Proceedings of the 2nd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/95/2024.21338

# Power allocation based on reinforcement learning in 5G/B5G multi-cell networks

**Kewei Ren**

School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu, P.R.China

2021100901014@std.uestc.edu.cn

**Abstract.** The fifth generation of wireless communication technology (5G), has revolutionized the digital landscape with its ultra-fast speeds, massive connectivity, and reduced latency; Beyond 5G (B5G), represents the evolutionary steps towards sixth-generation networks which aims to build upon 5G's capabilities by integrating advanced technologies like ultra-dense network, edge computing, and enhanced spectral efficiency. This paper investigates the application of two distinct exponential reward functions in reinforcement learning algorithms for power allocation in ultra-dense networked base stations. The primary objective is to maximize the overall network capacity and spectral efficiency. The performance of the proposed reinforcement learning algorithms is compared with the traditional water-filling algorithm, as well as against the other to highlight the differences in learning outcomes resulting from the choice of reward functions. The results show that the exponential function model with reciprocal exponent is superior to the previous two in spectral efficiency and convergence speed and provide valuable insights into the effectiveness of using reinforcement learning for complex resource allocation problems in modern communication networks.

**Keywords:** Reinforcement learning, ultra-dense networks, reward functions, power allocation

## 1. Introduction

The rapid growth of wireless data traffic and the increasing number of mobile devices in 5G/B5G (the fifth generation of wireless communication technology/Beyond 5G) network systems have posed significant challenges to efficient power allocation in multi-cell networks. Traditional optimization methods, such as water-filling algorithms, struggle to handle the complexities and dynamics of these modern communication systems. Consequently, there is a growing interest in the application of reinforcement learning (RL) for power allocation in 5G/B5G multi-cell networks. This paper provides an overview of the current state of research on power allocation based on RL, identifies the limitations and challenges in this area, and discusses the inspiration for this research topic.

DOI: 10.54254/2753-8818/95/2024.21338

RL is a learning paradigm that enables an agent to make decisions in an environment to maximize its cumulative reward [1]. In the context of power allocation in multi-cell networks, RL can be used to optimize the allocation of power across different base stations (BSs) and user equipment (UE), taking into account the dynamic nature of the wireless channel and the varying demands of users.

Research on power allocation based on RL in 5G/B5G multi-cell network has shown promising results.

Studies have proposed RL-based algorithms for power allocation, including Q-learning, actor-critic methods, and policy gradient methods [2]. These algorithms have been shown to achieve better performance than traditional optimization methods in terms of network capacity, spectral efficiency, and fairness among users [3].

However, there are several limitations and challenges in the current research on power allocation based on RL. First, the large state and action spaces in multi-cell networks make it difficult for RL algorithms to converge to an optimal solution [4]. Second, the non-convex and non-smooth nature of the optimization problem poses great challenges to the stability and convergence of RL algorithms. Third, the high computational complexity of RL algorithms limits their real-time application in practical communication systems [5].

This article investigates the dynamic power allocation problem in downlink cellular networks based on multi-agent reinforcement learning, where each BS user is modeled as a RL agent to learn optimal power allocation policy in order to maximize the total system capacity. The study addresses the expandability of reward function and state, in order to adapt the variation of network size, such as the number of BSs or UEs and the coverage area of cells [6].

To empirically substantiate the efficacy of our advanced proposal, we undertake exhaustive computational simulation experiments within a milieu representative of a 5G/B5G multicellular network infrastructure. Our findings illustrate that the proposed methodology surpasses contemporary state-of-the-art reinforcement learning algorithms in achieving superior network throughput, spectral utilization efficiency, and user-centric fairness indices.

In drawing our study to a close, we furnishes a panoramic survey of ongoing scholarly endeavors concerning reinforcement learning-guided power allocation strategies in the context of 5G/B5G multi-cell networks. It pinpoints the extant challenges confronted within this disciplinary terrain and contemplates the intellectual stimuli that animate this investigatory domain. By advancing a novel scheme for refining the reward mechanism, our simulative analyses corroborate the practical virtue and validity of the espoused approach, thereby enriching the corpus of knowledge in this nascent yet rapidly evolving frontier of telecommunications research.

