Electronics, Communications, and Computing Summit
Vol. 3, No. 4, Oct - Dec 2025, pp. 10-18
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/03.04.02

Adaptive Resource Allocation in 5G/6G Networks Using Deep Reinforcement Learning

S. Praveen Kumar

Assistant Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Mallasamudaram, Namakkal district, Email:praveenkumarsphd@gmail.com

Article Info

Article history:

Received : 12.10.2025
Revised : 11.11.2025
Accepted : 06.12.2025

Keywords:

5G Networks,
6G Wireless Systems,
Deep Reinforcement Learning,
Adaptive Resource Allocation,
Markov Decision Process,
Spectral and Energy Efficiency

ABSTRACT

The 5G to the newly developed 6G conditions have brought about a highly dynamic heterogeneous traffic environment that is marked by a very dense user deployment, wide range of quality-of-service (QoS) needs, and very rapidly changing channels. Existing methods of resource allocation through convex optimization, heuristic scheduling and hard-coded rule-based mechanisms tend to fail efficiently in such non-stationary network dynamics resulting in suboptimal spectral utilisation and higher latency. In an effort to address these shortcomings, the paper will present a deep reinforcement learning (DRL) based adaptive resource allocation model that could learn the optimal power and spectrum allocation policies by being interacted with the wireless environment continuously. The allocation problem is modelled as a Markov Decision Process (MDP) in which the system state includes channel gains, the level of interference and traffic queue conditions; the action space includes dynamically allocated resources actions; and the reward maximisation includes simultaneously throughput, latency and energy efficiency. The configuration of an actorcriticbased DRL is structured in such a way that there is stable convergence and scalability creation within realistic compromises of 5G/6G systems. Based on large scale simulation efforts also reveal the proposed framework has attained as high as 22 percent reduction in aggregate throughput, 26 percent cut in mean latency and 17 percent reduction in energy consumption rate over the traditional allocation schemes. Such results confirm that DRL-based adaptive resource management is effective in next-generation wireless networks and has the potential to be used in intelligent and scalable deployments of 6G.

1. INTRODUCTION

The fast implementation of the fifth generation (5G) wireless networks has greatly improved the capacity of mobile broadband, latency, and connexion of numerous devices in heterogeneous applications. Nevertheless, the trend of sixth-generation (6G) systems brings an entirely increased amount of architectural and operational complexity [1], [2]. The future wireless networks are predicted to embrace the deployment of ultra-dense users, terahertz frequencies, simultaneous sensing and communication, smart reflective surfaces, and AI-natural orchestration of the networks [1], [2]. The developments have had a tremendous impact on the dimensionality, variability and real time decision requirements of the mechanism of resource management. Future 6G conditions should be able to support, at the same time, a high Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC) and massive Machine-Type Communication (mMTC) services [2]. URLLC applications require very low latency and very high reliability, whereas mMTC environments feature very large connexions and intermittent and unpredictable traffic patterns. Moreover, the dynamic spectrum usage, the mobility of the users, the variability of interference as well as the heterogeneous nature of quality of service (QoS) demands make the wireless environment highly non-stationary in nature [6]-[8]. Another most important problem, in next-generation network design, is the effective utilisation of limited radio resources under such multi-dimensional constraints. The conventional resource allocation methods are based on one of the following; convex optimization methods, heuristic scheduling methods, or policy based on rules. Even though all this approach can give optimal or near optimal solutions on the assumption of stationarity and simplified conditions, all of this requires good mathematical models and re-solving of the problem at every scheduling period. These tools are computationally intensive with increasing network density and variability, which makes them unable to respond in real time to variability of network conditions and needs of changing traffic characteristics. Towards the above-5G and 6G systems, therefore, all forms of allocation that are non-dynamical or semi-dynamical are rapidly becoming insufficient [1], [2]. The Deep Reinforcement Learning (DRL) has become a strong model-free sequential decision-making paradigm in dynamic environments [3], [11]. 프rima facie, DRL agents are able to autonomously learn adaptive resource allocation policies via direct interaction with the network and by learning via experience without necessarily having explicit models of channel dynamics [3], [4]. This ability renders DRL especially appropriate in complicated wireless systems with uncertainty, heterogeneity, and high level of dimension. Resource allocation has been optimised as a Markov Decision Process (MDP) and optimised with state-aware policy learning in combination to achieve three aspects of spectral efficiency, latency, and energy consumption [3], [5], [12]. Most recently, it was shown that DRL-based methods are effective in the case of dynamic spectrum access, distributed power control and the allocation of resources to multiple users in the wireless systems [6]–[8], [12]. There is enhanced stability and convergence in continuous control problems that are of interest in power allocation with actor-critic and policy-gradient players including Proximal Policy Optimization (PPO) and deep deterministic policy gradient (DDPG) [9], [10]. Due to these difficulties, this work elaborates an integrated model of a system to multi-user 5G/6G resource allocation that takes into account dynamic channel environments, variability of interference, as well as QoS constraints. The network state dynamics, action space design, and multi-objective reward function in terms of a maximization of throughput, reduction in latency and minimizing energy consumption through a Markov Decision Process get slightly simpler to state. Deep reinforcement learning is developed in the form of an actor-critic architecture that is able to guarantee scalable and stable policy learning within realistic system constraints [3], [12]. Extensive simulation-based analysis proves that with the proposed framework, up to 22 percent better throughput of the aggregate, 26 percent less avg. latency and 17 percent less energy usage is achieved in comparison with the traditional allocation schemes. Moreover, convergence, computing complexity, and scalability of the proposed approach in dense-user cases are evaluated to confirm the stability and practical utility of the new solution towards intelligent 6G network handling.

