IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 January 2023, accepted 28 January 2023, date of publication 6 February 2023, date of current version 13 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243068

RESEARCH ARTICLE

A DRL-Based Automated Algorithm Selection Framework for Cross-Layer QoS-Aware Scheduling and Antenna Allocation in Massive MIMO Systems

CHIH-WEI HUANG[1], (Member, IEEE), IBRAHIM ALTHAMARY[1], YEN-CHENG CHOU[1], HONG-YUNN CHEN[2], AND CHENG-FU CHOU[3]

[1]Department of Communication Engineering, National Central University, Taoyuan 320317, Taiwan
[2]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan
[3]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan

Corresponding author: Chih-Wei Huang (cwhuang@ce.ncu.edu.tw)

This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 109-2221-E-008-054-MY3 and Grant 111-2218-E-008-004-MBK.

ABSTRACT Massive multiple-input-multiple-output (MIMO) systems support advanced applications with high data rates, low latency, and high reliability in next-generation mobile networks. However, using machine learning in massive MIMO resource allocation is challenging due to quality of service (QoS) and network complexity across layers. This work presents a novel framework for adapting the scheduling and antenna allocation in massive MIMO systems using deep reinforcement learning (DRL). Rather than directly assigning execution parameters, the proposed framework utilizes DRL to select combinations of algorithms based on the current traffic conditions. The DRL model is trained using a specialized Markov decision process (MDP) model with a componentized action structure and is evaluated in realistic traffic scenarios. The results show that the proposed framework increases satisfied users by 7.2% and 12.5% compared to static algorithm combinations and other cross-layer adaptation methods. This demonstrates the effectiveness of the framework in managing and optimizing resource allocation in a flexible and adaptable manner.

INDEX TERMS Mobile network, resource allocation, QoS, deep reinforcement learning, automated algorithm selection, massive MIMO.

I. INTRODUCTION

The rapid development of mobile networks proliferates the demands of high data rate, low latency, and high-reliability applications [1]. While the traditional mobile network confronts challenges on spectrum insufficiency, the multiple-input-multiple-output (MIMO) technology, which contributes to crucial progress in system capacity and reliability, is regarded as a necessary feature in the fifth-generation (5G) and beyond (B5G) wireless network systems [2]. However, when the quality of service (QoS) and upper-layer information are considered, the coordination problem becomes more complex under emerging MIMO technologies [3].

The massive MIMO system has achieved breakthroughs in practice by accessing a large number of antennas on active terminals [4]. It is characterized by base stations (BSs) equipped with a large number of antennas that simultaneously serve multiple users with shared time-frequency resources. In addition, the antennas steer energy in small regions to improve system throughput and energy efficiency. However, due to the scale of antenna and resource options, more advanced resource allocation is required for BSs in massive MIMO systems [5].

Cross-layer user scheduling and antenna allocation strategies in massive MIMO across media access control (MAC)

The associate editor coordinating the review of this manuscript and approving it for publication was Hosam El-Ocla.

and physical (PHY) layers have been actively investigated. Choi et al. [6] propose a joint user selection, power allocation, and precoder design algorithm for massive MIMO downlink systems providing gains in spectral efficiency. Kuo and Lu [7] propose a utility-based antenna allocation algorithm to efficiently allocate the number of antennas to user equipments (UEs) in a massive MIMO system. The work considered only scalable video streaming. Zhu et al. [8] propose a joint antenna and user scheduling solution with antenna/user deletion sub-problem solved by a low complexity rule-based algorithm named JAUS-LCC for massive MIMO scenarios. Though the sub-problem can be solved effectively, the overall QoS-aware algorithm is not as scalable. Also, when dealing with next-generation networks with significantly more control options, the rule-based algorithms require higher computational complexity to achieve high performance across layers due to their iterative nature and not accelerated by GPU-based platforms [9], [10]. We suggest developing a new framework capable of handling a more comprehensive range of factors across QoS requirements, scheduling, and antenna allocation.

With advances in artificial intelligence (AI), machine learning (ML) based techniques are adopted to deal with wireless network resource allocation problems. For example, Fiandrino et al. [11] lay out an ML-based framework containing a cross-layer orchestrator for 5G network optimization. The framework adopted deep learning models for traffic classification and forecasting. In [12], authors model massive MIMO user scheduling as a Markov decision process (MDP) and maximized the system sum-rate through a deep deterministic policy gradient (DDPG) based method, which is a landmark deep reinforcement learning (DRL) algorithm. In summary, current cross-layer resource allocation works appear not fully utilize machine learning for decision-making, while massive MIMO-focused works consider less on the application scenarios and QoS. To advance the state-of-the-art, i.e., [8], [12], we proposed to deal with the open problem of coordinating the interaction of cross-layer resource allocation algorithms in a MIMO system with QoS in mind.

While DRL is emerging as an essential resource management technique, practical reinforcement learning (RL) issues arise, such as challenging MDP properties, multi-objective reward design, and real-time feasibility [13]. Specifically, numerous control and condition parameters in a cross-layer massive MIMO system may result in high-dimensional MDP state and action spaces [3]. Also, diverse 5G QoS constraints [14] complicate the reward functions. Accordingly, RL-based *automated algorithm selection* is proposed to improve the performance and run-time feasibility when solving complex computational problems [15]. Studies have also shown that DL-based algorithm selection models that timely interact with environments have advantages in nonlinear and high complexity dynamic tasks [16]. With algorithm selection, the problem can be modeled as an MDP, and the instant action is a mixture of algorithms formed dynamically at run-time [17], [18]. Recently, the concept has been applied to 5G new radio (NR) resource allocation tasks to improve the training process but focused solely on user scheduling [19], [20]. We argue that the automated algorithm selection framework consisting of complementary algorithms can transform the cross-layer resource allocation problem to be more suitable for DRL-based solutions. For example, in B5G scenarios, when many UEs is under strict latency constraints, the systems can primarily benefit from providing higher priority to UEs with data expiring. When most UEs are traffic demanding, proportional fairness can be the preferred criterion. Therefore, joint resource allocation actions adapting a combination of feasible fundamental algorithms are worth investigating. It is an essential step toward a smart BS in an AI-managed open radio access network (O-RAN) [21] or general AI RAN [22].

This work proposes a DRL-based cross-layer user scheduling and antenna allocation framework for massive MIMO systems running 5G applications. Instead of directly assigning resource allocation parameters, such as selected users, the number of antennas, and precoding matrices, the task is transformed into an automated algorithm selection process. Combinations of algorithms in a flow of function components are dynamically determined for scheduling and antenna allocation. An MDP model is designed to meet the QoS requirements, including latency, data rate, and packet loss rate, and fit DDPG training processes. As a result, a novel adaptation framework incorporating automated algorithm selection, DRL, and cross-layer resource allocation is presented for QoS-aware resource management across a broad range of network layers. The main contributions can be summarized as follows:

- We formulate a QoS-aware radio resource allocation problem for joint scheduling and massive MIMO antenna allocation. The utility function integrates user requirements toward a long-term system-wide objective that matches the MDP return.
- A componentized MDP action structure for automated algorithm selection is proposed. The resource allocation functions and fundamental algorithms for user scheduling and antenna allocation are identified, so the dynamic algorithm selection policy can be effectively trained and executed.
- A training process based on DDPG [23] is developed for the problem with continuous or high dimensional states and actions. Action embedding [24] is also incorporated to convert the algorithm selection actions into a continuous space and take full advantage of DDPG.

The simulations are performed under realistic traffic scenarios referring to traffic types in the 5G QoS identifier (5QI) table [14]. Static algorithm combinations and baselines in the literature are implemented for comparison. Simulation results suggest that the proposed dynamic algorithm selection satisfied 7.2% and 12.5% more users against static algorithm selection and related joint allocation schemes under demanding scenarios.

In the rest of the paper, we present related works on resource management and machine learning in Section II. Then, Section III introduces the massive MIMO system model and problem formulation of cross-layer scheduling and antenna allocation. Section IV describes the proposed MDP model with componentized actions. The simulation results are demonstrated in Section V. Finally, Section VI concludes the article.

II. RELATED WORKS

A. CROSS-LAYER USER SCHEDULING AND ANTENNA ALLOCATION

User scheduling has been one of the primary resource allocation topics across generations of mobile communication technologies. With massive MIMO, the antenna allocation further controls the availability of underlying physical resources and can be jointly considered to enhance performance. Reference [25] presents an adaptive algorithm for joint user scheduling, precoding design, and beamforming in dynamic MIMO small cell networks. The transmit direction is optimized for every frame using conventional allocation strategies across scheduling, precoding, and power control. Sheikh et al. [26] propose user and antenna selection algorithms to maximize the system sum-rate of a massive MIMO system with various precoding schemes. Lagen et al. [27] present a procedure for joint user scheduling, precoder design, and transmission directing in MIMO small cell networks. Authors in [28] propose an improved user scheduling approach with low-rank channels and precoding design based on a two-stage precoding framework for large-scale MIMO systems with frequency division duplexing. As a result, both throughput gain and fairness were achieved. In [29] and [30], joint scheduling and precoding for matching MIMO cellular networks are investigated and analyzed. In [31], the authors propose an antenna and user selection algorithm for downlink massive MIMO transmission using zero-forcing (ZF) precoding. Singh et al. [32] develop an optimal resource fraction algorithm (ORFA) combining the proportional fair UE ranking and water filling resource allocation for MIMO networks with a minimum mean square error (MMSE) precoder. In [7], a utility-based layer and antenna allocation (UBLAA) algorithm is proposed to maximize the transmission efficiency for layered video streaming. The marginal utility is evaluated to determine the number of antennas assigned to UEs in a massive MIMO system.