In this paper, we aim to address the limitations and challenges in the current research on power allocation based on RL. Specifically, we focus on the optimization of RL for power allocation in 5G/B5G multi-cell networks:

- We propose a novel method for designing RL that takes into account the dynamic nature of the wireless channel and the varying demands of users.
- Our method is based on the concept of fairness and efficiency, and it aims to maximize the network capacity and spectral efficiency with the concurrent assurance of equitability amidst all network participants.
- Our modeling also ensures that the entire system can converge at a more accelerated rate compared to traditional models, thereby achieving optimal performance in a timelier manner.

## 2. System Model and Problem Formulation

### *2.1. System Model*

The system conceptualized in the ultra-dense network can be represented as a multi-cell OFDM cellular network whose serving scenario is mainly responsible for downlink services, characterized

DOI: 10.54254/2753-8818/95/2024.21338

by a high density of BSs. It operates on the reuse-1 principle, granting each BS access to the entire available bandwidth [7]. The network topology is visualized in Fig. 1. A central controller is responsible for aggregating comprehensive network metrics, including signal-to-interference-plus-noise ratio (SINR) and transmit power. The number of users associated with BS n at time t is denoted as

$$UE_{(n,t)} \in \{M_{(1,t)}, \dots, M_{(n,t)}, \dots, M_{(N,t)}\}, \quad (1)$$

where M represents the total user population. User mobility is modeled using a random walk approach, with the user's speed at time t denoted as

$$0 \le V_{m,t} \le V_{max}, \quad (2)$$

and the movement angle as

$$0 \le D_{m,t} \le 2\pi. \quad (3)$$

The spatial distribution of BSs adheres to a Poisson point process model. Each BSn reuses K orthogonal subcarriers in their entirety. When a user connects to a BS, the BS is activated and its transmit power is set to $p_{n,t}$ at any given time. A user can only connect to one BS, and each subcarrier is exclusively allocated to a single user. The six-path fading channel model is utilized for performance evaluation purposes, and the selection of this particular channel model does not influence the efficacy of the proposed approach [8].

### *2.2. Problem Formulation*

We can define that $\psi_{n,k,m}$ represents the received SINR of the n-th BS served user $m$ on the $k$-th subcarrier at time t , in which $k \in \{1, \dots, K\}$ and is given by

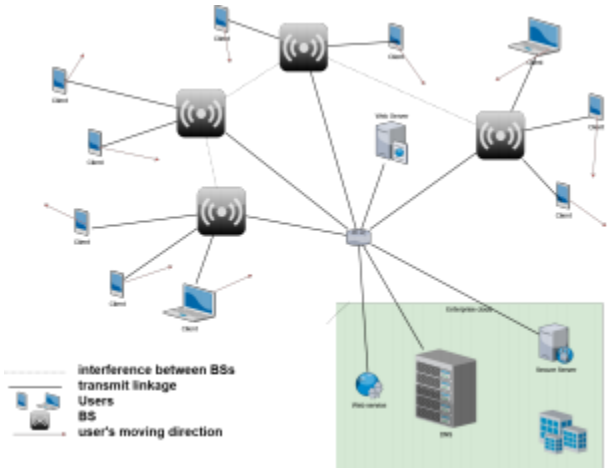



**Figure 1.** System model of ultra-dense network.

$$\psi_{n,k,m} = \phi_{n,m}\phi_{n,k,m}\frac{G_{n,m}^{(k)}p_{n,k}}{\sum_{n' \ne n} G_{n',m}^{(k)}p_{n',k}, +\sigma^2} \quad (4)$$

in which $G_{n,m}^{(k)}$ and $G_{n',m}^{(k)}$ signifies channel gains of the n-th and $n'$-th BSs to user $m$ on the $k$-th subcarrier. And $p_{n,k}$and $p_{n',k}$ denote the total transmit power of the $n'$ BSs on the $k$-th subcarrier respectively. $\phi_{n,k,m}$ denotes whether BS $n$ allocates subcarrier $k$ to user m $\phi_{n,k,m} \in [0, 1]$, and $\sigma^2$ represents the power of Gaussian white noise. $\phi_{n,m}$ indicates whether user m is connected to BS $n$:

DOI: 10.54254/2753-8818/95/2024.21338

$$\phi_{n,m} = \begin{cases} 1, & m \in M_i \\ 0, & m \notin M_i \end{cases} \tag{5}$$

The spectral efficiency is scrutinized through the prism of the aggregated capability, quantified in terms of (bps/Hz), thereby furnishing a metrical gauge of the communication efficiency underpinning the reinforcement learning-driven power allocation strategy. The capacity achieved by $BS_n^f$ at its associated user on subcarrierk is given by:

$$C_{(n,k)} = \frac{B}{K}\log_2(1 + \sum_{m=1}^{M} \psi_{n,k,m}) \tag{6}$$

Our endeavor is geared towards augmenting the aggregate capacitance of the holistic network, achieved through the meticulous calibration of BS transmission potencies on subcarrier $p_{n,k}$, guided by an approach verging on subcarrier allocation optimality. This endeavor of enhancement can be encapsulated in the following formulization of the optimization conundrum:

$$\begin{cases} \arg\max R_t \\ s.t.\ C1: p_t^{(n,k)} \geq p_{min} \quad \forall n,k \\ \quad C2: \sum_{n,k} p_t^{(n,k)} \leq p_{max} \quad \forall n,k, \end{cases} \tag{7}$$

within which $p_{max}$ denotes the paramount emission potency of the BS, whereas $p_{max}$ signifies the least possible transmission power allocated to a subcarrier.

In the preliminary phase of the network setup, users establish connections with BS by selecting the one that offers the highest SINR. During this workflow, the interference affecting users is predominantly caused by signals from neighboring cells, and not from users within the same cell [9]. The transmission rate of each user link is influenced by the power of the signal from the serving BS and the level of interference from other cells. Our objective is to fine-tune the BS transmit power to enhance the aggregate capacity of the network. We ensure equitable resource distribution among users by evenly distributing the BS's downlink subcarriers to all connected users. To kickstart the power allocation, we initially deploy the water-filling algorithm that could serve as the benchmark for subsequent comparison when employing ensuing reinforcement learning. The problem at hand presents a multi-objective non-convex optimization challenge. Conventionally, heuristic search algorithms, which are based on iterative local searches, have been employed to address such issues [10]. However, these algorithms often suffer from significant inefficiencies, characterized by extended computation times and an inability to adapt in real time. To overcome these limitations, the subsequent chapter will delve into the operationalization of reinforcement learning, a cutting-edge technique that harnesses neural networks to optimize policies through continuing trial and error.

## 3. Q-Learning Solution

### *3.1. Reinforcement Learning - Multi-Agent Q-learning*

The paradigm of disseminated cognitive BSs finds mathematical embodiment in stochastic games, wherein the educative progression of every BS is encapsulated by a quintuple

$$N, S, A, P, R(s, \vec{a}), \tag{8}$$

explicating thusly:

- $N$, an ensemble denoting $1, 2, \ldots, N_f$, symbolizes the collective of actors resembling BSs.
- $S$, a compendium comprising $S_1, S_2, \ldots, S_m$, delineates the gamut of potential systemic states, with m signifying the count thereof.

DOI: 10.54254/2753-8818/95/2024.21338

- $A$, a set encompassing $a_1, a_2, \ldots, a_l$, embodies the spectrum of feasible maneuvers.
- $P$, the probabilistic displacement function, quantifies the likelihood of systemic transition between states.
- $R(s, \vec{a})$, as the reward function, ascertains the boon accruing to actor n upon execution of the concerted action $\vec{a}$ amidst a state $s \subseteq S$.

