The method combines NOMA and "Time Division Multiple Access" (TDMA) to reduce system complexity. As an agent, the Secondary BS (SBS) aims to optimally allocate transmission energy to single users in each time slot through interaction with the system environment. Simulation results indicate that this approach outperforms conventional schemes, enhancing the performance of energy harvesting-powered CRNs over extended operations. The proposed method offers a solution for systems with large state space and action space, presenting the SBS with an optimal power allocation policy learned from environmental dynamics.

Subsequent to this introduction, Section 2 expounds upon our proposed methodology, delineating the Q-Learning approach and its applicability in optimizing power management within a MIMO-NOMA framework. This segment rigorously articulates the problem formulation and its associated mathematical constructs. Section 3 depicts an exhaustive compilation of discussions and results, substantiating the efficacy of our intended scheme in elevating the attainable aggregate data rate, EE, and SE. Comparative analyses with conventional

power management mechanisms are introduced to underscore the merits of integrating Q-learning. In Section 4, we synthesize our conclusions and suggest avenues for prospective research that could further fortify AI techniques' contribution in optimizing wireless communication paradigms. Finally, Section 5 enumerates all the references cited in this manuscript.

2. METHODOLOGY

This section outlines the research methodology, which entails the utilization of the Q-Learning algorithm for power assignment within the scheme of a MIMO-NOMA system.

2.1. THE SYSTEM MODEL OF MIMO-NOMA

In the scrutinized downlink MIMO-NOMA schemes, a multi-antenna is installed at the BS to facilitate serve for multi-user equipment sharing the same frequency and temporal resources. The users are partitioned into two distinct categories: proximal and distant, based on their distance to the BS. For efficacious allocation of transmitted Power, the BS employs a power allocation schema that apportions the available Power among these two classifications of users. The signals the user terminals receive are afflicted by path loss and contaminated by "Additive White Gaussian Noise" (AWGN). Figure 2 represents the system model under discussion of the proposed MIMO-NOMA scheme.

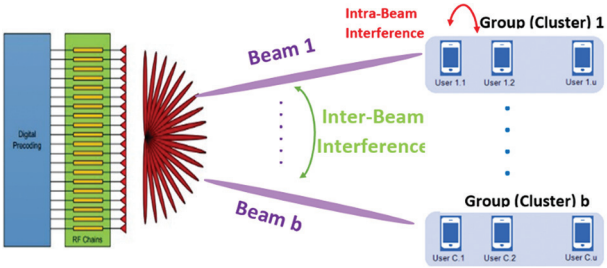

Fig. 2. The system model of the proposed work

2.2. PROBLEM FORMULATION

The intended paradigm seeks to fine-tune the power distribution among proximal and distant users within the wireless communication infrastructure of the MIMO-NOMA system. The overarching goal is to maximize the attainable aggregate data rate of the system. This optimization objective must be realized while concurrently preserving adequate levels of EE and SE.

Let's denote the assigned Power for the proximal and distant users as a_1 and a_2, such that $a_1 + a_2 = 1$. The proximal user also defines the channel gain as g_1 and the distant user as g_2. The Power of noise is represented as no. The achievable rate for the proximal user (R_{1n}) and the distant user (R_{2n}) are determined by the subsequent formulas of Shannon Capacity [3, 22]:

$$R_{1n} = log_2\left(\frac{1+p_t*a_1*g_1}{p_t*a_2*g_1+n_o}\right) \quad (1)$$

$$R_{2n} = log_2\left(\frac{1+p_t*a_2*g_2}{n_o}\right) \quad (2)$$

Where p_t is the total transmitted power from the BS, and log_2 represents the base-2 logarithm.

The total sum-rate (R_n) of the system can be calculated as the sum of R_{1n} and R_{2n} [23]:

$$R_n = R_{1n} + R_{2n} \quad (3)$$

The EE and SE of the system are defined as [13]:

$$EE = \frac{R_n}{p_t+p_c} \quad (4)$$

$$SE = \frac{R_n}{BW} \quad (5)$$

where the circuit power is denoted by P_c is, and the bandwidth (BW) is represented by BW.