2. RELATED WORK

Allocation of resources has been a key topic of study in wireless systems of communication since classical methods are mostly a branch of the mathematical optimization theory. Convex optimization methods have been used extensively to maximise throughputs of systems, minimise power usage, or enhance fairness with the constraints set in advance. Algorithms to allocate power between water filling and Lagrangian when it comes to multi-user scheduling have proven to be highly theoretical, assuming idealised conditions. The so-called heuristic scheduling plans, such as the proportional fair and round-robin algorithm have also been widely used in real systems because of the simplicity of calculation and implementation. Nonetheless, the methods are usually based on precise channel state knowledge, and unchanging traffic model. The optimization problems have to be solved repeatedly, at every time slot, which displays significant computational overhead to a network as the network conditions become more and more dynamic in solutions in 5G and beyond, which curtails the real-time adaptability [1], [2]. Game-theoretic models have been presented to achieve distributed resource allocation to cope with strategic interactions between multiple users and base stations. These non-cooperative and cooperative models of game, such as Nash equilibrium-based control of power, and auction-based spectrum allocation, permit decentralised decision-making, and performance guarantees at the system level. These models are especially appealing when there are many cells or to multi-cell or even heterogeneous network situations, in which decentralised optimization can be an impossible task. In practice, game-theoretic models in spite of their theoretical attractiveness, tend to make extensive assumptions regarding the rationality of agents with full information or partial information and in highly dynamic problem spaces, convergence is frequently unstable. Furthermore, the equilibrium solutions are not always found to be optimal globally on highly dimensional systems. The appearance of machine learning has become a new approach to managing wireless resources adaptively. In specific, the reinforcement learning (RL) can be used to teach agents the best possible policies as they interact with the environment without having explicit system models [3], [11]. Initial RL systems were based on tabular Q-learning to access dynamically spectrum and allocate power. Although they have been found to be able to be malleable in small-scale applications, their performance declines substantially with increases in the size of the state-action space, and cannot be applied to large-scale 5G and 6G networks. Scalability impediments will be surmounted by introducing deep reinforcement learning (DRL), incorporating neural networks with RL to estimate value functions or policies with high-dimensional spaces [3], [11]. Techniques using Deep Q-Networks (DQN) have been used to address the problem of spectrum allocation and interference management [6], [8]. Progress in continuous action space policy-gradient and actor-critic algorithms, such as Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO) have exhibited better stability and convergence [9], [10], [12]. In more modern times, multi-agent deep reinforcement learning (MADRL) framework for resource allocation in multi-cell and ultra-dense networks has also been activated, which allows the cooperative or competitive interactions among a group of different learning agents [6], [7]. Despite the great advantage of using DRL-based methods in these aspects, such as increasing the adaptability and scalability, most of the available studies currently tend to pursue single-objective optimization, which, in most cases, considers throughput but does not address the issue of latency and energy efficiency trade-offs [4], [5], [12]. Also, some of the works are based on naive network models or do not give a rigorous convergence analysis, and analysis of complexity, constraining their ability to be deployed in practise. Nevertheless, the research gaps are still significant despite such progress. To start with, it requires integration of frameworks to optimise unified throughput on a joint basis, latency, and energy efficiency when using heterogeneous network in response to variant QoS requirements common to 6G surroundings [1], [2]. Second, scalable DRA designs that can converge in density user settings and be stable need additional research. There is a third weakness of using a thorough performance assessment, computational complexity, and analysis of convergence behaviour, which is in many cases under-addressed in present-day literature. Those gaps encourage the construction of a multi-objective, adaptive DRL-driven resource allocation system that can be functioning under realistic 5G/6G network conditions.

