IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2023, accepted 26 December 2023, date of publication 28 December 2023,
date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347788

RESEARCH ARTICLE

Multi-Objective Reinforcement Learning for Power Allocation in Massive MIMO Networks: A Solution to Spectral and Energy Trade-Offs

YOUNGWOO OH, (Student Member, IEEE), ARIF ULLAH, (Member, IEEE), AND WOOYEOL CHOI, (Member, IEEE)
College of IT Convergence, Department of Computer Engineering, Chosun University, Gwangju 61452, Republic of Korea

Corresponding author: Wooyeol Choi (wyc@chosun.ac.kr)

This work was supported by the Research Fund from Chosun University, in 2023.

ABSTRACT The joint optimization of spectral efficiency (SE) and energy efficiency (EE) through power allocation (PA) techniques is a critical requirement for emerging fifth-generation and beyond networks. The trade-off between SE and EE becomes challenging in the massive multiple-input-multiple-output (MIMO) equipped base stations (BSs) in multi-cell cellular networks. Various algorithmic approaches including genetic algorithms and convex optimization have been considered to optimize the trade-offs between SE and EE in cellular networks. However, these methods suffer from high computational costs. A promising deep reinforcement learning technique is capable of addressing the computational challenges of single-objective optimization problems in wireless networks. Furthermore, multi-objective reinforcement learning has been employed for multi-objective optimization problems and can be utilized to jointly enhance the SE and EE in cellular networks. In this paper, we propose a downlink (DL) transmit PA method based on a multi-objective asynchronous advantage single actor-multiple critics (MO-A3Cs) architecture. The proposed architecture aims to optimize SE and EE trade-offs in massive MIMO-assisted multi-cell networks. Furthermore, we also propose a Bayesian rule-based preference weight updating mechanism, multi-objective advantage function, and balanced-reward aggregation method to effectively train and avoid biased objective reward during the training process of the proposed model. Extensive simulations depict that the proposed model is better capable of dealing with the joint optimization of SE and EE in dynamic changing scenarios. Compared to the existing benchmarks such as Pareto front approximation-based multi-objective, reinforcement learning-based single objective, and iterative methods, the proposed approach provides a better SE-EE trade-off by achieving a higher EE in multi-cell massive MIMO networks.

INDEX TERMS 5G and beyond networks, energy efficiency, massive MIMO, multi-objective reinforcement learning, power allocation, spectral efficiency.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is one of the key technologies of fifth-generation (5G) and beyond networks, capable of enhancing spectral efficiency (SE) and cell coverage. Massive MIMO utilizes multi-antenna transmissions at the base station (BS) to simultaneously serve multiple user equipment (UEs) [1], [2]. The impact of fading and interference in massive MIMO can be reduced through spatial diversity and multiplexing gain. Moreover, the link reliability and transmission rate are improved by leveraging the spatial domain to precisely focus energy toward the

The associate editor coordinating the review of this manuscript and approving it for publication was Walid Al-Hussaibi.

intended UE. The multi-user MIMO (MU-MIMO) systems utilize the same frequency resources to serve multiple users at the same time, which leads to more efficient use of scarce spectrum resources and provides more tolerance to propagation losses compared to single-user MIMO (SU-MIMO) systems. However, deploying a large number of antennas requires high transmission power which results in higher interference, degraded overall network performance, and significantly elevated energy consumption [3]. A remarkable advancement has been made to enhance the performance of downlink (DL) MU-MIMO systems, primarily focusing on tackling the high energy consumption of the cellular network by achieving a reasonable trade-off between SE and energy efficiency (EE) [4], [5]. Furthermore, the joint optimization of SE and EE in 5G and beyond 5G (B5G) networks is crucial due to the exponential increase in data traffic and the associated rise in energy consumption in multi-cell massive MIMO networks. This optimization is imperative to efficiently utilize the spectrum resources and address environmental impacts by reducing the network energy consumption [6], [7]. The resource allocation techniques, associated with the joint optimization of SE and EE, seriously influence the overall network performance and play a vital role in harnessing the full potential of multi-cell massive MU-MIMO networks. Optimal power allocation (PA) that efficiently utilizes the power resources while ensuring the quality-of-service (QoS) requirements of the UEs is a challenging task. The PA problem becomes more crucial in managing the high inter-cell and intra-cell interference in a densely deployed network scenario.