The maximization challenge is thus defined as:

Maximize: Optimize the function involving variables a_1, a_2 given a fixed transmission power pt while adhering to constraints C_1, C_2.This can be formally expressed as [23]:

$$Maximize \;\; R_n = log_2\left(\frac{1+p_t*a_1*g_1}{p_t*a_2*g_1+n_o}\right) + log_2\left(\frac{1+p_t*a_2*g_2}{n_o}\right) \; subjected\ to\ C_1, C_2 \quad (6)$$

Where C_1 Indicates that the aggregate of the transmitted Power across all user entities must equate to one. $a_1 + a_2$ =1. C_2 guarantees that the allocated power value of any user in the system must be a positive value.

The issue delineated in Eq. (6) constitutes a non-convex optimization quandary, particularly when the system encompasses more than two user entities. Under such circumstances, the problem escalates in complexity and becomes intractable through conventional optimization techniques. To address this challenge, the present study advocates the employment of the Q-Learning algorithm, a model-free RL methodology, as a viable solution strategy.

2.3. Q-LEARNING ALGORITHM

The Q-Learning algorithm serves as a model-free RL mechanism devised to address the maximization mentioned above challenge. The fundamental premise of the Q-Learning paradigm is to cultivate a policy that directs an agent in choosing apt actions contingent upon particular conditions or scenarios. The Q-learning methodology incorporates a Q-table, a data structure that retains the projected rewards associated with executing specific actions in defined states. The entries within the Q-table undergo iterative modification guided by a predetermined update equation [24]:

$$Q\,[s,a] = Q\,[s,a] + \alpha * (r + y * max\,Q\,[\dot{s},\dot{a}] - Q\,[s,a]) \quad (7)$$

Where,

- Q[s, a] designates the quantified Q-factor corresponding to a specific state-action pair (s, a) encompassing the aggregated reward accrued

through the execution of action (a) within the context of state (s).

- The learning coefficient, symbolized as α, stipulates the extent to which newly assimilated data supersedes extant information throughout the learning trajectory.
- The instant reward, annotated as r, constitutes the value realized subsequent to the state transition from (s) to (s') facilitated by the enactment of action (a).
- The discount coefficient, denoted by y, delineates the relative importance or weighting conferred upon prospective rewards within the framework of the RL procedure.
- The expression "$max\ Q[s', a']$" ignifies the maximal projected subsequent reward attainable upon transition to the subsequent state (s'), given the consideration of all plausible actions (a'). This corresponds to the peak Q-value amongst the set of feasible actions in the ensuing state.

Our Q-Learning approach in MIMO-NOMA systems is based on varying power allocation coefficients, path loss exponents, and critical system performance determinants. The reward structure is centered on the achievable sum rate, incentivizing the algorithm to optimize performance. The Q-Learning framework, a model-free RL technique, learns by interacting with the MIMO-NOMA system environment and updating the Q-table based on the received rewards. This process progressively guides future action selection, culminating in maximizing cumulative reward.

2.4. PERFORMANCE METRICS

The Q-learning performance is evaluated using three main measurements: achievable data rate, EE, and SE. These metrics help quantify how effectively the algorithm improves the system's performance.

Achievable Sum Rate: This metric denotes the total data rate the MIMO-NOMA scheme can uphold, computed as the sum of individual rates for proximal and distant users as per Eq. (3). The main goal is to maximize this sum rate.

Energy Efficiency (EE): This metric evaluates the system's effectiveness in utilizing energy for data transmission. It is calculated as the ratio of the achievable sum rate to the total power expenditure, encompassing both transmit Power and circuit power, as specified in Equation (4). An elevated EE value indicates the system's capability to sustain a greater data rate while maintaining identical power consumption.

Spectral Efficiency (SE): This metric gauges the system's adeptness in capitalizing on the available frequency spectrum for data transmission. Specifically, it evaluates the system's ability to use the allocated BW for information conveyance judiciously. It is mathematically determined as the ratio of the attainable aggregate data rate to the BW, in accordance with Equation (5). A superior SE index signifies the system's proficiency in supporting an elevated data rate while operating within identical BW constraints.