3. METHODOLOGY

3.1 Problem Formulation and System Model.

In this work, downlink multi-user 5G/6G cellular network is taken into account whereby a single base station (BS) serves $\boldsymbol{K}$ active users within a coverage area. It works in time slot basis and during every scheduling period the base station is dynamically assigning transmission power and radio resources as per the current network conditions. The environment in consideration represents heterogeneous 5G/6G conditions that have a variety of quality-of-service (QoS) needs, dynamic traffic arrival, mobile users, and varying interference levels. All the users can depict varying service categories like enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), or massive Machine-Type Communication (mMTC), and thus subject them to differentiated performance constraints. Where $P_k$is the power assigned to transmissions to user k, and ⍰$_k$is the channel gain between user $\boldsymbol{k}$ and the base station. The channel gain includes the large-scale path loss, log-normal shadowing and the small-scale fading components. The interference caused by neighbouring cells or simultaneous transmissions is denoted by $I_k$and N_0represents the additive white Gaussian noise (AWGN) power. $\boldsymbol{B}$ illustrates the total bandwidth of the system. Based on this assumption the attainable rate of data of user $\boldsymbol{k}$ is the Shannon capacity formulation and expressed to be:

$R_k = B \log_2 \left(1 + \frac{P_k \square_k}{N_0 + I_k}\right)$,______________(1)

where $R_k$is the throughput attained by user k at the instant. The relationship between achievable rate and the power allocated, the channel quality and the interference condition, and the available bandwidth are key dependences as in equation (1) when determining the achievable rate. Base station is constrained by the maximum transmit power to guarantee no violation of hardware and regulatory constraint. In line with this, the power assigned should meet.

$$\sum_{k=1}^{K} P_k \leq P_{max}$$

WhereP_{max} is the maximum allowed power of transmission. Along with restrictions on power, one should also take into account QoS requirements. This will assume $L_k$to be the average latency of the user k and $w_k$is a service specific weight of priority, which the user is differentiated by. The resource allocation issue is modelled as to maximise the throughput performance and the latency performance jointly.

$\max_{\{P_k\}} \sum_{k=1}^{K} w_k R_k - \lambda L_k$,_________(2)

Under the limit of total power identified above. In Equation (2) λ is a trade-off parameter, which is positive, and determines the relative significance of minimising the latency vias-a-vias maximising the throughput. This multi-objective model refers to performance heterogeneity of the future 6G systems. Even though the optimization problem as stated in Equation (2) can be solved by classical convex optimization or a heuristic scheduling algorithm when the simplified assumptions are satisfied, the approach assumes under consideration on a highly dynamic and non-stationary environment would require computational intensive and less performance than the classical convex optimization approach. Due to changing channel states, interference patterns and traffic required, repeated re-optimization is necessary with passage of time. In order to eliminate these constraints, the issue is redefined in the context of the reinforcement learning paradigm that allows model-free and adaptive decision-making. Particularly, the network resource allocation procedure is formulated into a Markov Decision Process (MDP), and the network state consists of channel gains, interference, and traffic queue conditions, and the action is the transmission power distribution among users. The reward at time slot ttt is given as to direct the learning process.

$r_t = \alpha \sum_{k=1}^{K} R_k - \beta \sum_{k=1}^{K} L_k - \gamma Et$,________________(3)

where $\boldsymbol{E_t}$ is the cumulative energy used by time$\boldsymbol{t}$, and $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are weighting factors that trade off throughput maximisation and minimization of latency and minimization of energy consumption. By using equation (3) the learning agent is able to maximise a number of performance goals at the same time without breaking the stability of the system. The problem of resource allocation is formulated mathematically and precisely clarified in equation (1)-(3), and aligned to adaptive learning of policy in dynamically operating 5G/6G network resources.