Q-learning, a model-extraneous reinforcement learning protocol grounded in value function theory, innately excels in negotiating the mutable landscape of wireless networks [10]. This algorithm facilitates each agent's convergence upon its proprietary behavioral value function via incessant recursive tutelage. Typically, this behavioral metric is instantiated as a tabular construct, $Q(s_m, a_l)$, with al inhabiting A and $sm$ nested within S, rendering a matrix dimensionality of $m \times l$. The Q-value, $Q(s_m, a_l)$, signifies the anticipated aggregation of rewards, spanning an infinite temporal horizon, derivable from executing action al in the state $s_m$, and is mathematically rendered by:

$$Q(S, A) = E_\pi \left[ R_a + \gamma \max_{a'} maxQ(S', a') | S_t = S, A_t = A \right] \tag{9}$$

where $R_{t+1}$ symbolizes the immediate gratification attained subsequent to the execution of an action a within the prevailing environmental condition, or state, $s$. The variable $\gamma$, meticulously confined to the interval [0,1], serves as the temporal discount factor, ascertaining the magnitude of emphasis placed upon the anticipated future returns, or the heuristic value, thereby acting as a modulator of the present reward's immediacy versus future rewards' prospective value in the algorithm's decision-making calculus. This coefficient $\gamma$ essentially reflects a balancing act between myopic and far-sighted strategies, infusing the reinforcement learning process with a temporal perspective that is pivotal for efficacious learning and decision optimization in dynamic environments.

The Q-value of the whole allocation system is updated in terms of the equation below:

$$Q(S, A) \overset{\pi}{\Rightarrow} \max_{A'} [R_a + \gamma Q(S', A')], \tag{10}$$

$$New\ Q_{S,A} = Q_{S,A} + \alpha \left( R_{S,A} + \gamma * maxQ' \left( S', A' \right) - Q_{S,A} \right). \tag{11}$$

In this formula, $\alpha$ means the learning rate whose range is $[0,1]$.

### *3.2. Power Allocation and Application of Q-Learning Mapping*

In the practical problem of power allocation of multi-agent, the conception in Q-learning, the agents, states, actions, and the reward function can be externalized:

- Agent: that is BS n, $1 \leq n \leq Nf$.
- $s_t^{n,k} = \{L_t^n, p_t^n\}$, where represents the number of users connected to the certain BS $n$ at certain time $t$, and represents the power of the n-th BS at time $t$. With the intent of mitigating the algorithmic intricacy and circumscribing the expanse of the network's state space, a stratagem of quantizing the BS's transmit power is henceforth enacted, proceeding as follows:

$$p_t^n = \tau \left( P_{max}^f - S_\tau \right) \leq \sum_{k=0}^{K} p_t^{n,k} < \left( P_{max}^f - S_{\tau+1} \right) \tag{12}$$

where $\tau \in \{0, 1, 2, 3, 4, 5\}$, $S_0 = P_{max}^f$, $S_6 = 0$, other $S$ values are selected threshold.

DOI: 10.54254/2753-8818/95/2024.21338

- Action: $A_n = \{i_n, \overrightarrow{pt}^{(n,k)}\}$, $i_n$ indicates the k-th subcarrier of the n-th BS, as well as $\overrightarrow{pt}^{(n,k)} \in \{-|P|, 0, |P|\}$. Its determination is contingent upon the metamorphosis in the aggregate throughput $C_{th}$ experienced by the environing $h$ users in proximity to the incumbent user, subsequent to the execution of an action. Hence, should C witness an amplification, the power increment $\overrightarrow{pt}^{(n,k)} = +|P|$ ensues; conversely, if a diminution occurs, the inverse adjustment is invoked. This adaptive modulation of power levels operates as a pivotal feedback loop within the reinforcement learning framework, fostering a strategy that dynamically aligns with the exigencies of maximizing network efficiency.
- Reward: the selection of reward function will be introduced in section C in detail.