These performance indicators are computed for each conventional MIMO-NOMA system and the intended MIMO-NOMA framework incorporating the Q-Learning methodology. The enhancement observed in these measurements substantiates the efficacy of the advanced Q-Learning-driven power allocation schema. Algorithm 1, delineated below, furnishes a structured procedure for implementing Q-Learning to optimize power allocation and the path loss exponent within the confines of a MIMO-NOMA system. By systematically selecting actions that elevate the Q-value across multiple episodes, the algorithm is poised to identify a policy that could potentially augment the system's performance in the domains of achievable data rates, EE, and SE.

Algorithm 1: Q-Learning Optimization for MIMO-NOMA System

1. Initialization:

- Set the simulation parameters, including distances d_1, d_2, number of users N transmission power p_t, Bandwidth BW, noise power n_o, and circuit power p_c.
- Initialize the Q-learning parameters: learning rate α, discount factor γ, exploration rate ϵ, and number of episodes for Q-learning $n_{episodes}$.
- Define the action space as the product of possible values for the power allocation coefficient a_1 and the path loss exponent η.
- Initialize the Q-table with zero values.

2. For each episode in $n_{episodes}$:

- Initialize the cumulative reward to zero.
- For each user in N
  - Select an action using the epsilon-greedy policy.
  - Extract the power allocation coefficient a_1 and the path loss exponent η from the selected action.
  - Compute the channel gains h_1, h_2 for each user.
  - Calculate the square of the absolute value of the channel gain g_1, g_2.
  - Calculate the achievable sum rate R_1, R_2 for each user and store the mean of the rates.
  - Compute the reward as the mean of the achievable rates.
  - Update the Q-table using the Q-learning update rule.
  - Add the reward to the cumulative reward.
- Store the cumulative reward for the episode.

3. After all episodes:

- Select the optimal action as the one with the maximal Q-value.
- Extract the optimal power allocation coefficient $a_{1_{opt}}$. and the optimal path loss exponent $\eta_{opt.}$ from the optimal action.

- Using the optimal action, calculate the achievable rates, energy efficiency, SE, and signal-to-noise ratio.
- Plot the results and compare them with the system without Q-learning.

3. RESULTS AND DISCUSSION

The ensuing section will elucidate and critically evaluate the results garnered from our investigation, focusing specifically on the efficacy of the Q-Learning paradigm in improving the MIMO-NOMA system. In the context of our examination, we make reference to the system parameters encapsulated in Table 1, enumerating the values for diverse parameters. We postulate that the BS is endowed with flawless Channel State Information (CSI) for all constituents of the user network. This assumption signifies that the BS possesses precise awareness of the CSI for each user, thereby facilitating optimum decision-making and RA paradigms.

Table 1. The simulation system parameters of the proposed system

Parameter	Description	Value(s)
d_1	Separation metric among user '1' and BS	200
d_2	Separation metric among user '2' and BS	500
N	The system's user count	1000
P_t	Transmission power magnitudes expressed in dBm	[-30, -25, -20, -15, -10, -5, 0, 5, 10, 15, 20, 25, 30]
BW	Bandwidth	1 MHz
N_o	The power of Noise Magnitude Specified in dBm	-114 dBm
P_c	Circuit consumed Power	100 W
alpha	The learning rate of Q-learning	0.5
gamma	Discount coefficient of Q-learning	0.95
epsilon	The Exploration rate of Q-learning	0.1
episodes	Number of Q-learning episodes	1000
etas	Path loss exponent (eta) discretized values	[2.0, 3.0, 4.0, 5.0, 6.0]
a_{1s}	Power allocation coefficient (a1) discretized values	[0.5, 0.6, 0.7, 0.8, 0.9]
Original	The conventional NOMA standard from reference [13] is used as the benchmark for comparing our results.	-

Our Q-learning approach for MIMO-NOMA systems provides adaptability through learning from system experiences, simplifies complex optimization by learning a policy mapping states to actions, and ensures scalability by efficiently handling large action spaces.

3.1. ACHIEVABLE SUM RATE

The inaugural graph delineates the attainable aggregate data rates corresponding to each of the conventional and Q-Learning methodologies, plotted corresponding to transmission power. Examination of the graphical representation reveals a direct correlation between the sum rates and transmit Power levels for both methods. Notably, the Q-Learning algorithm consistently surpasses the performance metrics of the conventional approach across the entire spectrum of transmission power magnitudes. This observation substantiates the efficacy of the Q-Learning paradigm in the realm of power assignment for sum-rate optimization. Figure 3 elucidates the achievable data rate corresponding to varying magnitudes of transmission power; the intended scheme manifests a convergent augmentation of 140% in comparison to conventional NOMA techniques when evaluated at identical transmitted power levels.