Different algorithmic approaches such as genetic algorithm and convex optimization are used to solve multi-objective optimization (MOO) problems. However, using these conventional techniques for MOO-based PA faces limitations such as scalability and high computational complexity which increases exponentially with the number of antennas in massive MIMO systems [8], [9]. In this regard, deep learning-based PA schemes are proposed that can achieve near-optimal performance while addressing the computational complexity issues inherited by the iterative algorithm-based PA techniques [10], [11], [12]. However, deep learning-based approaches require additional datasets and do not perform well in dynamically changing wireless network scenarios.

Deep reinforcement learning (DRL) is an emerging technique that employs the Markov decision process (MDP) framework to solve optimization problems. Through a trial-and-error strategy, DRL algorithms utilize interactions between agents and environments to determine optimal policies for solving problems. The DRL algorithm has the potential to effectively deal with computationally complex optimization problems in dynamic wireless networks [13], [14], [15], [16], [17], [18], [19], [20]. Keeping in view, the escalating importance of the joint optimization of SE and EE in the 5G and B5G networks, this work aims to design an efficient multi-objective reinforcement learning (MORL) framework for PA which ensures effective training and convergence while addressing the MOO problem in multi-cell massive MIMO systems.

A. RELATED WORKS

DRL has been applied to multiple problems in wireless networks. The authors in [13] proposed a deep Q-network (DQN)-based PA method to enhance the sum rate in multi-cell networks, using the maximized sum rate as the reward. The states considered for the actions selected by the DQN agents include normalized interference, DL rate, and transmit power. Similarly, the authors in [14] defined the MDP for sum rate maximization, considering the previous transmission power and channel gain as states. However, this approach results in a high dimensionality problem, as noted in [21]. The DQN algorithm employs deep neural networks to estimate the expected rewards of actions and can be computationally intensive due to the complexity of training deep neural networks. This complexity arises from the need to address large neural networks with many layers and parameters, which requires significant computational resources for training and updating. To address these challenges, the actor-critic (A2C) algorithm is employed in [15]. The A2C algorithm improves upon the limitation of the DQN algorithm by using separate networks for the actor, which learns the policy, and the critic, which estimates the value function. The authors in [16] consider continuous action space for the DL max-min power control problem in cell-free (CF) massive MIMO systems and propose a deep deterministic policy gradient (DDPG) method. Furthermore, the objective function is maximized considering max-min fairness [22] and the maximum product signal-to-interference-plus-noise ratio (SINR) [23] methods. However, single agent-based PA strategies in DRL algorithms require extensive training to determine optimal policies in case of optimization in large-scale complex environments.

To deal with training overhead in single agents-based DRL technique for PA in dynamic wireless networks, the multi-agent reinforcement learning (MARL) approach was adopted with enhanced training strategy, scalable distributed learning, and execution in [17] and [18]. The authors in [17] introduce a multi-agent DQN-based PA technique to maximize the sum rate in multi-cell networks. The sum rate is maximized using local agents with uniform target parameters while the global network updates the replay buffer gathered by these local agents. Furthermore, the authors demonstrate that the multi-agent DQN outperforms the single DQN in model training efficiency. Similarly, a multi-agent double DQN (DDQN)-based PA framework is proposed in [18] to maximize the capacity in multi-cell massive MIMO networks. The multi-agent DDQN model is split into sub-networks, i.e., the target Q-network and the evaluation Q-network, to avoid overestimating the Q-value in the DQN model. It is concluded that the proposed multi-agent DDQN provides improved convergence stability compared to the conventional DQN approach.

TABLE 1. Comparison with existing power allocation studies based on various reinforcement learning strategies.