While cross-layer user scheduling and precoding can be executed to some extent with conventional methods, the challenging application QoS requirements and rising complexity of B5G systems lead to performance degradation [9]. Therefore, machine learning-based approaches are worth studying to integrate cross-layer functions for future networks [11].

Machine learning-based techniques have been actively developed for next-generation network resource management. Wei et al. [33] propose an actor-critic-based model for user scheduling and resource allocation to efficiently utilize radio resources in HetNets. The training was performed without deep neural networks. Authors in [34] address the benefits of artificial intelligence-aided wireless systems and categorized primary machine learning algorithms in next-generation networks. Applicable wireless communication technologies include massive MIMO, cognitive radios, heterogeneous networks, small cells, and device-to-device networks. Reference [35] builds a resource management model with DRL for network slicing and demonstrates that DRL could incorporate demand and supply, enhancing network slicing agility. Reference [36] propose a DRL-enabled coverage and capacity optimization algorithm for massive MIMO systems considering perfect channel estimation. DRL is used to coordinate the inter-cell interference and support user scheduling dynamically. Reference [37] applies a DRL model for resource allocation in vehicle-to-vehicle (V2V) communications. The agents determine the sub-band and power levels for transmission with local observations. Zhang et al. [38] propose DRL-based control for resource management in spectrum sharing. With dynamic power control, primary and secondary users efficiently meet QoS requirements. Chien et al. [39] propose the PowerNet using large-scale fading information to predict the pilot and data powers with a varying number of active users. Furthermore, it is independent of small-scale fading and allows long-term throughput/spectral efficiency (SE) optimization.

Overall, DRL has been applied to various resource management tasks in wireless networks. However, cross-layer coordination is more complex, less studied, and requires specifically designed ML structures to be effectively solved [40].

B. DEEP REINFORCEMENT LEARNING AND DDPG

RL is a machine learning technique to solve decision-making problems typically modeled as an MDP, a mathematical framework to describe the target environment [41]. In RL, an agent learns through interacting with the environment and iteratively improves its ability to achieve a pre-defined goal. An MDP problem consists of states $\mathbf{s}_t \in \mathcal{S}$, actions $\mathbf{a}_t \in \mathcal{A}$, transition probabilities $Pr(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, and rewards $r_t = r(\mathbf{s}_t, \mathbf{a}_t) \in \mathbb{R}$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces. At time step t, an agent recognizes $\mathbf{s}_t$ from the environment and chooses a suitable $\mathbf{a}_t$. After the action is applied, the next state $\mathbf{s}_{t+1}$ and reward r_t are observed from the environment. This model aims to learn the stochastic policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$, which maximizes the long-term return

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(\mathbf{s}_i, \mathbf{a}_i), \tag{1}$$

where T and $\gamma \in (0, 1)$ are termination time and the discount factor. The action-value represents the expected return when executing action $\mathbf{a}_t$ in state $\mathbf{s}_t$ following π as

$$Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\mathbf{a}\sim\pi}[R_t|\mathbf{s}_t, \mathbf{a}_t], \tag{2}$$

where $\mathbb{E}$ is the expectation operator.

When modeled by MDP, the massive MIMO resource allocation is a high complexity problem with large state and action spaces to present possible situations. Therefore, a DRL-based approach, specifically DDPG [23], is proposed to integrate with the resource allocation process. DDPG is a landmark scheme in the policy gradient family and is more suitable for applications with complex actions than the deep Q-network (DQN) [42]. The deterministic policy gradient (DPG) [43] concept, experience replay, slow-learning target networks from DQN, and the actor-critic structure are integrated into DDPG.

The DDPG algorithm utilizes the recursive Bellman equation to evaluate action-value functions differently from DQN. Thus the deterministic policy $\mu : \mathcal{S} \rightarrow \mathcal{A}$ provides the action $\mathbf{a}_t = \mu(\mathbf{s}_t)$, and the action-value function becomes

$$Q(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{r_t, \mathbf{s}_{t+1} \sim E}[r(\mathbf{s}_t, \mathbf{a}_t) + \gamma Q(\mathbf{s}_{t+1}, \mu(\mathbf{s}_{t+1}))]. \quad (3)$$

Furthermore, we can optimize the action-value function by training deep neural networks considering function approximators parameterized by θ^Q and θ^μ. The actor network updates the policy with aids from the critic network, where the policy gradient is [43]

$$\begin{aligned} \nabla_{\theta^\mu} J &\approx \mathbb{E}_{r_t, \mathbf{s}_{t+1} \sim E}[\nabla_{\theta^\mu} Q(\mathbf{s}_t, \mathbf{a}_t)] \\ &= \mathbb{E}_{\mathbf{s}_t \sim E}[\nabla_{\mathbf{a}} Q(\mathbf{s}_t, \mathbf{a}_t | \theta^Q) \nabla_{\theta_\mu} \mu(\mathbf{s}_t | \theta^\mu)]. \end{aligned} \quad (4)$$

Accordingly, the training procedures using samples from experience replay, E, can be realized.

III. SYSTEM MODEL AND PROBLEM FORMULATION

This section describes the massive MIMO network structure and problem formulation. A cross-layer user scheduling and antenna allocation problem is proposed to be modeled as an MDP.

A. SYSTEM MODEL

We consider a single-cell massive MIMO system consisting of an M-antenna BS, K single-antenna UEs, and cross-layer controls [8]. Thus we have $m \in \mathbf{M} = \{1, 2, \ldots, M\}$ and $k \in \mathbf{K} = \{1, 2, \ldots, K\}$. The BS assigns $N_{k,t}$ number of antennas to UE k at the t-th transmission time interval (TTI), where TTI is T_I-second long. Each UE is associated with a traffic type, referring to the 5QI table [14]. UE properties include channel quality indicators, requested data, and a traffic type. The channel quality indicator, **CQI**, can be obtained from the table defined in [44], given the signal to interference and noise ratio (SINR). The requested data packets, $\mathbf{D}_k$, is the set of packet identification numbers (IDs) to transmit for UE k. Finally, the properties attached with a traffic type, **TYPE**, are packet size, mean packet arrival time, latency constraint, guarantee bit-rate, and error rate constraint.

The cross-layer user scheduling and antenna allocation framework is illustrated in Figure 1, with three *function components* considered for automated algorithm selection. The components are interdependent resource allocation functions to be jointly adapted across layers; several complementary

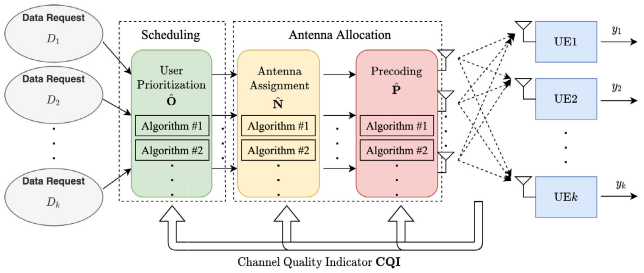


FIGURE 1. The cross-layer user scheduling and antenna allocation framework with automated algorithm selection. User prioritization, antenna assignment, and precoding function components are adapted across layers; each includes several algorithms as options.

algorithms should be included in each component. Thus, the framework is extensible with more function component as long as it takes advantage from automated algorithm selection. In addition, to be practical, other functions in the transmission system run algorithms typically, if not included in the cross-layer adaptation. In this work, user prioritization is included in the framework as the core of the *scheduling* function [32]. The antenna assignment and precoding, part of antenna resource allocation and dependent on other function components [45], are included and called *antenna allocation* later in the article.

The outcome of user prioritization is defined as $\hat{\mathbf{O}} = \{\mathbf{O}_t | 1 \leq t \leq T\}$, which ranks UEs every TTI. $\mathbf{O}_t$ is an *ordered* subset of UEs containing requested data to be transmitted at t. The antenna assignment results, $\hat{\mathbf{N}} = \{\mathbf{N}_t | 1 \leq t \leq T\}$, record the number of antennas assigned to prioritized UEs, where $\mathbf{N}_t = \{N_{k,t} | k \in \mathbf{O}_t\}$. Also, the precoding matrix set, $\hat{\mathbf{P}} = \{\mathrm{P}_t | 1 \leq t \leq T\}$, is the set of precoding matrices evaluated given antenna assignment $\mathbf{N}_t$. Given user data requests and traffic types, the algorithm selected in each function component is determined according to channel quality indicator **CQI**. Finally, a UE decodes the received signal y_k for the data.