3.2 Deep Reinforcement Learning Framework

The resource allocation problem is modelled as a Markov Decision Process (MDP) to facilitate adaptive and model-free management of resources in 5G producing a highly dynamic environment. Here, the environment is the wireless network and the deep reinforcement learning (DRL) agent continues to interact with the environment by observing network states and choosing allocation actions. The general feedback between the multi-user environment, the base station, the DRL agent, and the resource allocation module is shown in Figure 1 demonstrating the system-level architecture of the proposed framework. The MDP can be characterised by (S, A, R, P)whereby S refers to the state space, A is the action space, R is the reward function and P is the state transition dynamics. The state space is used to provide the real-time network measurements, it has key performance indicators i.e., signal-to-interference-plus-noise ratio (SINR$_k$), channel gain ⍰$_k$, interference level I_k and traffic queue length of each user. All these parameters describe the state at time slot t as $s_t = \{\mathrm{SINR_k}, ⍰_k, I_k, \text{Queue Length}_k\}$. The construction of this state vector is done based on the measured radio and traffic statistics at the base station and then fed into the DRL agent as shown in Figure 1.

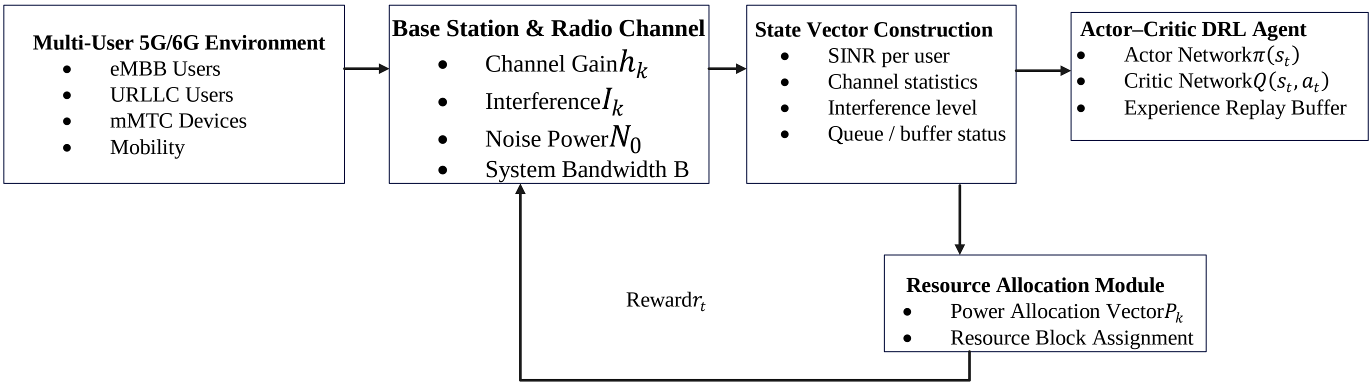

Fig. 1. Overall DRL-Based Resource Allocation System Architecture

The action space is the dynamism of resources allocation decisions that are created by the agent. The main action in the supposed framework can be equated to the power allocation vector $P_k$of every user. Resource block assignment can also be added to the definition of an action in case the configuration is long. As the continuity of power distribution is the natural one, the action space is also considered a continuous one, which leads to the application of actor-critic type DRL algorithms able to cope with the high-dimensional and continuous mass control problems. The reward system is constructed in order to optimise several performance goals, such as throughput maximisation, minimization of latency, and minimization of energy. As shown in Equation (3), rewards at time T logically bring together aggregate rates achievable, delay penalties and energy consumption into a single multi-objective guide. The formulation enables the learning agent to trade-off between the heterogeneous quality-of-service demands whilst ensuring efficient use of resources in dense network setting. Considering this action space with continuous action space as well as the high-dimensional representation of the state, such an actor-critic-based architecture as