Types	References	Network layouts	Reinforcement learning algorithms	Optimization goals
DRL	[13]	Multi-cell	DQN with replay memory	Sum rate maximization
	[14]	Small cell	DQN with replay memory	Sum rate maximization
	[15]	Single cell	A2C without replay memory	EE maximization
	[16]	Cell-free massive MIMO	DDPG with replay memory	Solve the max-min power control problems
MARL	[17]	Mobile ad-hoc network	Multi-agent DQN with replay memory	Sum rate maximization
	[18]	Multi-cell massive MIMO	Multi-agent DDQN with replay memory	Capacity maximization
MORL	[19]	Cell-free massive MIMO	TD3 with replay memory and scalarization	Joint optimization of sum rate and fairness
	[20]	Single cell	A2C with replay memory and PFA policy	Joint optimization of capacity and fairness
	Our work	Multi-cell massive MIMO	Proposed MO-A3Cs with preference updates	Solve the trade-off between SE and EE

Recently, the emerging MORL algorithm has been used to solve MOO problems in the massive CF MIMO networks. The authors in [19] use a reward vector, defined as the sum rate and user fairness. In addition, to solve the MOO problem by transforming the problem into a single objective optimization (SOO). Moreover, the twin-delayed DDPG (TD3) algorithm with a replay buffer effectively maximizes the sum rate and fairness. These replay buffer-based training strategies can enhance sampling diversity and efficiency in CF massive MIMO networks. The authors in [20] harness the A2C algorithm combined with replay memory to design and learn a Pareto front approximation (PFA) policy [24], [25]. This proposed methodology simultaneously optimizes channel capacity and user fairness. However, the existing MORL model does undergo training through buffer memory-based training strategies. It relies on old data saved in the buffer with limited memory size and uses it for future training. Furthermore, instead of using weight adjustment among multiple objectives, interpolation preference weights are considered, which are scenario-limited. These training strategies can lead to sub-optimal policies in the case of massive MIMO networks.

To solve this problem, it is crucial to develop an advanced MORL algorithm designed to optimize transmit PA, thereby enhancing the overall SE and EE in massive MIMO networks. This paper introduces a novel MORL algorithm that leverages a MARL strategy for efficient training. This strategy enables interactions between each local agent and independent environments, leading to the acquisition of diverse and immediate experience data to train joint optimization policy. Furthermore, we implement a Bayesian rule-based preference weight updating mechanism that dynamically adjusts the weightings of multi-objectives, including SE and EE, informed by the trajectories collected from each local agent. These innovations ensure that our proposed MORL algorithm not only trains from a diversity of experience data but also improves both SE and EE in multi-cell massive MIMO networks.

The main contributions of the paper are as follows:

- We propose a PA technique based on the novel MORL algorithm for the DL multi-cell massive MIMO networks. The proposed MO-A3Cs algorithm utilizes MORL to optimize a trade-off between SE and EE in a massive MIMO network. MO-A3Cs follow Bayesian rule-based preference updating, the multi-objective advantage function, and the balanced-reward aggregation methods to solve the trade-off problem effectively. The proposed PA technique optimally allocates the transmission power in a massive MIMO network while ensuring an overall SE and EE balanced increase.
- We define a multi-objective MDP (MOMDP) for the proposed MO-A3Cs model comprising the state space, action space, and the extended reward vector. In addition, We provide the proposed MO-A3Cs model-based DL transmit PA strategies in multi-cell massive MIMO networks. This procedure offers insights into the MORL algorithm for optimizing trade-offs, a critical aspect of 5G networks and next-generation wireless communications.
- Extensive simulations are conducted to analyze the performance of the proposed MO-A3Cs for DL PA in multi-cell massive MIMO networks. Compared with other benchmark schemes, the proposed MO-A3Cs provide better performance regarding average SE and power consumption in the massive MIMO networks. Furthermore, the simulation results depict the effectiveness of the proposed MO-A3Cs in achieving a joint-optimized SE and EE.

The rest of the paper is organized as follows. Section II presents the system model for the DL multi-cell massive MIMO networks. Section III presents the background and problem formulation, while Section IV presents the proposed MO-A3Cs model for DL PA in multi-cell massive MIMO networks. The simulation setup and the detailed discussion related to simulation results are presented in Section V. Finally, the paper is summarized and concluded in Section VI.

II. SYSTEM MODEL

In this section, we present the network layout followed by the main system assumptions, SINR and SE, the network power consumption model, and an overview of the joint spectral-energy optimization problem.

A. NETWORK LAYOUT

A DL multi-cell massive MIMO network is considered with L number of cells as shown in Fig. 1. The BS is deployed at the center of each cell where j-th BS $\forall j \in \{1, 2, \ldots, L\}$ in the network is equipped with M number of antennas. The

UEs are assumed to be located randomly in the l-th cell [10]. Furthermore, we assume that each BS simultaneously serves a K number of UEs by sharing the same frequency band.