B. MASSIVE MIMO TRANSMISSION MODEL

The precoding influences the spectral efficiency by evaluating the precoding matrix and providing raw capacity for resource allocation. In massive MIMO communications, the conventional realization of linear precoder requires a complex, high-resolution phase shifter network. Furthermore, each radio frequency (RF) chain is coupled to all antennas resulting in an expensive and energy-demanding system. Therefore, hybrid precoding/beamforming architecture [46] is utilized in this work to mitigate hardware constraints while realizing linear precoding. Assuming there are K_r UEs simultaneously receiving data at a TTI, the received signal vector $\mathbf{y} = [y_1, y_2, \ldots, y_{K_r}]^T \in \mathbb{C}^{K_r \times 1}$ can be expressed as

$$\mathbf{y} = \mathrm{H}\mathbf{x} + \mathbf{n}. \quad (5)$$

$\mathrm{H} \in \mathbb{C}^{K_r \times M}$ denotes the channel matrix of all K_r users; $h_k \in \mathbb{C}^{1 \times M}$ is user k's channel vector. $\mathbf{n} = [n_1, n_2, \ldots, n_{K_r}]^T$ is the additive white Gaussian noise (AWGN) with $n_k \sim \mathcal{CN}(0, \sigma^2)$, where $\mathcal{CN}$ denotes a complex Gaussian distribution and σ is the noise standard deviation. Processed by

TABLE 1. Summary of notations.

Notation	Description
t	Index of the t-th TTI
$\mathbf{M}/M$	BS antenna set / total number of antennas
$\mathbf{K}/K/K_r$	UE set / total number of UEs / simultaneously receiving UEs
$\hat{\mathbf{O}}/\mathbf{O}_t$	Ordered UE set / the ordered UE set at t
$\hat{\mathbf{N}}/\mathbf{N}_t/N_{k,t}$	Antenna assignment set / the antenna assignment set at t / number of antennas assigned to UE k at t
$\hat{\mathbf{P}}/\mathrm{P}_t/\mathbf{p}_{k,t}$	Precoding matrix set / the precoding matrix at t / k-th column vector of P_t
$\mathrm{H}/\mathbf{h}_{k,t}$	Channel matrix / channel vector of UE k at t
$\Phi_{k,t}$	Data transmission rate of UE k at t
$U_k/U_{k,t}$	Overall utility of UE k / utility of UE k up to t
$\nu_{k,t}$	Number of successfully received packets by UE k at t
E_k	Packet error rate requirement of UE k
$\mathbf{D}_k$	Set of requested data packet IDs of UE k
$\mathbf{t}(d)$	TTIs assigned to transmit data packet d in time
$u^d_{k,t}$	Receiving status of data packet d of UE k up to t
ε_d	Packet size of the data d
W	System bandwidth
ρ	BS transmission power
y_k	Received signal of UE k
$\mathbf{x}$	Transmitted signal for UE
$\mathbf{n}$	The additive white Gaussian noise
T_I	The duration of a TTI
T	Termination time in number of TTIs
B	Number of transitions in a mini-batch

a power amplifier, the transmitted signal vector $x \in \mathbb{C}^{M\times 1}$ is transmitted through the antennas. It is given as

$$\mathbf{x} = \mathrm{P}\chi. \tag{6}$$

$\mathrm{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{K_r}] \in \mathbb{C}^{M\times K_r}$ with column vectors $p_k \in \mathbb{C}^{M\times 1}$ is the hybrid precoding matrix. $\chi = [\chi_1, \chi_2, \ldots, \chi_{K_r}]^T \in \mathbb{C}^{K_r\times 1}$ is the modulated user signals with $\mathbb{E}[\chi\chi^H] = (\frac{\rho}{K_r})\mathrm{I}_{K_r}$, where I and ρ refer to the unit matrix and the total transmission power. With the system model defined in Figure 1, the received signal of user k is

$$y_k = \mathbf{h}_k \sum_{i=1}^{K_r} \mathbf{p}_i\chi_i + n_k. \tag{7}$$

Therefore, after antenna assignment at the time index t, we have $K_r = \|\mathbf{O}_t\|$. The precoding matrix $\mathrm{P}_t = f_p(\mathbf{N}_t) = [\mathbf{p}_{1,t}, \mathbf{p}_{2,t}, \ldots, \mathbf{p}_{K_r,t}] \in \mathbb{C}^{M\times\|\mathbf{O}_t\|}$ is evaluated by precoding algorithm f_p given antenna assignment $\mathbf{N}_t$. The signal-to-interference-plus-noise ratio (SINR) of UE k at t is [47]

$$SINR_{k,t} = \frac{\frac{\rho}{\|\mathbf{O}_t\|}|\mathbf{h}_{k,t}\mathbf{p}_{k,t}|^2}{\sigma^2 + \frac{\rho}{\|\mathbf{O}_t\|}\sum_{j\in\mathbf{O}_t, j\neq k}|\mathbf{h}_{k,t}\mathbf{p}_{j,t}|^2}. \tag{8}$$

Hence, the data transmission rate Φ of UE k at time t is

$$\Phi_{k,t} = \frac{W}{\|\mathbf{O}_t\|}\cdot\log\left(1 + SINR_{k,t}\right), \tag{9}$$

where W is the system bandwidth. Instead of ergodic expressions, the adopted instantaneous SINR is more suitable for the utility-based problem introduced later. We also assume the SINR is estimated accurately to concentrate on the cross-layer resource allocation in this work. Nevertheless, the proposed framework can be augmented with advanced channel estimation [48] for non-ideal situations.

C. QoS-AWARE USER SCHEDULING AND ANTENNA ALLOCATION

User resource allocation maximizes the total system utility by actively distributing resources. Packet-level transmission utility is first defined to describe the QoS status in requested data. Assuming set $\mathbf{t}(d)$ is the TTIs assigned to transmit data packet d within its latency constraint. $u^d_{k,t}$ indicates the receiving status of data packet $d \in \mathbf{D}_k$ and is defined as

$$u^d_{k,t} = \begin{cases} 1, & \text{if } \sum_{i\in\mathbf{t}(d)}\Phi_{k,i}\cdot T_I \geq \varepsilon_d \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

A packet is successfully received, i.e., $u^d_{k,t} = 1$, if sufficient resources are allocated to a packet in time. ε_d is the packet size. Consequently, the number of successfully received packets by UE k up to time t is

$$\nu_{k,t} = \sum_{d\in\mathbf{D}_k} u^d_{k,t}. \tag{11}$$

Based on the receiving status defined above, the user resource allocation problem maximizes the total utility every TTI. The utility gain of UE k up to the t-th TTI, $U_{k,t}$, is a function of data received over time. Simultaneously, application requirements are satisfied, including guarantee bit rate (GBR), packet loss rate, and latency. The GBR constraint can therefore be derived as

$$\frac{1}{t}\sum_{i=1}^{t}\Phi_{k,i} \geq GBR_k, \quad \forall k \in \mathbf{K}. \tag{12}$$

Also, the packet error rate constraint is

$$1 - \frac{\nu_{k,t}}{\|\mathbf{D}_k\|} \leq E_k, \quad \forall k \in \mathbf{K}, \tag{13}$$

where E_k is the packet error rate requirement from UE k's traffic type.

It is challenging to adapt all options from scheduling to antenna allocation effectively. Given the antenna assignment, the precoding matrix determines the resulting throughput, while the antenna assignment is based on user prioritization. We model the complex interaction with a utility function integrating requirements and dependencies toward a long-term system-wide objective. The problems are jointly processed under the componentized structure and automated algorithm selection.

The QoS-aware cross-layer resource allocation objective maximizes the number of satisfied users in the system given their application requirements. Therefore, we propose a utility function with integrated requirements until the termination time T. The received utility of UE k, U_k, is set to 1 when

GBR, loss, and latency requirements are satisfied given allocated antenna resources. The utility function follows the conditional concept for QoS [49] can be expressed as

$$U_k \equiv U_{k,T} = \begin{cases} 1, & \text{if } \frac{1}{T}\sum_{i=1}^{T} \Phi_{k,i} \geq GBR_k \\ & \wedge 1 - \frac{v_{k,T}}{\|\mathbf{D}_k\|} \leq E_k \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The joint problem is formulated by QoS requirements embedded utility and resource constraints as

$$\max_{\hat{\mathbf{O}},\hat{\mathbf{N}},\hat{\mathbf{P}}} \sum_{k\in\mathbf{K}} U_k, \tag{15}$$

$$\text{subject to} \sum_{k\in\mathbf{O}_t} N_{k,t} \leq M, \quad 0 \leq t \leq T \tag{16a}$$

$$\sum_{k\in\mathbf{O}_t} |\mathbf{p}_{k,t}|^2 \leq 1, \quad \forall m \in \mathbf{M}. \tag{16b}$$

The objective is to maximize the number of satisfied UEs by determining the optimal $\hat{\mathbf{O}}, \hat{\mathbf{N}}, \hat{\mathbf{P}}$ over time. $\hat{\mathbf{O}}, \hat{\mathbf{N}}, \hat{\mathbf{P}}$ are the outcomes of scheduling and antenna allocation algorithms that influence the data transmission rate $\Phi_{k,t}$ and, thus, the utility function. (16a) limits the total number of the allocated antennas. (16b) is the constraint of the precoding matrix gain. At the same time, the problem features a utility function depending on complex criteria and long-term returns. Therefore, an MDP-based solution, which models complex agent-environment interaction and optimizes future return during the process, adequately fits the problem.

IV. DEEP REINFORCEMENT LEARNING FOR AUTOMATED ALGORITHM SELECTION

This section presents the MDP model with states, actions, and rewards. Also, the resource allocation function components and the DDPG training procedures are detailed.