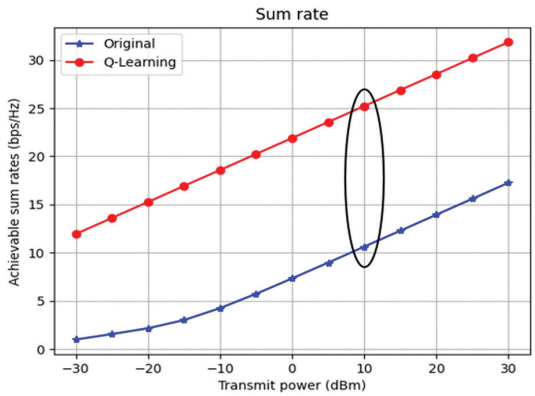

Fig. 3. The achievable data rate parameterized corresponding to the Power transmitted

3.2. SIGNAL-TO-NOISE RATIO

The following diagram illustrates the SNR for each conventional and Q-Learning scheme, plotted against transmit Power, assuming perfect CSI for all users. Both plots reveal a comparable direct correlation between transmitted Power and SNR, with no significant differences in SNR between the methods. Fig. 4 compares the SNR for the benchmark and proposed systems, confirming that SNR is predominantly contingent upon the transmission power and channel conditions, variables that remained invariant across both methodologies. This suggests that the Q-Learning approach can improve specific performance metrics without negatively affecting other metrics.

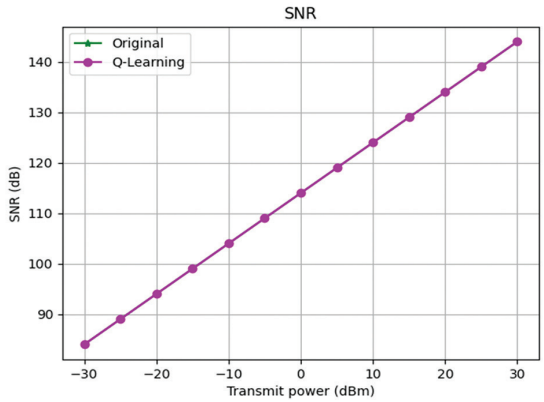

Fig. 4. SNR parameterized as a function of Power transmitted

3.3. ENERGY EFFICIENCY

The third chart compares the EE for the conventional and Q-Learning methods, with EE plotted against transmit Power in Fig. 5. The empirical findings reveal that the Q-Learning methodology uniformly eclipses the traditional approach in the realm of EE, irrespective of the magnitude of transmit Power. This suggests that the Q-Learning-based power management yields higher data rates while keeping power consumption constant. Notably, at 20 dB of transmission power, the intended Q-Learning method outstrips the conventional by an appreciable performance differential of approximately 93%. These outcomes underscore the advantages of integrating Q-Learning into power management, enabling more efficient power utilization while achieving superior data rates.

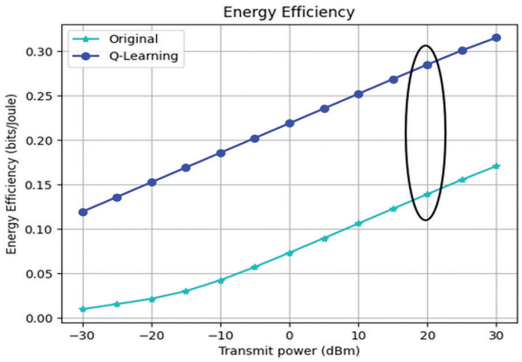

Fig. 5. The EE of the proposed scheme in contrast to Traditional NOMA

3.4. SPECTRAL EFFICIENCY

The next diagram delineates the SE metrics for both the benchmark and proposed systems employing the Q-Learning algorithm. Figure 6 portrays that the Q-Learning methodology consistently registers higher SE across all levels of transmitted Power. This attests to the efficacy of the Q-Learning algorithm in maximizing the data-rate within the identical BW allocation, highlighting its utility in improving MIMO-NOMA systems. Specifically, at a transmit power setting of 30 dB, the SE of the proposed Q-Learning framework surpasses that of the conventional NOMA approach by an appreciable margin of approximately 88.6%.