The channel matrices between the j-th BS and k-th UE located in l-th cell is denoted by $\mathbf{h}^l_{j,k} \in \mathbb{C}^M$ and can be expressed as

$$\mathbf{h}^l_{j,k} \sim \mathcal{N}_{\mathbb{C}}\left(0, \mathbf{R}^l_{j,k}\right), \tag{1}$$

where $\mathbb{C}^M$ and $\mathbf{R}^l_{j,k} \in \mathbb{C}^{M\times M}$ denote the complex-valued vector space of dimension M and the spatial correlation matrix, respectively. Furthermore, we assume the BSs and UEs are perfectly synchronized and operate under the time division duplex (TDD) protocol. Before performing DL the channel at the BS. The UEs reuse the pilot signal in the cell, and the reuse factor $\tau_p = K$ is employed to reduce interference in the adjacent cells [26]. Based on this assumption, we utilize the minimum mean-square error (MMSE) estimation method at the BS to effectively estimate the imperfect channel condition corrupted by the interference and noise in the network [27]. The estimated channel between the j-th BS and k-th UE computed from the uplink pilot signal ρ^{UL}, is denoted by $\hat{\mathbf{h}}^l_{j,k}$. The MMSE-based estimated channel is given by

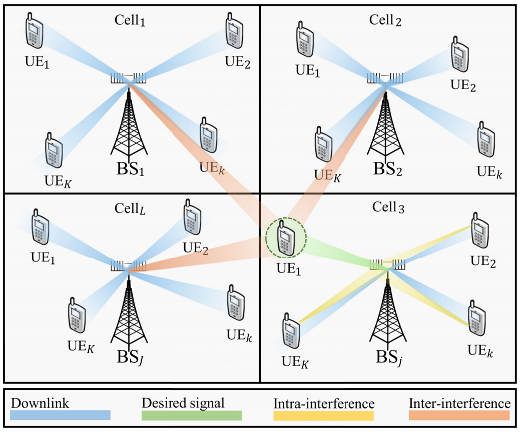


FIGURE 1. Illustration of the DL multi-cell massive MIMO networks.

$$\hat{\mathbf{h}}^l_{j,k} = \mathbf{R}^l_{j,k}\mathbf{Q}^{-1}_{j,k}\left(\sum_{l\neq j}^{L}\mathbf{h}^l_{j,k} + \frac{1}{\tau_p}\frac{\sigma^2}{\rho^{\mathrm{UL}}}\mathbf{n}_{j,k}\right), \tag{2}$$

where $\mathbf{Q}_{j,k} = \sum_{l\neq j}^{L}\mathbf{R}^l_{j,k} + \frac{1}{\rho^{\mathrm{UL}}}\mathbf{I}_M$, $\mathbf{I}_M$ denotes the identity matrix, and σ^2 is the noise variance. The noise added by the system is represented as $\frac{1}{\tau_p}\frac{\sigma^2}{\rho^{\mathrm{UL}}}\mathbf{n}_{j,k}$. Based on the MMSE technique, the channel estimation is performed by minimizing the estimation error between the actual and estimated channels and is expressed as $\mathbf{e}^l_{j,k} = \mathbf{h}^l_{j,k} - \hat{\mathbf{h}}^l_{j,k}$.

B. RECEIVED SINR AND SPECTRAL EFFICIENCY

The DL signal received at k-th UE contains the desired signal transmitted from the j-th BS, inter-cell and intra-cell interference, and the system-added noise. The DL signal received at the k-th UE from the BS located in the j-th cell can be expressed as [3] and [28]

$$y_{j,k} = \underbrace{z'_{j,k}s_{j,k}}_{\text{Desired signal}} + \underbrace{\sum_{l=1,l\neq j}^{L}\sum_{i=1}^{K} z'_{l,i}s_{l,i}}_{\text{Inter-cell interference}} + \underbrace{\sum_{i'=1,i'\neq k}^{L} z'_{l,i'}s_{l,i'}}_{\text{Intra-cell interference}} + \underbrace{n_{j,k}}_{\text{Noise}}, \tag{3}$$

where $s_{j,k}$ denote the transmitted signal from the j-th BS to each k-th UE, $z'_{j,k} = \mathbb{E}\{z^{\mathrm{H}}_{j,k}\hat{\mathbf{h}}^j_{j,k}\}$ denote the regularized zero-forcing (RZF) precoding vector [29], and $z'_{j,k}s_{j,k}$ represents the actual transmitted DL signal to k-th UE.