A. MARKOV DECISION PROCESS FORMULATION

Figure 2 illustrates the massive MIMO resource allocation problem in the DDPG structure. The control agent in the BS collects *state* information to determine resource allocation actions during the RL process. In addition, the state aims to assess system *statistics* regarding UEs channel quality levels, the amount of data to transmit, and QoS requirements. The channel quality level set, $\hat{\mathbf{CQI}}$, records the distribution of UE channel quality. The number of elements in $\hat{\mathbf{CQI}}$ equals the number of modulation and coding scheme (MCS) levels mapped from CQI by specifications in [44]. The values of each element are the number of UEs at the MCS level. The amount of data to transmit, $\hat{\mathbf{D}}$, is the requested data remaining in the queue. The set of traffic types, $\hat{\mathbf{TYPE}}$, presents the distribution of UE QoS requirements. The number of elements in $\hat{\mathbf{TYPE}}$ equals the number of considered data types, and the values of each element are the number of UEs that belong to the type. Therefore, the model can deal with varying numbers

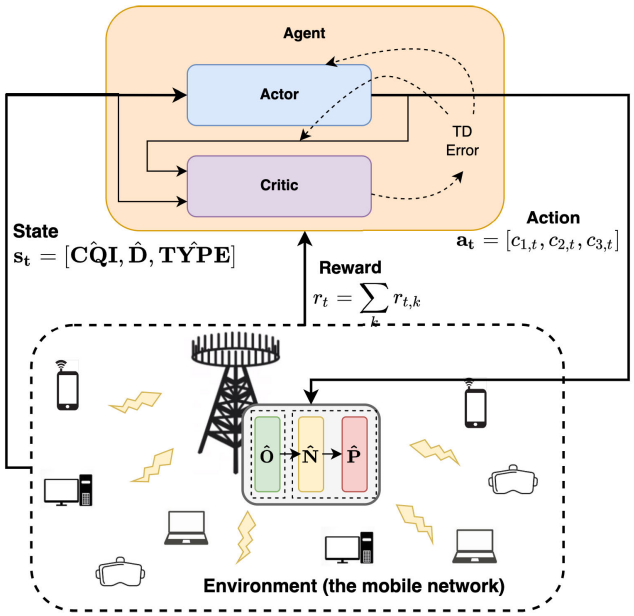


FIGURE 2. DDPG structure for massive MIMO resource allocation. The control agent in the smart BS collects state and reward information to determine resource allocation actions during the process.

of UEs without retraining. The state at the t-th TTI is defined as

$$\mathbf{s}_t = [\hat{\mathbf{CQI}}, \hat{\mathbf{D}}, \hat{\mathbf{TYPE}}]. \tag{17}$$

Based on the problem formulated in Section III-C, the resource allocation *action* is formed as a combination of user prioritization, antenna assignment, and precoder algorithms. In addition, fundamental schemes proven helpful in specific scenarios are included in a function component. The action dynamically selects an algorithm in each component according to the state observed every TTI and is expressed as

$$\mathbf{a}_t = [c_{1,t}, c_{2,t}, c_{3,t}]. \tag{18}$$

The details of included components are described later in Section IV-B.

The reward keeps the data transmission on pace, considering traffic type-specific GRB and latency requirements. However, due to higher uncertainty in quantifying the advantage of proactive transmission, we adopt negative rewards to discourage situations with transmission progress falling behind [50]. The reward of UE k is formulated as

$$r_{k,t} = \left(\frac{v_{k,t}}{\|\mathbf{D}_k\|} - 1\right)\left(1 + \alpha \cdot \left(1 - \frac{\sum_{i=1}^{t} \Phi_{k,i}}{t \cdot GBR_k}\right)\right). \tag{19}$$

The first term reflects the incompletion ratio of requested data up to time t. As a penalty, the value is negative if requested data from UE k are not fully transmitted. On the other hand, if all request data are transmitted, the first term and thus the reward becomes zero. The second term is the adjustment to keep the transmission data rate on GBR_k. α is the penalty weight, and $\alpha = 0$ when the traffic type has no GBR assigned.

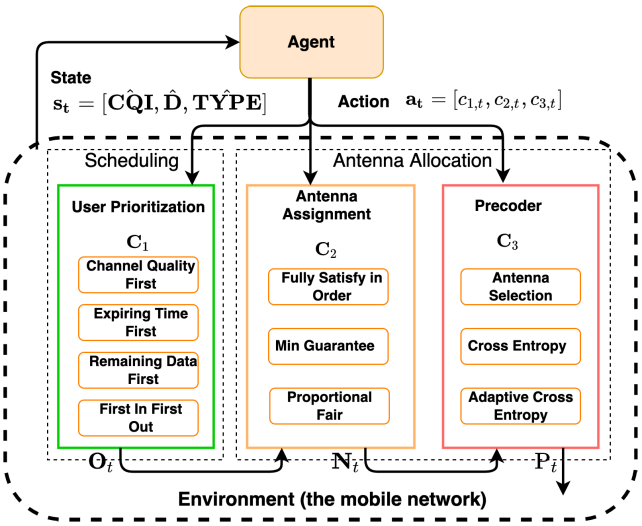


FIGURE 3. Componentized actions and fundamental algorithms in each component. After the algorithms are selected, execution results, $\mathbf{O}_t$, $\mathbf{N}_t$, and $\mathbf{P}_t$, are sent to the following components or the transmitter.

Therefore, the reward function is

$$r_t = \sum_{k \in \mathbf{K}} r_{k,t}. \tag{20}$$

Therefore, the reward function is R. With the reward function design, the learning process fits the utility optimization problem (15) and realizes future return optimization in MDP.

B. COMPONENTIZED ACTIONS

The componentized action is the concept introduced to facilitate dynamic resource allocation via algorithm selection and improve DDPG training. As shown in Figure 3, we decompose the scheduling and precoding process into three function components; a component contains several fundamental algorithms as options. After an action of selected algorithms is generated, the corresponding algorithms in each component are then executed. Execution results, $\mathbf{O}_t$ and $\mathbf{N}_t$, are sent to the following components, while the transmitter takes $\mathbf{P}_t$ as inputs. The included algorithms are diverse in design concepts for meaningful selection. That is, Component 1 ($\mathbf{C}_1$) prioritizes UE according to specific criteria. Component 2 ($\mathbf{C}_2$) decides the number of antennas assigned to each UE per TTI. Finally, the precoding method is determined in Component 3 ($\mathbf{C}_3$). The adopted algorithms are introduced as follows.

The UE prioritization component ranks UEs in the system. Four implemented sorting methods are:

- Channel quality first (CQI): sorts UEs according to channel conditions. A UE with higher channel quality is ranked higher.
- Expiring time first (Delay): ranks UEs on how close its oldest requested data is expired. It depends on the traffic-specific time constraints and how long the transmission is delayed.
- Remaining data first (Remain): sorts UEs according to the size of requested data remaining in the queue, i.e., $\|\mathbf{D}_{k,t}\| - \nu_{k,t-1}$. A UE receives higher priority with more untransmitted data.
- First-in-first-out (FIFO): prioritizes UEs with the arrival time of the earliest arrival packet.

Thus, component $c_{1,t} \in \mathbf{C}_1 = \{\text{CQI, Delay, Remain, FIFO}\}$. An ordered UE set $\mathbf{O}_t$ is generated every TTI.

The second component is to assign the system resources, i.e., the number of antennas $N_{k,t}$, to UEs based on the ordered set $\mathbf{O}_t$. As a result, c_2 also controls the final number of UE, which can be granted a transmission opportunity. In addition to algorithms, a percentage parameter ι is integrated to extend the options. The fundamental assignment methods implemented are:

- Fully satisfy in order (FSO): assign sufficient antennas to fully transmit the remaining requested data of each UE, $\|\mathbf{D}_k\| - \nu_{k,t-1}$, in the order of $\mathbf{O}_t$ until exhausting the system resource. The number of antennas to fully satisfy a UE, $N^{fs}_{k,t}$, is defined as

$$N^{fs}_{k,t} \equiv (N_{k,t} | \Phi_{k,t} \cdot T_I \geq (\|\mathbf{D}_k\| - \nu_{k,t-1}) \cdot \varepsilon_k). \tag{21}$$

- Minimum guarantee (MinG) [51]: evenly distributes a portion of antennas to a subset of UEs, $\mathbf{O}^G_t \subseteq \mathbf{O}_t$, and applies FSO on the remaining resources. Therefore, several UEs can receive a minimum share of antennas, and the portion of resources reserved for even distribution, ι^G, is a key parameter to consider. We determine the number of UEs receiving guaranteed resources according to the smallest $N^{fs}_{k,t}$ and can be expressed as

$$\|\mathbf{O}^G_t\| = \frac{\iota^G \cdot M}{\min_{k \in \mathbf{O}_t} N^{fs}_{k,t}}, \tag{22}$$

  where $\iota^G = \{25\%, 50\%, 75\%, 100\%\}$. Thus, there are four MinG-based options in $\mathbf{C}_2$. For example, the option with $\iota^G = 50\%$ is denoted as MinG50.
- Proportional fair (PF) [52]: considers a subset of UEs and allocates antenna resources proportional to the ratio of currently available data rate, $\Phi_{k,t}$, to the historical transmission rate. In practice, the historical transmission rate can be updated through moving averages. The parameter $\iota^{pf} = \{25\%, 50\%, 75\%, 100\%\}$ determines the percentage of UEs in $\mathbf{O}_t$ to be included.

The complete option set $\mathbf{C}_2$ has nine elements with all the schemes and parameters.