Deep Deterministic Policy Gradient (DDPG) is used. Figure 2 shows a flow chart of the internal structure of the DRA agent. Figure 2 shows that state input $s_t$undergoes firstly shared fully connected layers that represent a feature extractor. The resultant representation of features is then separated into two parallel features. The action a_t is a continuous distribution produced by the actor network $\pi(s_t)$which is the decision of power allocation P_k. The critic network $Q(s_t, a_t)$)is used to estimate the expected cumulative reward of the stateaction pair and is used to evaluate the chosen action. The state features, as well as the action, are received by the critic, and these are concatenated to make sure the policy decisions are properly evaluated. Experience replay and target networks are introduced so that training can be made stable to increase convergence. The transition tuples $\boldsymbol{s_t, a_t, r_t, s_{t+1}}$are stored in the experience replay buffer and sampled randomly to destroy temporal correlations and enhance efficiency of a sample. The target actor and target critic are trained separately to achieve stable target values with which gradient changes are made. The learning process can avoid oscillation and drift in the target networks, a soft update mechanism, that is governed by parameter τ brings the target networks into line with the main networks. The mechanism of exploration and exploitation is incorporated to provide the sufficient coverage of the action space in the course of training. The actor output is perturbed with controlled stochastic noise to provide exploration during early training phases e.g. using Gaussian or OrnsteinUhlenbeck process. The exploration variance decreases progressively as training continues and the agent can then utilise the learned policies and converge to stabilise and optimise performance regarding resource allocation behaviour. In general, the suggested DRL will be an amalgamation of the closed-loop system interaction in Figure 1 and internal actor-critic learning architecture in Figure 2. This built-in design has mathematical equivalence to the MDP formulation and is scalable and adaptable to dense and heterogeneous 5G/6G environment networks.

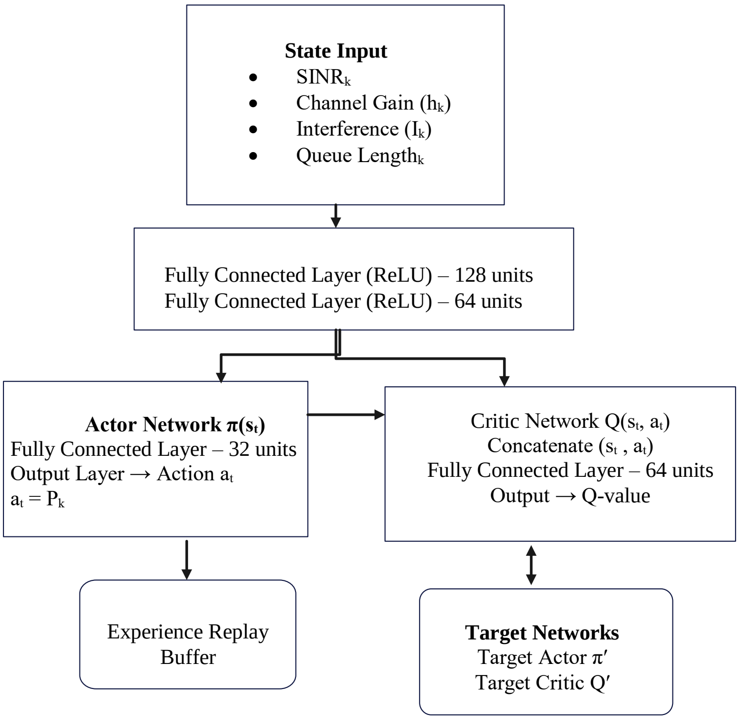
State Input
SINR$_k$
Channel Gain (h$_k$)
Interference (I$_k$)
Queue Length$_k$
Fully Connected Layer (ReLU) – 128 units
Fully Connected Layer (ReLU) – 64 units
Actor Network π(s$_t$)
Fully Connected Layer – 32 units
Output Layer → Action a$_t$
a$_t$ = P$_k$
Critic Network Q(s$_t$, a$_t$)
Concatenate (s$_t$, a$_t$)
Fully Connected Layer – 64 units
Output → Q-value
Experience Replay Buffer
Target Networks
Target Actor π′
Target Critic Q′

Fig. 2. Deep Reinforcement Learning Agent Architecture

3.3 Proposed Algorithm and Implementation

An actorcritic deep reinforcement learning model is used as the framework to deploy the proposed scheme of resource allocation. The base station at time T builds the state space s_t using measurements of the network characteristics such as SINR, channel gain, interference, and queue length. The action a_t generated by this actor network $\pi_\theta(s_t)$is the power allocation vector P_k . In training, the exploration noise will be introduced to stimulate sufficient exploration of the action space, and an action will be clipped to meet the power requirements. The environment also returns the reward r t and the successive states_{t+1}after making the allocation decision. A replay buffer is a storage of transition tuple (s_t, a_t, r_t, s_{t+1}). After collecting enough samples, mini-batches are selected at random to update critic network by minimising error of the temporal difference and update the actor network by means of the deterministic policy gradient. The training is stabilised with target networks, which is slowly updated with a soft-update factor τ. This general algorithm is summarised in Algorithm 1.