The received SINR at the k-th UE from the j-th BS is written as

$$\lambda_{j,k} = \frac{p_{j,k}\alpha_{j,k}}{\sum_{l=1}^{L}\sum_{i=1}^{K}p_{l,i}\beta_{l,i} + \sigma^2}, \tag{4}$$

where $p_{j,k}$, $\alpha_{j,k}$, and $\beta_{l,i}$ denote the DL transmit power, the channel gain between the j-th BS and the k-th UE, and the interference signal power received at the k-th user from the l-th BS. The channel gain is given as $\alpha_{j,k} = \left|\mathbb{E}\{z^{\mathrm{H}}_{j,k}\hat{\mathbf{h}}^j_{j,k}\}\right|^2$ while the interference term is given by

$$\beta_{l,i} = \begin{cases} \mathbb{E}\left\{\left|z^{\mathrm{H}}_{j,k}\hat{\mathbf{h}}^l_{j,k}\right|^2\right\}, & \text{if } (l,i) \neq (j,k) \\ \mathbb{E}\left\{\left|z^{\mathrm{H}}_{l,i}\hat{\mathbf{h}}^l_{j,k}\right|^2\right\} - \left|\mathbb{E}\left\{z^{\mathrm{H}}_{l,i}\hat{\mathbf{h}}^j_{j,k}\right\}\right|^2, & \text{if } (l,i) = (j,k) \end{cases} \tag{5}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose.

C. JOINT SPECTRAL-ENERGY OPTIMIZATION

According to Shannon's theorem, the channel capacity is defined as the maximum amount of information that can be transferred over a channel [10]. The achievable channel capacity of the established link between the k-th UE and the j-th BS is expressed as

$$C_{j,k} = \frac{\tau_d}{\tau_c}\log_2(1 + \lambda_{j,k}), \tag{6}$$

where τ_d and τ_c denote the number of samples used for DL data transmission and per coherence block, respectively.

1) DL SPECTRAL EFFICIENCY

The DL SE is defined as the total achievable data rate over the available bandwidth in massive MIMO networks and is measured in bits per second per Hertz (b/sec/Hertz). Based on the received SINR in (4) and the achievable channel capacity in (6), the total achievable SE in multi-cell massive MIMO networks can be formulated as [11]

$$\mathrm{SE}_{\mathrm{DL}} = \sum_{j=1}^{L}\sum_{k=1}^{K}C_{j,k}. \tag{7}$$

2) NETWORK POWER CONSUMPTION MODEL

The total power consumption in the DL multi-cell massive MIMO networks is the sum of the effective transmit power $p_{j,k}$ allocated based on the PA technique and the circuit power consumption P_{CR}. The total consumed power can be mathematically expressed as [3]

$$P_{\mathrm{total}} = \underbrace{\sum_{j=1}^{L}\sum_{k=1}^{K} p_{j,k}}_{\text{Effective transmit power}} + \underbrace{\sum_{j=1}^{L} P_{\mathrm{CR}}}_{\text{Circuit power}}. \tag{8}$$

The circuit power consumption of each BS in the massive MIMO network comprises the constant power consumed at BS denoted by P_{FIX} and the constant power incurred during the signal processing denoted by P_{SP}. Therefore, the total circuit power consumption of a BS can be expressed as

$$P_{\mathrm{CR}} = \underbrace{P_{\mathrm{CH}} + P_{\mathrm{CE}} + P_{\mathrm{BH}} + P_{\mathrm{ED}}}_{\text{Operating circuit power}} + \underbrace{P_{\mathrm{FIX}} + P_{\mathrm{SP}}}_{\text{Fixed circuit power}}. \tag{9}$$