In light of the environmentally dynamic context, the state space of the system is relatively extensive, and the dimensions of the $Q$-value table for each agent vary. If one were to establish a $Q$-value table of considerable size with a static dimensionality, the complexity of computations would ascend dramatically.

Consequently, a strategy for dynamically incorporating states is imperative, wherein novel states are automatically integrated into the state sets as they emerge. This obviates the necessity of tailoring a $Q$ table for each single agent, a decided advantage of the dynamic approach. Additionally, it facilitates a more efficient search of $Q$ tables and optimizes the utilization of storage space [11]. What's more, the reward table updates in $Q$-learning are relatively slow, hence the choice of the reward function critically impacts performance. The selection of the reward function can alter the learning path and approach of the agent, consequently affecting the rate of convergence and overall capacity of the model.

### *3.3. Comparison on the Attribution of Different Reward Functions*

Within the ambit of the Q-learning algorithm, paramountcy resides in the reward function, whose office and stature are of great significance. Its formulation bears a direct umbilical to the algorithm's triumph, acting as the lodestar in the navigational journey towards optimal strategizing. This function, in essence, serves as an arbiter, evaluating the instant aftermaths of an agent's singular maneuver within a prescribed environmental condition, thereby steering the agent's erudition towards actions of superlative efficacy. It imparts to the agent an instantaneous missive, echoing the ecosystem's resonant feedback —— a cornerstone for the agent's calibration and juxtaposition of disparate actions' worthiness.

The agent's actions revolve around the maximization of accrued rewards, rendering the reward function a crucible that molds the agent's discernment of praiseworthy and censurable maneuvers. Furthermore, the sinew of the recompense function's blueprint reverberates in the Q-learning algorithm's velocity of convergence and its steadfastness. A flawed reward signal's architecture may engender a sluggish, capricious learning odyssey, or worse, an inability to attain the paragon of strategies [12].

In the purview of the study, an exponential function has been elected as the archetype for $t$ he reward function, a choice lauded for its prowess in hastening convergence and ensuring a polished trajectory, thereby enhancing the algorithmic voyage with a refined, efficacious compass.

$$y_1 = 1 - e^{(-\alpha x)}, \tag{13}$$

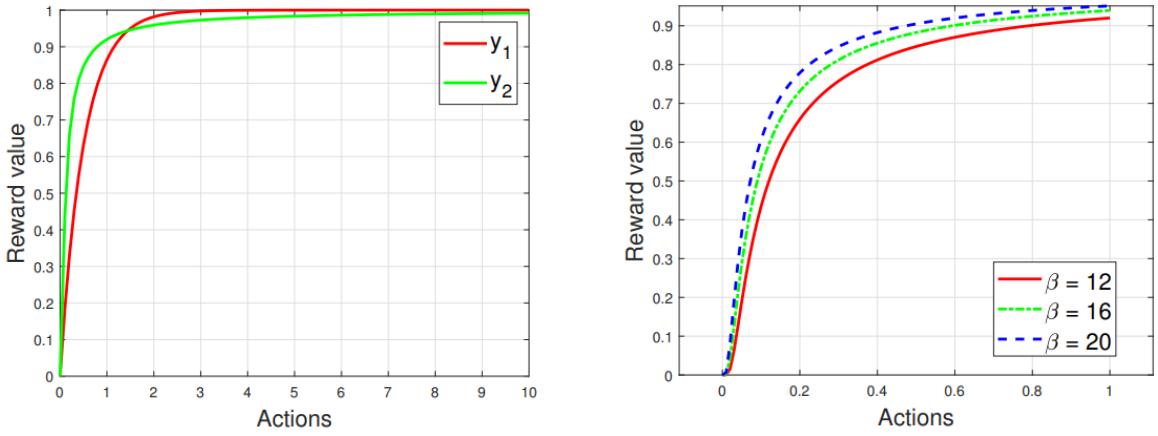
$$y_2 = e^{\left(-\frac{1}{\beta x}\right)}, \tag{14}$$

In this problem, two different typical exponential function models are selected above for Q-learning training.