The third component selects a precoding algorithm for high spectrum efficiency in massive MIMO transmission to evaluate the precoding matrix. The implemented precoders are:

- Antenna selection (AS) [53]: greedily chooses antennas to achieve high single antenna efficiency.
- Cross entropy (CE) [54]: is a probabilistic model-based algorithm iteratively solving the combining problem. The algorithm computes the achievable sum-rate of each candidate and selects the best candidates as "elites." The probability distribution is updated based on the selected

elites by minimizing the cross entropy. CE precoding performs well with sufficient resources and a less saturated system.
- Adaptive cross entropy (ACE) [55]: is a variation based on the CE algorithm. The ACE algorithm weights "elites" adaptively based on their achievable sum-rates. This precoding method can gain better SINR than CE in saturated situations.

The component $c_{3,t} \in \mathbf{C}_3 = \{\text{AS, CE, ACE}\}$. Overall, the action $\mathbf{a}_t$ is one of 108 component combinations with all options considered.

C. ACTION EMBEDDING AND TRAINING PROCEDURES

As introduced in Section II-B, DDPG takes advantage of DQN, DPG, and the actor-critic structure [23]; it is utilized to make resource allocation decisions for MDP problems with continuous or high dimensional states and actions. Furthermore, we extend the actions in this work to a continuous space through action embedding [24], where the original discrete actions are embedded in continuous upon which the actor can generalize. The function $\upsilon : \mathbb{R}^{\dim(\mathcal{A})} \rightarrow \mathcal{A}$ is defined to convert the *continuous* action $\check{\mathbf{a}}_t$ used for training into the *discrete* action $\mathbf{a}_t$ applied to the environment, with $\dim(\mathcal{A})$ denoting the dimension of action space $\mathcal{A}$. Therefore, the converting function is expressed as

$$\mathbf{a}_t = \upsilon(\check{\mathbf{a}}_t), \tag{23}$$

where $\check{\mathbf{a}}_t = [\check{c}_{1,t}, \check{c}_{2,t}, \check{c}_{3,t}]$ is the action formed by the continuous component values. Also, the deterministic policy generating continuous action $\check{\mu} : \mathcal{S} \rightarrow \mathbb{R}^{\dim(\mathcal{A})}$ is applied in the model as the actor network $\check{\mu}(\mathbf{s}_t|\theta^{\mu})$.

The training process is described in Algorithm 1. First, networks are initialized. Then, for every TTI, the agent generates continuous action $\check{\mathbf{a}}_t = \check{\mu}(\mathbf{s}_t|\theta^{\mu}) + \mathbf{n}_t^{exp}$ using the actor network with exploration noise $\mathbf{n}_t^{exp}$ from random process $\mathcal{N}$. The discrete action $\mathbf{a}_t$ is obtained from (23) and applied to the environment for the reward r_t and the next state $\mathbf{s}_{t+1}$ as feedbacks. In order to reuse execution experiences, DDPG stores transition $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t)$ in the replay buffer. After that, DDPG samples B number of transitions from the replay buffer to form a mini-batch $\mathcal{B}$. With mini-batch inputs, the target actor network $\check{\mu}'(\mathbf{s}_{t+1}|\theta^{\mu'})$ outputs the action to the target critic network Q'. The resulting action-value can be evaluated based on (3). Therefore, the critic network is updated by minimizing the loss function

$$L(\theta^{Q}) = \frac{1}{B}\sum_{i\in\mathcal{B}}\Big[r_i + \gamma Q'(\mathbf{s}_{i+1}, \upsilon(\check{\mu}'(\mathbf{s}_{i+1}|\theta^{\mu'}))|\theta^{Q'}) - Q(\mathbf{s}_i, \mathbf{a}_i|\theta^{Q})\Big]. \tag{24}$$

The actor network is updated following the deterministic policy gradient theorem modified from (4) as [24]

$$\nabla_{\theta^{\mu}} J(\theta^{\mu}) \approx \frac{1}{B}\sum_{i\in\mathcal{B}} \nabla_{\check{\mathbf{a}}} Q(\mathbf{s}_i, \check{\mathbf{a}}_i|\theta^{Q}) \nabla_{\theta^{\mu}} \check{\mu}(\mathbf{s}_i|\theta^{\mu}). \tag{25}$$

Finally, DDPG uses the soft-update to improve critic and actor target networks with the constant τ as

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}, \\ \theta^{\mu'} &\leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}, \end{aligned} \tag{26}$$

where $\leftarrow$ represents the assignment operator in the algorithm. As a result, the parameters in target networks change slowly and considerably improve the learning stability.

Algorithm 1 The DDPG Training With Action Embedding

```
Randomly initialize critic network Q and actor network μ̌ in the DDPG agent
Initialize target network Q' and μ̌' with weights θ^{Q'} ← θ^{Q}, θ^{μ'} ← θ^{μ}.
Initialize replay buffer
for episode = 1 to end do
    Initialize a random process N for action exploration
    Receive initial observation state s_1
    for t = 1 to T do
        Generate continuous action ǎ_t = μ̌(s_t|θ^{μ}) + n_t^{exp} from actor in DDPG
        Convert the action form continuous to discrete a_t = υ(ǎ_t) to embedding on three components [c_{1,t}, c_{2,t}, c_{3,t}]
        Execute action a_t and observe reward r_t and new state s_{t+1}
        Store transition (s_t, a_t, r_t, s_{t+1}) in replay buffer
        Sample a random mini-batch from the replay buffer
        Update the critic by minimizing the loss (24)
        Update the actor using the gradient (25)
        Update the actor and critic network with the equation (25)(26)
    end for
end for
```

The computational complexity of DDPG can be evaluated using floating-point operations per second (FLOPS). Let l^{actor} and l^{critic} be the number of fully connected layers of actor and critic networks, including hidden and output layers. Then, the asymptotic upper bound, $\mathcal{O}$, of training time is [56]

$$\mathcal{O}\left(\sum_{i=0}^{l^{actor}-1} \eta_i^{actor}\eta_{i+1}^{actor} + \sum_{i=0}^{l^{critic}-1} \eta_i^{critic}\eta_{i+1}^{critic}\right), \tag{27}$$

where η_i^{actor} and η_i^{critic} are the numbers of units in the i-th layer of networks. Also, since only the actor network is involved during the executing phase, the executing time complexity is

$$\mathcal{O}\left(\sum_{i=0}^{l^{actor}-1} \eta_i^{actor}\eta_{i+1}^{actor}\right). \tag{28}$$

The sizes of input layers, η_0^{actor} and η_0^{critic}, equal the state's dimension and increase with number of MCS levels and QoS types. Therefore, it does not scale with the number of UEs because only UE statistics are recorded in our state design.

TABLE 2. Traffic type parameters [14].

Type (5QI Value)	Application	Packet Size (Bytes)	Mean Packet Arrival Time (ms)	Latency (ms)	GBR (Mb/s)	Error Rate
A (1)	VoIP	200	15	100	0.112	10^{-2}
B (2)	Video	1250	5	150	0.8	10^{-3}
C (3)	Online Game	500	8	50	0.72	10^{-3}
D (80)	VR/AR	1250	2	10	-	10^{-6}
E (7)	Video Streaming	1250	10	100	-	10^{-3}
F (8)	FTP	1250	6	300	-	10^{-2}

TABLE 3. Scenarios in various UE traffic type ratios.

Scenario	A : B : C : D : E : F
Scenario 1	1 : 1 : 1 : 1 : 1 : 1 (Balanced)
Scenario 2	1 : 1 : 1 : 2 : 1 : 1 (Double VR/AR traffic)
Scenario 3	1 : 1 : 1 : 1 : 1 : 2 (Double FTP traffic)
Scenario 4	3 : 3 : 3 : 1 : 1 : 1 (More UEs with GBR)
Scenario 5	1 : 1 : 2 : 2 : 1 : 1 (More UEs with short latency)
Scenario 6	1 : 1 : 1 : 2 : 1 : 2 (More UEs with high data rate)

The output size of the actor network $\eta^{actor}_{|actor}$ is the dimension of action, which is the number of function components. The output of the critic network $\eta^{critic}_{|critic}$ is the Q value which has a dimension of 1. Furthermore, after automated DDPG-based algorithm selection, the selected fundamental algorithms are executed in their conventional way with no extra computation demands. Therefore we focus on DDPG complexity analysis in this article.

V. NUMERICAL RESULTS

This section introduces simulation settings for traffic scenarios, the massive MIMO environment, and DDPG training. Numerical results compare the proposed learning-based method with baselines, including static combinations of fundamental methods and related works.

A. SIMULATION SETUP

The simulation scenarios are built as mixes of applications in a massive MIMO system. Table 2 shows six selected traffic types based on 5QI specifications [14], including voice over IP (VoIP), video streaming, gaming, and virtual reality (VR) / augmented reality (AR). The properties attached to a traffic type include latency, GBR, packet size, mean packet arrival time, and error rate requirements. A UE is a traffic session with a predetermined type and properties to generate requested data. Traffic sessions from all types are mixed in various UE ratios listed in Table 3 with specific focuses to form scenarios. For the communication system, COST2100 [57] is used to model the MIMO channel, and a varying number of active UEs are distributed following the Poisson point process (PPP). The channel is also assumed to be under block fading. The massive MIMO antenna allocation model [7] is also simulated under a small cell VR environment [49], [58]. Table 4 lists the complete communication system parameters.

TABLE 4. Communication system parameters [7], [49], [57], [58].