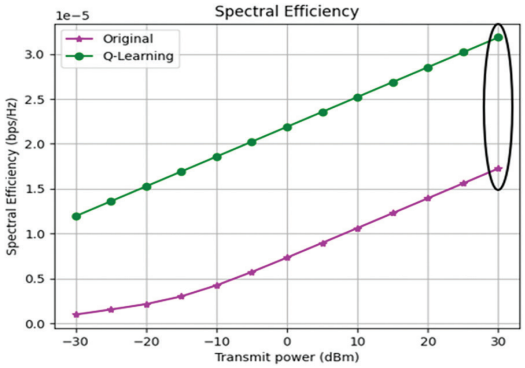

Fig. 6. The SE of the Intended Paradigm in Comparison with Conventional NOMA

3.5. IMPROVEMENT WITH Q-LEARNING

Two graphical representations, shown in Figs. 7 and 8, elucidate the advancements achieved via the Q-Learning algorithm. These figures empirically corroborate the pronounced improvements in both performance metrics, juxtaposing the Q-Learning methodology with the conventional approach across a diverse range of transmit power levels. The Q-Learning algorithm manifests the most salient augmentation in EE within the power range of 0 to 20 dB. Concurrently, the most notable enhancement in the sum rate is observed within the power spectrum of -5 to 30 dB.

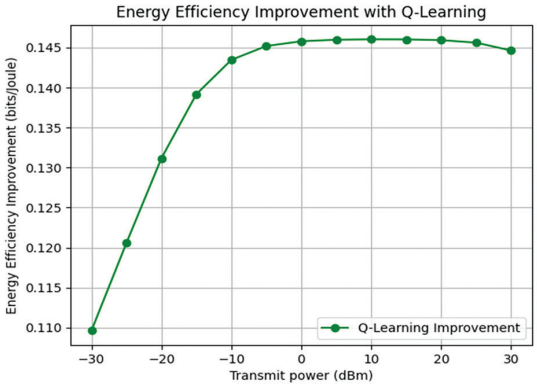

Fig. 7. The EE Enhancement Attributable to the Implementation of the Q-Learning Algorithm

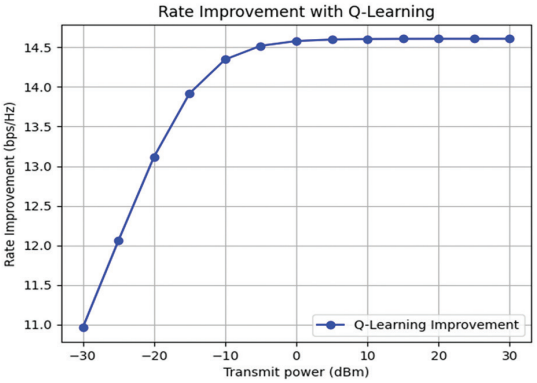

Fig. 8. The Increment in Achievable Sum Rate by Adopting the Q-Learning Algorithm

3.6. ACCUMULATED REWARD PER EPISODE

The final figure, denoted as Fig. 9, exhibits the accumulated reward accrued per episode within the context of the Q-Learning methodology. The figure reveals an ascending trajectory of rewards as a function of the episode count, substantiating that the algorithm is undergoing a learning process and progressively enhancing its operational performance. The incremental trend in the cumulative reward serves as empirical evidence of the Q-Learning algorithm's effectiveness in discerning an optimal power allocation strategy.

To sum up, the intended Q-Learning-based power management methodology engenders substantial enhancements in the operational execution of the MIMO-

NOMA paradigm, particularly in the domains of achievable data-rates, EE, and SE. Despite the constancy in SNR, this attribute does not undermine the merits of the Q-Learning Algorithm. Prospective study initiatives may contemplate the investigation of alternative RL algorithms or scrutinize diverse architectural variations of MIMO-NOMA systems.