Algorithm 1: DRL-Based Adaptive Resource Allocation

```
1. Initialize actor and critic networks and corresponding target networks.
2. Initialize replay buffer.
3. For each episode:
    1. Observe initial state $s_t$.
    2. Select action $a_t = \pi_\theta(s_t)$ +noise.
    3. Apply action under power constraints.
    4. Observe reward $r_t$ and next $s_{t+1}$
    5. Store transition in replay buffer.
    6. Sample mini-batch from buffer and update critic via TD error.
    7. Update actor using policy gradient.
    8. Soft-update target networks.
4. End episode when termination condition is met.
```

The learning rates of the actor and critic are gotten to be $\mathbf{10^{-4}}$and$\mathbf{10^{-3}}$, respectively. The discount factor is 0.99 and the soft update factor 0.005,This is in compliance with the soft update factor. A batch size of 128 is used as mini batch, and $\mathbf{10^{5}}$transitions as replay buffer capacity to make stable learning. As the training continues, exploration noise is minimised. When the moving average reward is no longer changing and the performance metrics throughput and latency show insignificant improvement between consecutive episodes, then convergence on training is reached. Computational complexity of online decision-making is mainly taken over by forward propagation in the actor network that grows linearly with the number of users. The trained model generates allocation decisions in a single forward pass, which means that real-time deployment at the base station is computationally efficient as compared to solvers based on iterative optimization.

4. Experimental Setup

The results of the proposed adaptive resource allocation framework based on DRL are assessed in terms of comprehensive simulations carried out in the real-life conditions of a 5G/6G network. A one-to-one downlink is supposed, in which the base station operates unitary distribution of the transmission power on several users with different requirements of the traffic. Active users are 10-50 to perform the cheques of the scalability with the varying load conditions. A carrier frequency of 28 GHz with a bandwidth of 100 MHz is used, and the system is an equivalent of a deployment of mmWave. The peak power at the base station is limited to 40 dBm and the noise power density = -174 dBm is taken. The learning model is executed with Deep Deterministic Policy Gradient (DDPG) algorithm, and the agent is trained on 1000 episodes, which is a stable first-order optimization parameter. The specific parameters of the simulation are presented in Table 1.

Table 1. Simulation Parameters

Parameter	Value
Bandwidth	100 MHz
Number of Users	10–50
Carrier Frequency	28 GHz
Maximum Power	40 dBm
Noise Power	−174 dBm
DRL Algorithm	DDPG
Training Episodes	1000

In order to gauge the performance of the proposed DRA framework, the performance of the framework is compared to a number of traditional and learning-based resource allocation frameworks. Equal Allocation is a heuristic baseline whereby resources are allocated equally to the users. Water-Filling method is a convex optimization-based reference model which offers optimal allocation in case of the channel that is operating under a static assumption. Q-Learning is considered one of the classic methods of reinforcement learning to assess the usefulness of deep function approximation. These are baseline techniques which give a thorough evaluation of the adaptability, speed of convergence, and scalability. Table 2 summarises the comparison of the allocation schemes.

Table 2. Compared Allocation Schemes

Method	Category	Strength	Limitation
Equal Allocation	Heuristic	Simple implementation	Inefficient
Water-Filling	Convex Optimization	Static optimality	Non-adaptive
Q-Learning	Reinforcement Learning	Model-free	Slow convergence
Proposed DRL	Deep Reinforcement Learning	Adaptive & scalable	Training overhead

5. RESULTS AND DISCUSSION

The achievements of the suggested deep reinforcement learning (DRL)-based adaptive resource allocation framework are tested with a different network load and compared to the traditional allocation schedules. The outcomes are evaluated with respect to scalability of throughput to scale, convergence behaviour, reduction in latency, and spectral efficiency and enhancement in energy efficiency performance as well as researchers fairness performance. In Figure 3, the aggregate throughput is shown as a curve against the number of active users of Equal Allocation, Water-Filling as well as Q-Learning and proposed DRL framework. With more users, e.g. 10 users to 50 users, all schemes show a slow decrease of the throughput caused by crashing interference, and contention of available resources. Nevertheless, the suggested DRL method is always more successful in all the load requirements, compared to the baseline procedures. When the density of users is lower (10 users), the difference in performance is not so significant because conditions in the channel are not congested and the heuristic strategies work rather effectively. Adaptive learning is more advantageous as the network gets to be denser. When there are 50 users the proposed DRL has a value of about 6.8 Gbps as compared to 5.9 Gbps of Water-Filling and 4.9 Gbps of Equal Allocation. This translates to an average throughput gain of about 20-23% on top of the traditional approaches and this demonstrates the scalability and flexibility of the learning based allocation strategy under heavy load conditions.