Parameter	Value
Operating Frequency	3.5GHz
Number of Antennas per BS (M)	128
Total Number of UEs (K)	100 to 600
Average Number of Co-existing UEs (K_r)	8.3 to 50
Bandwidth (B)	20 MHz
Transmit Power (ρ)	24 dBm
Transmission Time Interval (TTI, T_I)	1 ms
Channel Model	COST2100
Antennas Array Type	Cylindrical Array
UEs Distribution	Poisson Point Process
Cell Range	radius of 100m
Termination Time (T)	60000 TTI

TABLE 5. DDPG parameters.

Parameter	Value
Replay Buffer Size	600000 TTI
Mini Batch Size	60000 TTI
Actor Learning Rate	$2e^{-3}$
Critic Learning Rate	$1e^{-3}$
Reward Discount, γ	0.9
Soft Replacement τ	$1e^{-2}$
Dropout Rate	0.5
Explore Rate	0.1
Actor Network Size	41 (input)
	512, 512, 512 ReLU
	3, tanh (output)
Critic Network Size	41 (input)
	512, 512, 512 ReLU
	1 (output)
Optimizer	Adam

The simulation datasets are formed by 60000-TTI-long data blocks containing CQIs and requested data of UEs every TTI. We generate four data blocks for each of six traffic types for training, resulting in 24 distinguish traffic data blocks. The training goes through 24 data blocks in random orders in an epoch. Therefore the resulting model can handle traffic scenarios in an arbitrary mix of data types. The testing is performed on ten separately generated data blocks for each scenario. Also, the penalty weight α in (19) is set to 0.5 when GBR is available. The continuous component values in (23) are set in $[-1, 1]$ and evenly distributed for discrete actions with $\dim(\mathcal{A}) = 3$. The training and decision-making models are implemented using TensorFlow [59] library version 1.14 on a desktop machine with an Intel i7-3770 CPU and Nvidia RTX 2080Ti GPU. The under hyperparameters are determined by grid search [60] and listed in Table 5. As shown in Figure 4, the model converges to stable total rewards after 75 epochs of training in 102 minutes on average.

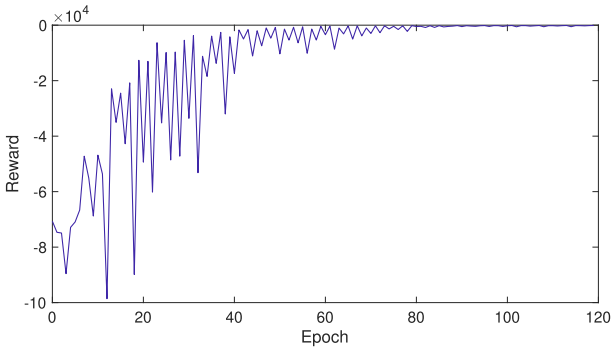


FIGURE 4. Convergence plot. The proposed DDPG model converges to stable total rewards after 75 epochs of training in 102 minutes on average.

Several static fundamental method combinations and algorithms in the literature are compared with the proposed learning method (Learning). The static actions, denoted in the abbreviations introduced in Section IV-B, are

- CQI-MinG75-AS: applies channel quality first, minimum guarantee with 75% resources reserved, and antenna selection precoder.
- CQI-PF50-ACE: applies channel quality first, proportional fair with 50% of UEs included, and adaptive cross entropy precoder.
- Delay-MinG75-ACE: applies expiring time first, minimum guarantee with 75% resources reserved, and adaptive cross entropy precoder.
- Remain-MinG50-ACE: applies remaining data first, minimum guarantee with 50% resources reserved, and adaptive cross entropy precoder.

These benchmark combinations are the most frequently selected ones from the learning results and are kept static during simulation runs.

Related joint resource allocation works are also compared. Due to variations in system setups, we extract algorithms from related works for our environment with proposed concepts maintained. The selected algorithms have to cover user prioritization and are compatible with the antenna allocation process. Also, if necessary, we supplement the algorithms with the best-performing precoding methods for a fair evaluation. The comparing algorithms are

- ORFA [32]: specified all three function components: proportional fair UE ranking, water filling resource allocation, and linear MMSE precoding.
- UBLAA [7]: defines the marginal utility in a massive MIMO video streaming system to prioritize UEs for antenna allocation. Then, the AS precoder, which matches the greedy-based UBLAA algorithm, is applied.
- LWDF-JAUS: is a combination of LWDF-PF [61] QoS scheduling and JAUS-LCC [8] massive MIMO antenna allocation algorithms. LWDF-PF is a landmark proportional fair algorithm that adopts weighted delay fairness for UE scheduling, while JAUS-LCC determines the number of antennas used for receiving UEs. The combination is necessary to extend the JAUS-LCC for fair comparison in our scenarios.

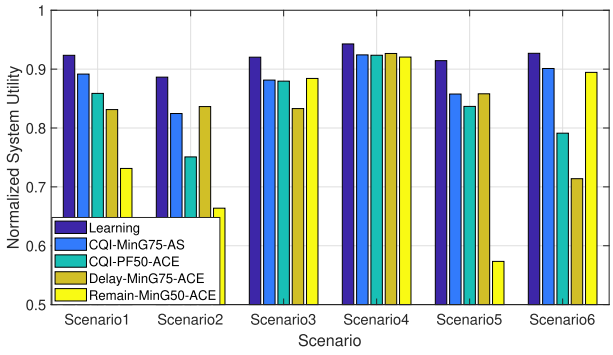
B. DYNAMIC VS. STATIC ALGORITHM COMBINATIONS

In this section, we compare the proposed learning-based method against static combinations to demonstrate the advantages of automated algorithm selection. In addition, performance metrics in system utilities and throughputs are illustrated.

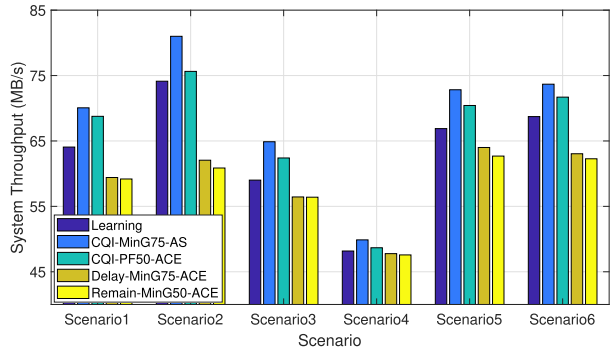
Figure 5a illustrates the normalized system utility defined as the percentage of satisfied UEs through termination time T. Due to its adaptive nature, the proposed learning-based approach gains 2.2% to 7.2% more system utility than the best static scheme across all scenarios. The most significant advantage appears in Scenario 2 with doubled VR/AR traffic, showing that the learning method can achieve high bandwidth and low latency simultaneously. The performances of static schemes are inconsistent across application scenarios. For example, the delay emphasizing scheme, Delay-MinG75-ACE, ranks second in Scenarios 2 and 5, where more latency demanding VR/AR or gaming traffic exists. On the other hand, the scheme Remain-MinG50-ACE is comparable with the best ones in data rate demanding Scenario 3, 4, and 6. However, it achieves significantly less in others because UEs with more remaining data are ranked higher. Furthermore, CQI-MinG75-AS is a more versatile static combination because CQI provides high system throughput while MinG75 forces even distribution of most antenna resources. The greedy nature of AS precoder also fits well with CQI and MinG75.

From the system throughputs presented in Figure 5b, we observe that greater throughput not necessarily reflects greater utility. Schemes that apply the CQI method for c_1, CQI-MinG75-AS and CQI-PF50-ACE, result in the highest throughputs because UEs are ranked according to channel quality. The proposed learning-based method is ranked only behind CQI methods in throughput and outperforms them in system utility. When the overall traffic demand and throughput are lower in Scenario 4, all schemes achieve system utility greater than 0.9.

Algorithm selection details in Figure 6 can further reveal the advantage of the learning method. The figure breaks down the proportion of top actions chosen by the DDPG agent in simulated scenarios. We observe that the learned strategies are adjusted accordingly to the scenarios. In Scenario 2 and 5, where the proposed method gains more than 7% utility, CQI-Min75-AS and Delay-MinG75-ACE are most frequently selected with more than 70% TTIs combined. It shows that timely switching between throughput and delay priority timely is an effective strategy for simultaneously fulfilling throughput and latency requirements. In Scenario 3 and 6, CQI-Min75-AS and Remain-MinG50-ACE are most applied for high data rate applications. When the demand is low in Scenario 4, CQI-PF50-ACE is more applied to emphasize proportional fairness. In balanced Scenario 1, most action combinations other than the top five are selected. Overall,



(a) Normalized System Utility.



(b) Throughput

FIGURE 5. The learning based automated algorithm selection method compares with best-performing static algorithm combinations.