Both functions have the following properties in common that are suitable for reinforcement learning of power allocation :

DOI: 10.54254/2753-8818/95/2024.21338

- Non-negative: Since the exponential function always has a value in the range (0,1), the value of function $y$ is always non-negative.
- Increment: As the independent variable $x$ increases, the value of the two exponent gradually decreases, so $y$ gradually increases. This means that $y_1$ and $y_2$ are a monotonically increasing function.
- Saturation and convergence: As $x$ approaches infinity, the value of exponent goes verge to zero, so the function value approaches one. This indicates that function $y_1$ and $y_2$ have an initial value of zero as well as an upper limit of one.




(a) Comparison of two reward function model. (b) Comparison of the model y2 with distinct β.

**Figure 2.** Comparison of different reward function.

As can be seen in Fig. 2(a). above, these properties allow reinforcement learning to proceed smoothly: since these functions are monotonically increasing, the agent can receive a clear reward signal, that is, as the effect of the action improves, the reward gradually increases. However, there is a certain difference in the rate of rise of these two exponential function models, which typically determines the learning path of the agent:

- Pure exponential form $y_1$: The model promotes early exploration, with a faster reward growth step when $x$ is small to incentivize the agent to quickly achieve a larger reward in the initial learning stage. As $x$ increases, the reward growth gradually saturates and slows down, allowing the agent to stabilize its behavior after reaching a certain level of performance and preventing overfitting by avoiding an unlimited pursuit of higher reward.
- Exponential form including reciprocals type $y_2$: The variable $y_2$ provides the agent with a robuster early reward signal, and as x increases, the rate at which the reward signal diminishes surpasses that of $y_1$. This prompts the agent to swiftly acquire rewards while exploring new actions, but subsequently reduces exploration more rapidly as it approaches saturation.

Furthermore, these two models also demonstrate adaptability. By adjusting the parameters in the model (such as α and β), the shape of the reward function can be modified to accommodate various learning tasks and environments. Moreover, by altering these parameters, the smoothness of the reward function and the distribution of rewards can be flexibly adjusted. In this study, we only modify the second exponential model $y_2$, and choose the parameter $\beta = 12,16,20$ for training and comparison. Their plots are showed in Fig. 2(b):

When reflected in the practical problem of BS power resource allocation, the reward function of the n-th BS on the k-th subcarrier can be respectively defined as:

DOI: 10.54254/2753-8818/95/2024.21338

$$r_t^{n,k} = \begin{cases} 1 - e^{-\left(\alpha C_t^{(n,k)}\right)}, & if \sum_{k=1}^{K} p_t^{(n,k)} \le P_{max}^f \\ -1, & otherwise \end{cases} \quad (15)$$

$$r_t^{n,k} = \begin{cases} e^{-\frac{1}{\beta C_t^{(n,k)}}}, & if \sum_{k=1}^{K} p_t^{(n,k)} \le P_{max}^f \\ -1, & otherwise \end{cases} \quad (16)$$

## 4. Simulation and Evaluation

### *4.1. Settings of training*

The entire cellular network environment includes $N_f$ BSs. Each user connects to the BS with the highest SINR. All BSs share the spectrum bandwidth. In the simulation, we set the noise power $\sigma^2 = 10^{-7}$, and the power regulation amplitude of each BS on each subcarrier, where $S_1 - S_5 = [25, 20, 15, 10, 5]$. Some other parameters settings are shown in Table 1 and $\beta$ is updated as the method in Table 2. At each step, the network updates the topology once due to the node mobility.

### *4.2. Consequence and comparison of training model*

The convergence of the frequency efficiency models of the first model, the second model, and the water injection algorithm is illustrated in Fig. 3(a). With each iteration, the user's movement triggers an update in network topology. The figure demonstrates that following reinforcement learning for power allocation updates, Q-learning achieves significantly higher spectral efficiency compared to the default water-filling algorithm. Furthermore, employing the exponential function model $y_2$ with reciprocal yields the highest spectral efficiency, proving to be more efficient and stable than both $y_1$ of Q-learning and the water-filling algorithm.