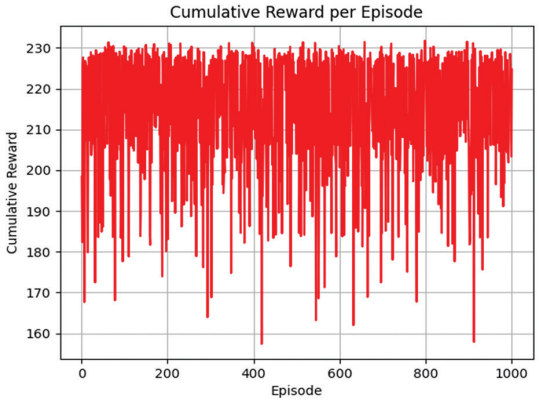

Fig. 9. The number of cumulative rewards per episode

4. CONCLUSION

In this research, we proposed and meticulously investigated an AI-based reinforcement Q-Learning approach for performance optimization in MIMO-NOMA wireless communication systems. Our Q-Learning approach in MIMO-NOMA systems is based on varying power allocation coefficients, path loss exponents, and key system performance determinants. The reward structure is centered on the achievable sum rate, incentivizing the algorithm to optimize performance. The Q-Learning framework, a model-free RL technique, learns by interacting with the MIMO-NOMA system environment and updating the Q-table based on the received rewards. This process progressively guides future action selection, culminating in maximizing cumulative reward.

Our results showed a substantial enhancement in the achievable sum rate, EE, and SE, compared to traditional MIMO-NOMA systems. We got 140%, 93%, and 88.6% higher performance for each previously mentioned performance metric, respectively, compared with traditional NOMA systems; this underscores the capacity of our Q-Learning approach to adaptively manage power between users, effectively optimizing the trade-off between these critical system performance metrics.

Based on the successes of this study, future work could expand in several potential directions. Investigating other RL methodologies beyond Q-Learning, including deep reinforcement learning or policy gradient techniques, could be compelling in improving the performance of MIMO-NOMA systems to a greater extent. Second, our approach could be applied to more complex communication scenarios, including massive MIMO systems or cooperative NOMA networks, offering insights into the scalability and adaptability of AI-based power allocation. Lastly, with the advent of 5G and beyond wireless networks, integrating our AI-based methodology with other emerging technologies like edge computing or the Internet of Things (IoT) could be examined to foster comprehensive and efficient solutions for future wireless communication systems. Furthermore, The implications of the research on practical applications and real-world deployments could be explored.

5. REFERENCES

[1] S. P. Yadav, "Performance Optimization of Universal Filtered Multicarrier Technique for Next Generation Communication Systems", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 2, 2023, pp. 119-127.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G", IEEE Communications Surveys & Tutorials, Vol. 20, No. 3, 2018, pp. 2294-2323.

[3] A. A. Majeed, I. Hburi, "Beamspace-MIMO-NOMA Enhanced mm-Wave Wireless Communications: Performance Optimization", Proceedings of the International Conference on Computer Science and Software Engineering, Dohuk, Iraq, 15-17 March 2022, pp. 144-150.

[4] A. F. Banob, F. W. Zaki, M. M. Ashour, "The effect of quantized ETF, grouping, and power allocation on non-orthogonal multiple accesses for wireless communication networks", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No, 8, 2022, pp. 681-693.

[5] C. Ben Issaid, C. Anton-Haro, X. Mestre, M.-S. Alouini, "User Clustering for MIMO NOMA via Classifier Chains and Gradient-Boosting Decision Trees," IEEE Access, Vol. 8, 2020, pp. 211411-211421.

[6] A. A. Majeed, I. Hburi, "Energy-Efficient Optimization of mm-Wave Communication Using a Novel Approach of Beamspace MIMO-NOMA", Wasit Journal of Engineering Sciences, Vol. 10, No. 2, 2022, pp. 223-239.

[7] Z. Shi, H. Wang, Y. Fu, G. Yang, S. Ma, F. Hou, T. A. Tsiftsis, "Zero-Forcing Based Downlink Virtual MIMO-NOMA Communications in IoT Networks", IEEE Internet of Things Journal, Vol. 7, No. 4, 2020, pp. 2716-2737.

[8] U. Ghafoor, M. Ali, H. Z. Khan, A. M. Siddiqui, M. Naeem, "NOMA and future 5G & B5G wireless networks: A paradigm", Journal of Network and Computer Applications, Vol. 204, 2022.