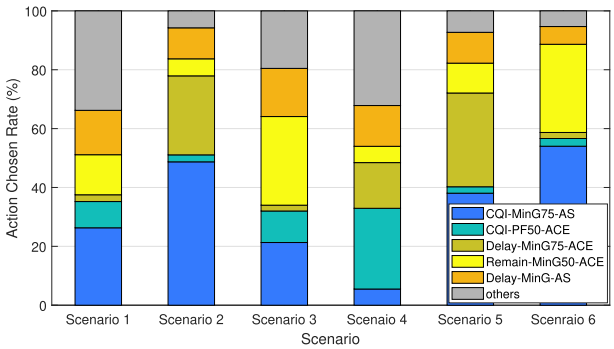
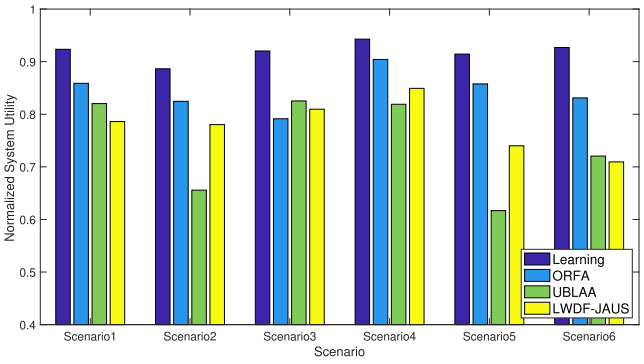


FIGURE 6. The proportion of top actions chosen by the DDPG agent in simulated scenarios.

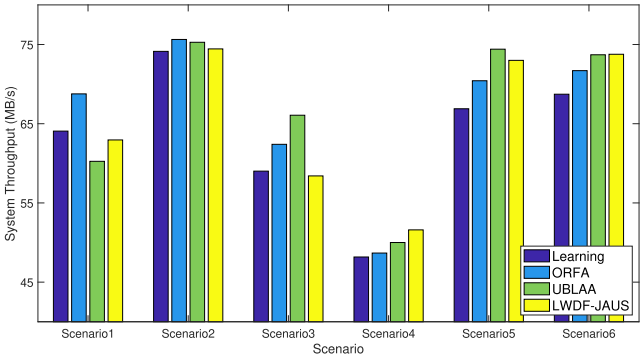
the simulations demonstrate the effectiveness of forming an advanced resource allocation solution by switching between fundamental algorithms. The proposed componentized action structure is the key to realizing this concept under advanced DRL training.

C. COMPARISON WITH JOINT RESOURCE ALLOCATION ALGORITHMS

Figures 7a and 7b present the normalized system utility and system throughput compared with joint resource allocation algorithms in the literature. The proposed learning approach outperforms ORFA, UBLLA, and LWDF-JAUS algorithms in normalized system utility, though not providing the highest throughputs. The largest utility gap is at 12.5% in Scenario 6, with heavy traffic on high data rate and low latency types. The smallest gap presents in the less loaded Scenario 4 at



(a) Normalized System Utility

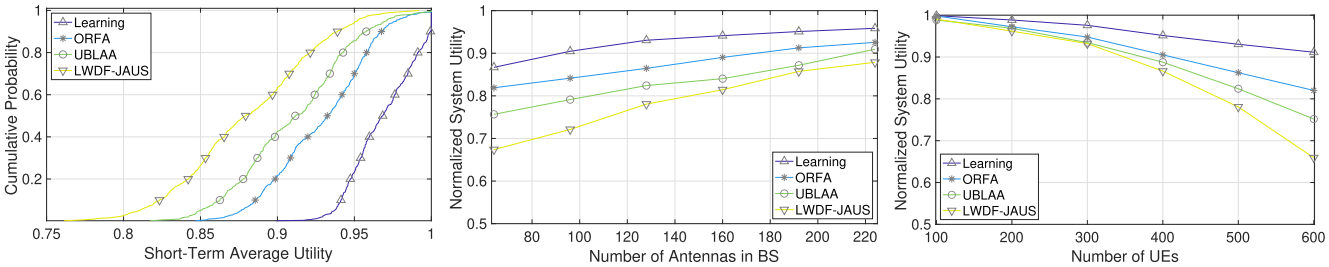


(b) Throughput

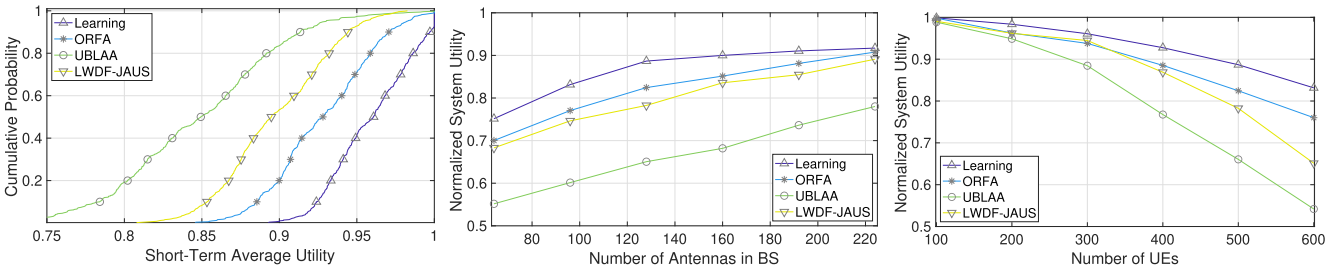
FIGURE 7. The learning based automated algorithm selection method compares with joint resource allocation algorithms in the literature.

4.4%. ORFA consistently achieves greater than 0.8 in utility due to the optimality of the water-filling algorithm. However, its' general-purpose proportional fair scheduling suffers from degraded performance under diverse application requirements. UBLAA fulfills data rate requirements and results in high throughput in all scenarios. However, since latency is not effectively presented via marginal utility, system utility performance is not satisfactory in Scenarios 2, 5, and 6, with latency-demanding VA/AR applications. LWDF-JAUS performs worse than the learning and ORFA methods but achieves top throughput results, because the proportional fair strategy and effective antenna allocation provide system-wide throughput advantages. However, it also suffers significant utility drops in Scenarios 5 and 6 due to less cross-layer QoS consideration.

Figure 8 shows detailed results for two representative scenarios. Scenario 1 with a balanced traffic mixture and Scenario 2 emphasizing VR/AR applications are selected. Suppose we divide the whole simulation into 100-TTI windows. In that case, the average utility of UE sessions ending in the same 100-TTI windows is evaluated as *short-term average utility* to analyze the system condition over time. Figures 8a and 8d illustrate the cumulative distribution function (CDF) of short-term average utilities with 128 antennas and 500 UEs. We can see that the learning method spread mainly to 0.9 and above. ORAF has samples lower than 0.85. LWDF-JAUS outperforms UBLAA in Scenario 2 and is close to ORFA when the resources are sufficient, i.e., more



(a) CDF of short-term average utility for the balanced traffic scenario (b) System utility vs. number of antennas for the balanced traffic scenario (c) System utility vs. number of UEs for the balanced traffic scenario



(d) CDF of short-term average utility for the AR/VR emphasized scenario (e) System utility vs. number of antennas for the AR/VR emphasized scenario (f) System utility vs. number of UEs for the AR/VR emphasized scenario

FIGURE 8. Simulation results of short-term average utility and the system utility under the various number of antennas and UEs. The representative scenarios are: Scenario 1 with a balanced traffic mixture and Scenario 2 emphasizing VR/AR applications.

antennas or fewer UEs. UBLAA keeps all samples greater than 0.83 in the balanced condition, while some samples fall below 0.75 when there is more latency-sensitive traffic in the system, like in Scenario 2. Figures 8b and 8e present the system utility trend using 64 to 224 BS antennas with 500 total (42 average coexist) UEs. In general, systems gain more utility with more antennas. When the resources are limited to 64 antennas, the learning method gains 6.2% to 40% more utility than others in the balanced cases and 7.1% to 22% more in VR/AR emphasized cases. Figures 8c and 8f show the system utility trend with 100 to 600 total (8.3 to 50 average coexist) UEs at 128 antennas. The advantage of learning-based algorithm selection grows with the saturation level resulting from more UEs. Also, overall utilities drop faster in VR/AR emphasized Scenario 2 than Scenario 1.

To summarize, the comparing joint methods fulfill the user scheduling and resource allocation problem objective (15) in general, where the decision is made to maximize the instant utility $U_{k,t}$. In contrast, the proposed MDP-based method maximizes the long-term utility (14) and thus joint objective (15), because maximizing the long-term return (1) is the nature of MDP. Furthermore, the cross-layer integration of scheduling and precoding also shows effectiveness.

VI. CONCLUSION

A DRL-based radio resource allocation approach for joint scheduling and precoding in a massive MIMO system is investigated in this work. We suggest an architecture decomposing the cross-layer adaptation decision as a combination of algorithms and learning a dynamic algorithm selection policy in challenging 5G traffic scenarios. Comprehensive simulations are carried out to justify the effectiveness of the proposed method. Overall, the proposed automated algorithm selection framework can be the core of an extensible smart agent to deal with complex decision-making problems in future mobile networks. Future works will include the self-adaptation of machine learning models and decisions under imperfect channel estimation.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[2] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G wireless communications: Vision and potential techniques," *IEEE Netw.*, vol. 33, no. 4, pp. 70–75, Jul. 2019.

[3] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.

[4] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.

[5] X. Chen, F.-K. Gong, H. Zhang, and G. Li, "Cooperative user scheduling in massive MIMO systems," *IEEE Access*, vol. 6, pp. 21910–21923, 2018.

[6] J. Choi, N. Lee, S.-N. Hong, and G. Caire, "Joint user scheduling, power allocation, and precoding design for massive MIMO systems: A principal component analysis approach," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 396–400.

[7] W.-H. Kuo and Y.-H. Lu, "Antenna allocation scheme for providing scalable video coded streams over multiple-base-station massive MIMO cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2314–2323, Mar. 2018.