In the course of examining the secondary reward schema, we ventured to manipulate the hyper-parameters' valuations within the model's architecture, leading to nuanced variations in both the ascension pace and the potency of the reward. A meticulous inspection of Fig. 3(b). divulges that the zenith of spectral efficiency, accompanied by heightened stability, is attained when the $\beta$ value assumes the figure of 20. An escalation in the $\beta$ value corresponds with a proportionate amplification in performance, affirming a direct correlation.

**Table 1.** Simulation Parameters

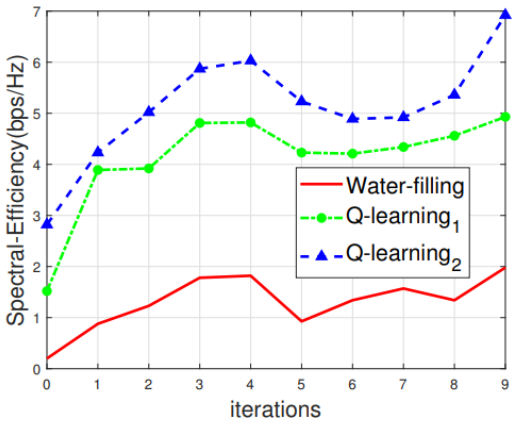
| Parameters | Values |
|---|---|
| Bandwidth B (MHz) | 10 |
| The number of BSs | 32 |
| The number of UEs | 200 |
| Radius r (m) | 15 |
| Initial power p (dbm) | 0-5 |
| Network Size m | 100×100 |
| pmax (dbm) | 30 |
| pmin (dbm) | -100 |
| h | 5 |
| α | 0.5 |
| γ | 0.9 |
| ϵ | 0.2 |

Nevertheless, the law of diminishing returns, as an influential factor, comes into play; with each successive increment in $\beta$, the incremental enhancement dwindles, gradually nearing a plateau of negligible change.

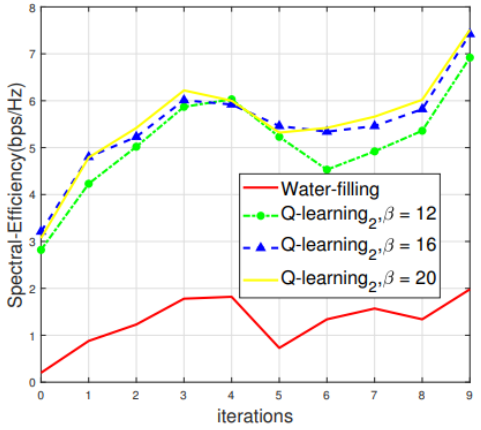
DOI: 10.54254/2753-8818/95/2024.21338

**Table 2.** β REVISE

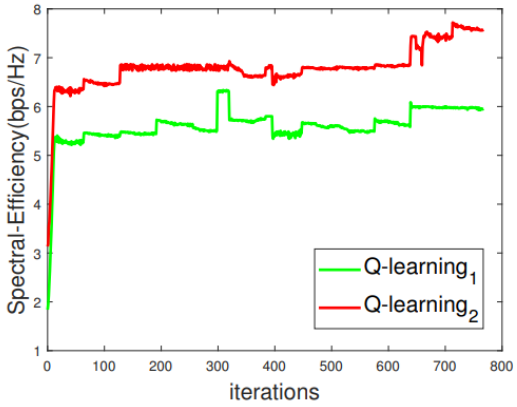
| Parameters | Values |
|---|---|
| Iteration step (0-30) | 0.8 |
| Iteration step (30-60) | 0.6 |
| Iteration step (60-100) | 0.3 |
| Iteration step (100-) | 0.1 |




(a) Comparison of spectral efficiency between different models.




(b) Comparison of spectral efficiency in the same model $y_2$.




(c) Comparison of convergence speed.