[9] S. Mahyar, M. Dohler, S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations", IEEE Communications Magazine, Vol. 55, No. 9, 2017, pp. 55-61.

[10] A. Akbar, S. Jangsher, F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques", Computer Networks, Vol. 190, 2021.

[11] J. Clifton, E. Laber, "Q-Learning: Theory and Applications", Annual Review of Statistics and Its Application, Vol. 7, No. 1, 2020, pp. 279-301.

[12] C. Yanmei, W. Lin, L. Jiaqing, Z. Jun, Y Lin, Z, Guomei, "Joint Resource Allocation Scheme Based Multi-agent DQN for Massive MIMO-NOMA Systems", Proceedings of the 14th International Conference on Communication Software and Networks, Chongqing, China, 10-12 June 2022, pp. 51-55.

[13] I. Hburi, H. F. Khazaal, N. M. Mohson, T. Abood, "MISO-NOMA Enabled mm-Wave: Sustainable Energy Paradigm for Large Scale Antenna Systems", Proceedings of the International Conference on Advanced Computer Applications, Missan, Iraq, 25-26 July 2021, pp. 45-50.

[14] Y. Cao, G. Zhang, G. Li, J. Zhang, "A Deep Q-Network Based-Resource Allocation Scheme for Massive MIMO-NOMA", IEEE Communications Letters, Vol. 25, No. 5, 2021, pp. 1544-1548.

[15] W. Ahsan, W. Yi, Y. Liu, A. Nallanathan, "A Reliable Reinforcement Learning for Resource Allocation in Uplink NOMA-URLLC Networks", IEEE Transactions on Wireless Communications, Vol. 21, No. 8, 2022, pp. 5989-6002.

[16] S. A. H. Mohsan, Y. Li, A. V. Shvetsov, J. V. Aldás, S. M. Mostafa, A. Elfikky, "A Survey of Deep Learning Based NOMA: State of the Art, Key Aspects, Open Challenges and Future Trends", Sensors, Vol. 23, No. 6, 2023.

[17] T. Lu, H. Zhang, K. Long, "Joint Beamforming and Power Control for MIMO-NOMA with Deep Reinforcement Learning", Proceedings of the IEEE International Conference on Communications, Montreal, Canada, 14-23 June 2021, pp. 1-5.

[18] L. Xiao, Y. Li, C. Dai, H. Dai, H. V. Poor, "Reinforcement Learning-Based NOMA Power Allocation in the Presence of Smart Jamming", IEEE Transactions on Vehicular Technology, Vol. 67, No. 4, 2018, pp. 3377-3389.

[19] M. V. da Silva, R. D. Souza, H. Alves, T. Abrão, "A NOMA-Based Q-Learning Random Access Method for Machine Type Communications", IEEE Wireless Communications Letters, Vol. 9, No. 10, 2020, pp. 1720-1724.

[20] A. A. Hammadi, L. Bariah, S. Muhaidat, M. Al-Qutayri, P. C. Sofotasios, M. Debbah, "Deep Q-Learning-Based Resource Allocation in NOMA Visible Light Communications", IEEE Open Journal of the Communications Society, Vol. 3, 2022, pp. 2284-2297.

[21] H. T. Huong Giang, P. Duy Thanh, I. Koo, "Dynamic Power Allocation Scheme for NOMA Uplink in Cognitive Radio Networks Using Deep Q Learning", Proceedings of the International Conference on Information and Communication Technology Convergence, Jeju Island, Korea, 21-23 October 2020, pp. 137-142.

[22] C. L. Wang, Y. C. Wang, P. Xiao, "Power Allocation Based on SINR Balancing for NOMA Systems with Imperfect Channel Estimation", Proceedings of the International Conference on Signal Processing and Communication Systems, Surfers Paradise, Australia, 16-18 December 2019, pp. 1-6.

[23] D. Tse. P. Viswanath, " Fundamentals of Wireless Communication", Cambridge University Press, 2005, pp. 228-289.

[24] B. H. Abed-Alguni, D. J. Paul, S. K. Chalup, F. A. Henskens, "A comparison study of cooperative Q-learning algorithms for independent learners", International Journal of Artificial Intelligence, Vol. 14, No. 1, 2016, pp. 71-93.