[8] Y.-X. Zhu, D.-Y. Kim, and J.-W. Lee, "Joint antenna and user scheduling in the massive MIMO system over time-varying fading channels," *IEEE Access*, vol. 9, pp. 92431–92445, 2021.

[9] S. Han, T. Xie, I. Chih-Lin, Li Chai, Z. Liu, Y. Yuan, and C. Cui, "Artificial-intelligence-enabled air interface for 6G: Solutions, challenges, and standardization impacts," *IEEE Commun. Mag.*, vol. 58, no. 10, pp. 73–79, Oct. 2020.

[10] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. Zhang, "Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 212–217, Apr. 2019.

[11] C. Fiandrino, C. Zhang, P. Patras, A. Banchs, and J. Widmer, "A machine-learning-based framework for optimizing the operation of future networks," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 20–25, Jun. 2020.

[12] X. Guo, Z. Li, P. Liu, R. Yan, Y. Han, X. Hei, and G. Zhong, "A novel user selection massive MIMO scheduling algorithm via real time DDPG," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[13] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis," *Mach. Learn.*, vol. 110, no. 9, pp. 2419–2468, Sep. 2021.

[14] *System Architecture for the 5G System*, 3GPP, document TS 23.501, 2016.

[15] M. G. Lagoudakis and M. L. Littman, "Algorithm selection using reinforcement learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 511–518.

[16] A. Loreggia, Y. Malitsky, H. Samulowitz, and V. Saraswat, "Deep learning for algorithm portfolios," in *Proc. 30th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 1–7.

[17] L. Kotthoff, "Algorithm selection for combinatorial search problems: A survey," in *Data Mining and Constraint Programming* (Lecture Notes in Computer Science), vol. 10101. Cham, Switzerland: Springer, Dec. 2016, pp. 149–190.

[18] P. Kerschke, H. H. Hoos, F. Neumann, and H. Trautmann, "Automated algorithm selection: Survey and perspectives," *Evol. Comput.*, vol. 27, no. 1, pp. 3–45, Mar. 2019.

[19] S.-C. Tseng, Z.-W. Liu, Y.-C. Chou, and C.-W. Huang, "Radio resource scheduling for 5G NR via deep deterministic policy gradient," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–6.

[20] P.-C. Chen, Y.-C. Chen, W.-H. Huang, C.-W. Huang, and O. Tirkkonen, "DDPG-based radio resource management for user interactive mobile edge networks," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Levi, Finland, Mar. 2020, pp. 1–5.

[21] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 21–27, Oct. 2021.

[22] M. Yao, M. Sohul, V. Marojevic, and J. H. Reed, "Artificial intelligence defined 5G radio access networks," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 14–20, Mar. 2019.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2016, pp. 1–14.

[24] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," 2015, *arXiv:1512.07679*.

[25] G. Kwon and H. Park, "A joint scheduling and millimeter wave hybrid beamforming system with partial side information," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[26] T. A. Sheikh, J. Bora, and M. A. Hussain, "Capacity maximizing in massive MIMO with linear precoding for SSF and LSF channel with perfect CSI," *Digit. Commun. Netw.*, vol. 7, no. 1, pp. 92–99, Feb. 2021.

[27] S. Lagen, A. Agustin, and J. Vidal, "Joint user scheduling, precoder design, and transmit direction selection in MIMO TDD small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2434–2449, Apr. 2017.

[28] A. Almradi, M. Matthaiou, P. Xiao, and V. F. Fusco, "Hybrid precoding for massive MIMO with low rank channels: A two-stage user scheduling approach," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4816–4831, Aug. 2020.

[29] M. Al-Saedy, M. Al-Imari, M. Al-Shuraifi, and H. Al-Raweshidy, "Joint user selection and multimode scheduling in multicell MIMO cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10962–10972, Dec. 2017.

[30] T. Xie, L. Dai, D. W. K. Ng, and C.-B. Chae, "On the power leakage problem in millimeter-wave massive MIMO with lens antenna arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4730–4744, Sep. 2019.

[31] M. Olyaee, M. Eslami, and J. Haghighat, "An energy-efficient joint antenna and user selection algorithm for multi-user massive MIMO downlink," *IET Commun.*, vol. 12, no. 3, pp. 255–260, Nov. 2018.

[32] S. Singh, M. Geraseminko, S.-P. Yeh, N. Himayat, and S. Talwar, "Proportional fair traffic splitting and aggregation in heterogeneous wireless networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1010–1013, May 2016.

[33] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.

[34] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[35] R. Li, Z. Zhao, Q. Sun, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[36] Y. Yang, Y. Li, K. Li, S. Zhao, R. Chen, J. Wang, and S. Ci, "DECCO: Deep-learning enabled coverage and capacity optimization for massive MIMO systems," *IEEE Access*, vol. 6, pp. 23361–23371, 2018.

[37] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[38] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209–4219, Jun. 2020.

[39] T. V. Chien, T. N. Canh, E. Björnson, and E. G. Larsson, "Power control in cellular massive MIMO with varying user activity: A deep learning solution," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5732–5748, May 2020.

[40] D. Zeng, L. Gu, S. Pan, J. Cai, and S. Guo, "Resource management at the network edge: A deep reinforcement learning approach," *IEEE Netw.*, vol. 33, no. 3, pp. 26–33, May/Jun. 2019.

[41] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[42] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[43] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, vol. 1, 2014, pp. 605–619.

[44] *NR; Physical Layer Procedures for Data*, 3GPP, document TS 38.214, 2020.

[45] Y. Zhang, P. Mitran, and C. Rosenberg, "Joint resource allocation for linear precoding in downlink massive MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3039–3053, May 2021.

[46] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

[47] X. Gao, L. Dai, Z. Gao, T. Xie, and Z. Wang, "Precoding for mmWave massive MIMO," in *mmWave Massive MIMO*, S. Mumtaz, J. Rodriguez, and L. Dai, Eds. New York, NY, USA: Academic, 2017, pp. 79–111.

[48] A. Ghazanfari, T. Van Chien, E. Bjornson, and E. G. Larsson, "Model-based and data-driven approaches for downlink massive MIMO channel estimation," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 2085–2101, Mar. 2022.

[49] M. Chen, W. Saad, and C. Yin, "Virtual reality over wireless networks: Quality-of-service model and learning-based resource management," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5621–5635, Nov. 2018.

[50] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. 15th ACM Workshop Hot Topics Netw.*, New York, NY, USA, Nov. 2016, pp. 50–56.

[51] N. Guan, Y. Zhou, L. Tian, G. Sun, and J. Shi, "QoS guaranteed resource block allocation algorithm for LTE systems," in *Proc. IEEE 7th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2011, pp. 307–312.

[52] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 210–212, Mar. 2005.

[53] R. Mendez-Rial, C. Rusu, A. Alkhateeb, N. Gonzalez-Prelcic, and R. W. Heath Jr., "Channel estimation and hybrid combining for mmWave: Phase shifters or switches?" in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2015, pp. 90–97.

[54] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Cham, Switzerland: Springer, 2013.

[55] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[56] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.
[57] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. D. Doncker, "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.
[58] C. Wang, R. C. Elliott, D. Feng, W. A. Krzymien, S. Zhang, and J. Melzer, "A framework for MEC-enhanced small-cell HetNet with massive MIMO," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 64–72, Aug. 2020.
[59] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[60] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 473–480.
[61] T. Girici, C. Zhu, J. R. Agre, and A. Ephremides, "Proportional fair scheduling algorithm in OFDMA-based wireless systems with QoS constraints," *J. Commun. Netw.*, vol. 12, no. 1, pp. 30–42, Feb. 2010.

CHIH-WEI HUANG (Member, IEEE) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, in 2001, the M.S. degree in electrical engineering from Columbia University, New York, NY, USA, in 2004, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2009.

From 2006 to 2009, he was an Intern Researcher at Siemens Corporate Research and Microsoft Research. He joined the Department of Communication Engineering, National Central University, Taoyuan, Taiwan, in 2010. He is currently an Associate Professor heading with the Information Processing and Communications (IPC) Laboratory, National Central University. He is the author of articles in areas including wireless networking, multimedia communications, machine learning, digital signal processing, and information retrieval. He received the best paper awards from the IEEE ICCE 2020, IEEE ICC 2018, and WOCC 2015 conferences.

IBRAHIM ALTHAMARY received the bachelor's degree in computer science from Thamar University, in 2010, and the master's degree in computer network from the King Fahd University of Petroleum and Minerals (KFUPM), in 2017. He is currently pursuing the Ph.D. degree in communication engineering with the National Central University. He has been a Researcher with the Information Processing and Communications (IPC) Laboratory, National Central University, since 2017. His research interests include machine learning, C-V2X, computer security, fog computing, 5G, 6G, and the IoT.

YEN-CHENG CHOU received the B.S. degree in communication engineering from Yuan Ze University, Taoyuan, Taiwan, in 2018, and the M.S. degree in radio resource management for 5G networks in communication engineering from the National Central University, in 2020.

HONG-YUNN CHEN is currently pursuing the Ph.D. degree with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan. He has been a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA. His current research interests include massive MIMO systems and cloud computing systems.

CHENG-FU CHOU received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 2002. After graduation, he joined the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, where he is currently a Professor. He has been a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA. His current research interests include peer-to-peer networks, distributed multimedia systems, 5G communication networks, sensor networks, and their performance evaluation.