**Figure 3.** Consequence of Experiment.

Also, the convergence behavior of diverse model ensembles is depicted in Fig. 3(c), wherein Q-learning manifests notable oscillations, a consequence of the motility inherent in its interconnected nodes. Upon the occurrence of topological alterations, Q-learning necessitates recalibration and a renewed convergence process. Nonetheless, a conspicuous upsurge in initial recompense swiftly elevates the performance of the subsequent exponential paradigm, rendering it superior to its unadulterated exponential counterpart. Concurrently, the strategic constancy imbued by the reciprocal model significantly bolsters spectral efficacy and fosters an enhanced stability profile.

## 5. Conclusion

Tackling the complex issue of power allocation in multi-cell wireless systems, we've adopted a Q-learning reinforcement method, informed by machine learning insights. This integration,

DOI: 10.54254/2753-8818/95/2024.21338

alongside refined exponential models, has significantly improved both the speed and stability of our optimization strategy, enhancing network capacity. Our focus on maximizing capacity holds broader implications for ultra-dense networks, with potential applications spanning energy efficiency and user experience. Ultimately, this reinforcement learning approach paves the way for comprehensive performance improvements in wireless systems.

**Acknowledgment**

This project is largely supported by School of Aeronautics and Astronautics, University of Electronic Science and Technology of China.

**References**

[1] Yi Yang, Fenglei li, Xinzhe Zhang, Zhixin Liu, Dynamic power allocation in cellular network based on multi-agent double deep reinforcement learning, School of Electrical Engineering, China, Version of Record 11 September 2022.

[2] F. Meng, P. Chen, L. Wu and J. Cheng, Power Allocation in Multi-User Cellular Networks: Deep Reinforcement Learning Approaches, in IEEE Transactions on Wireless Communications.

[3] Y. Oh, A. Ullah and W. Choi, Multi-Objective Reinforcement Learning for Power Allocation in Massive MIMO Networks: A Solution to Spectral and Energy Trade-Offs, in IEEE Access.

[4] K. Antevski et al., A Q-learning strategy for federation of 5G services, ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICC40277.2020.9149082.

[5] Y. Zhang, C. Kang, T. Ma, Y. Teng and D. Guo, Power Allocation in Multi-Cell Networks Using Deep Reinforcement Learning, 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 2018, pp. 1-6, doi: 10.1109/VTCFall.2018.8690757.

[6] Sun, M., Hu, L., Wang, Y., Zhang, H., Wang, J. (2022). Deep Reinforcement Learning for Resource

[7] Allocation in Multi-cell Cellular Networks. In: Hassanien, A.E., Xu, Y., Zhao, Z., Mohammed, S., Fan, Z. (eds) Business Intelligence and Information Technology. BIIT 2021.

[8] H. V. Hoang Phuc and H. Hoang Kha, Massive MIMO Power Allocation Using Deep Reinforcement Learning, 2023 International Symposium on Electrical and Electronics Engineering (ISEE), Ho Chi Minh, Vietnam, 2023, pp. 97-102.

[9] F. B. Mismar, B. L. Evans and A. Alkhateeb, Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination, in IEEE Transactions on Communications.

[10] F. H. Costa Neto, D. C. Ara´ujo, M. P. Mota, T. F. Maciel and A. L. F. de Almeida, Uplink Power Control Framework Based on Reinforcement Learning for 5G Networks, in IEEE Transactions on Vehicular Technology.

[11] ZHOU Shuo, QIU Runhe, TANG Minjun, Power allocation algorithm for CR-NOMA system based on tabu search and Q-learning

[12] HU Langtao, BI Songjiao, LIU Quanjin, WU Jianlan, YANG Rui. Multi-Cell NOMA Energy Efficiency Optimization Power Allocation Algorithm Based on Reinforcement Learning[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(3): 384-391.

[13] D. Ma, Y. Wang, S. Wu and W. Wang, Using Reinforcement Learning for 5G RAN Slicing Resource

[14] Allocation in New Power Load Management System, 2023 9th International Conference on Computer and Communications (ICCC), Chengdu, China, 2023, pp. 1164-1169.