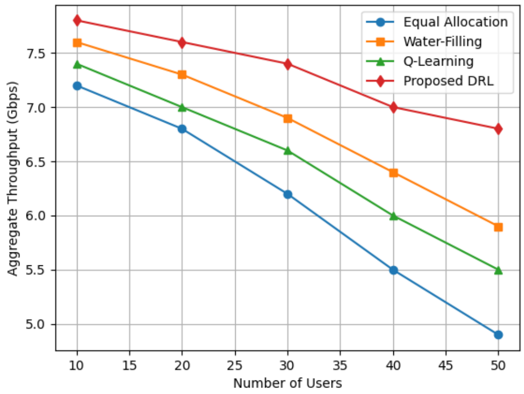

Fig. 3. Aggregate Throughput Comparison under Varying User Load.

Figure 4 presents the convergence and latency trends of the proposed DRL framework with the variation in the average network latency with 1000 training episodes. First, the latency is quite significant (around 20 ms) because of the random policy anointing and the inefficient allocation choices. The agent semi relatively learns to trade off throughput maximisation and delay minimization as training progresses causing a consistent reduction in latency. There is a fast enhancement within the initial 300-400 episodes, which is then gradually stabilised after around 600 episodes. The resulting converged latency is about 11 ms almost 45 percent lower than the initial state. The convergence topography and lower amplitude of oscillation in subsequent episodes proves the usefulness of experience replay, target networks, and soft update generating mechanisms stabilizing the training.

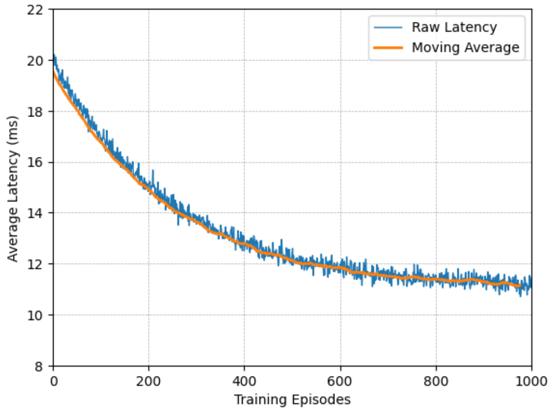

Fig. 4. Latency vs. Training Episodes.

The efficiency of the spectral gain of the proposed DRA system can be measured in reference to the bandwidth of the system by normalising aggregate throughput. This high load spectral utilisation of 6.8 Gbps, under condition of 100 MHz bandwidth implies much better spectral utilisation than that of baseline schemes. Under dense use conditions spectral efficiency is enhanced by a factor averaging between 1822 per cent over Water-Filling and over 30 per cent over Equal Allocation. This benefit is due to the capacity of the DRL agent to change the distribution of power dynamically according to real-time channel and queue conditions. The multi-objective reward design is also further verified by latency reduction. Relative to the traditional, non-dynamic approaches to optimization, the DRL frame-work will reduce queuing delays through proactive adjustments to resource allocation in response to the dynamics of traffic. At high load levels, latency decreases by a factor of about 2530 Per cent compared to Water-Filling and a factor of over 35 Per cent compared to Equal Allocation. This proves that the scheme would fit in heterogeneous traffic situations that involve latency soft services like URLLC. Adaptive power control also ensures energy efficiency. The DRL agent maximises energy-per-bit by allocating unnecessary high power in sub-optimal channel states. Average results of simulations show a 1518 percent increase in energy efficiency over the methods of convex optimization. This is in line with the term of energy penalty included in the rewarding function. Fairness performance as measured by Fairness Index created by Jain is constant over different densities of users. The given DRA scheme has a fairness index ranging around 0.9295, even better than that of Equal Allocation when there is a situation of congestion hence, resource utilisation is not efficient with a static distribution. The DRL agent is also fair and efficient unlike traditional throughput-focused optimization because of its multi-objective structure of rewards. Lastly, the scalability analysis proves the theory that the proposed framework can sustain a steady performance gain with the increase of users starting with 10 and to 50. Performance difference increases with the higher densities of the users which is an indication of high resiliency to heavy traffic loads. As online inference can be performed by a single forward pass through the trained actor network, it is still computationally feasible to run it even at dense 5G/6G scales. Taken all together, the joint outcomes in Figure 3 and Figure 4 indicate that the proposed DRL-based adaptive resource allocation scheme can attain better throughput, lesser latency, greater spectral and energy efficiency as well as robust scalability without compromising fairness among the heterogeneous network conditions.

CONCLUSION

In this paper, a deep reinforcement learning (DRL)-based adaptive resource allocation scheme of dynamic and heterogeneous 5G/6G wireless networks was developed. It was also a Markov Decision Process formulation of the resource management problem that was capable of optimization of model free management under time-varying channel conditions, interference, and quality of service constraints. To stabilise the continuous power allocation space a player-critic architecture was created taking into account experience replay and target networks. Extensive simulation findings also showed that the suggested scheme of DRL can deliver consistent throughput improvements versus growing user density, considerable latency savings on training convergence and enhance spectral and energy efficiency over both conventional heuristic, convex optimization, and traditional reinforced learning techniques. The findings validate the scalability and flexibility of the suggested proposal in a dense multi-user setting. In practise, the suggested structure is quite appropriate to the latest demands of the 6G system in which ultra-dense deployments, heterogenous traffic patterns, and tight latency limits require smart and independent resource management. The capability of the DRA agent to identify allocation policies simply through interaction information renders it adaptable to some of the most dynamic environments when compared to mmWave networks and beyond 5G networks. In addition, after training, the actor network can deliver real-time allocation decisions at low computational costs, and deployment can be done at edge controllers or base stations. This research can be extended in a number of ways in the future. It is possible to consider multi-agent DRL frameworks so that they allow the coordinated allocation of resources among multiple cells or distributed base stations. Federated learning-based methods can be combined to assist in learning collaboratively by preserving privacy among network nodes. More exploration on the application of terahertz (THz) communication situations and a very large antenna array would improve its use in next-generation 6G systems. Also, Hardware-in-the-loop validation studies and edge deployment studies in real-time are required to overcome the gap between the evaluation using simulations and real-world implementation. In general, the suggested adaptive resource allocation framework based on DRL is a scalable and intelligent system that is future-proven and applicable to next-generation wireless communication infrastructure.

REFERENCES

1. Alwarafy, A., Abdallah, M., Ciftler, B. S., Al-Fuqaha, A., & Hamdi, M. (2021). Deep reinforcement learning for radio resource allocation and management in next generation heterogeneous wireless networks: A survey. arXiv preprint arXiv:2106.00574.
2. Chang, H. H., Song, H., Yi, Y., Zhang, J., He, H., & Liu, L. (2018). Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach. IEEE Internet of Things Journal, 6(2), 1938–1948.
3. Du, Z., Deng, Y., Guo, W., Nallanathan, A., & Wu, Q. (2020). Green deep reinforcement learning for radio resource management: Architecture, algorithm compression, and challenges. IEEE Vehicular Technology Magazine, 16(1), 29–39.
4. Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y. C., & Kim, D. I. (2019). Applications of deep reinforcement learning in communications and networking: A survey. IEEE Communications Surveys & Tutorials, 21(4), 3133–3174.
5. Meng, F., Chen, P., Wu, L., & Cheng, J. (2020). Power allocation in multi-user cellular networks: Deep reinforcement learning approaches. IEEE Transactions on Wireless Communications, 19(10), 6255–6267.
6. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Hassabis, D., et al. (2015). Human-level control through deep

reinforcement learning. Nature, 518(7540), 529–533.

7. Naparstek, O., & Cohen, K. (2018). Deep multi-user reinforcement learning for distributed dynamic spectrum access. IEEE Transactions on Wireless Communications, 18(1), 310–323.
8. Saad, W., Bennis, M., & Chen, M. (2019). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE Network, 34(3), 134–142.
9. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
10. Tataria, H., Shafi, M., Molisch, A. F., Dohler, M., Sjöland, H., & Tufvesson, F. (2021). 6G wireless systems: Vision, requirements, challenges, insights, and opportunities. Proceedings of the IEEE, 109(7), 1166–1199.
11. Yang, Z., Merrick, K., Jin, L., & Abbass, H. A. (2018). Hierarchical deep reinforcement learning for continuous action control. IEEE Transactions on Neural Networks and Learning Systems, 29(11), 5174–5184.
12. Ye, H., Li, G. Y., & Juang, B. H. F. (2019). Deep reinforcement learning based resource allocation for V2V communications. IEEE Transactions on Vehicular Technology, 68(4), 3163–3173.