A large fraction of the power consumed in the network comprises the power consumed at the BS [30]. The power consumption of the BS comprises circuit powers required in operations such as the number of transmit antennas, channel estimation, and encoding and decoding [31]. In particular, the circuit power due to the transceiver chain, which accounts for the most power consumption, includes components such as filters, mixers, digital-to-analog converters (DAC), and analog-to-digital converters (ADC). The power consumption of the transceiver chain component can be written as

$$P_{\mathrm{CH}} = \underbrace{Mp_{\mathrm{BS}} + p_{\mathrm{LO}}}_{\text{BS circuit components}} + \underbrace{Kp_{\mathrm{UE}}}_{\text{UE circuit components}}, \tag{10}$$

where p_{BS}, p_{LO}, and p_{UE} denote the transmission power of a single BS antenna, the local oscillator (LO), and the circuit power coefficient of the UE, respectively. From (10), the power consumption of the BS is proportional to the number of antennas. Furthermore, the power consumed during the channel estimation at the BS for each coherent block is also taken into consideration [2]. The power consumption in terms of the channel estimation can be calculated as

$$P_{\mathrm{CE}} = \frac{3B}{\tau_c L_{\mathrm{BS}}} K \underbrace{M\tau_p + M^2}_{\text{MMSE}}, \tag{11}$$

where B and L_{BS} denote the bandwidth and the computational efficiency of the BS, respectively [32].

The circuit power consumed in the backhaul during the uplink and DL data transmission can be expressed as

$$P_{\mathrm{BH}} = p_{\mathrm{BT}}\mathrm{TP}, \tag{12}$$

where p_{BT} denotes the backhaul traffic power and TP represents the achievable throughput within a cell. The value of TP is calculated as $B\sum_{k=1}^{K} C_{j,k}$. Similarly, the circuit power consumed in channel encoding and decoding is denoted by P_{ED} and is given by

$$P_{\mathrm{ED}} = (p_{\mathrm{ENC}} + p_{\mathrm{DEC}})\,\mathrm{TP}, \tag{13}$$

where p_{DEC} and p_{DEC} represent the power consumption coefficients incurred during the encoding and decoding processes, respectively.

3) ENERGY EFFICIENCY

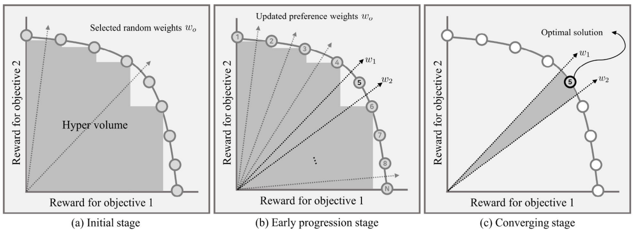
The EE of the DL massive MIMO network is the ratio of SE to the total power consumption and can be formulated as

$$\mathrm{EE}_{\mathrm{DL}} = \frac{\mathrm{SE}_{\mathrm{DL}}}{P_{\mathrm{total}}}. \tag{14}$$

To evaluate the joint optimization in the multi-cell massive MIMO networks, simultaneously SE and EE must be optimized. Let us define a joint objective function of SE and EE by $\mathcal{F}\left(\mathrm{SE}_{\mathrm{DL}}, \mathrm{EE}_{\mathrm{DL}}\right)$. Thus, the joint optimization problem can be formulated as [5]

$$\begin{aligned} &\max_{p_{j,k}} \mathcal{F}\left(\mathrm{SE}_{\mathrm{DL}}, \mathrm{EE}_{\mathrm{DL}}\right) \\ &\text{s.t. } 0 < p_{j,k} \leq P_{\max}, \ \forall j, k, \end{aligned} \tag{15}$$

where $P_{\max}$ denote the maximum transmit power. The transmit power ($p_{j,k} \geq 0$) that affects both SE and EE is defined as a constraint and is required in the joint optimization problem [33]. The joint optimization problem in (15) is classified as multi-objective non-convex and NP-hard and requires high computations [13], [18].



(a) Initial stage (b) Early progression stage (c) Converging stage

FIGURE 2. Illustration of policy convergence in the MORL algorithms driven by preference weights for solving the MOO problem.

III. MORL ALGORITHMS AND TRANSFORMATION TO SOO

In this section, we first briefly present the background of the MORL techniques for MOO problems. Then, the detailed description of the MOMDP is presented to transform the MOO problem into the SOO problem, which is later utilized in allocating power using the proposed MO-A3Cs model.

A. MORL TECHNIQUE FOR MOO PROBLEM

Numerous studies have adopted and validated the DRL algorithms for PA to achieve enhanced performance in wireless networks. However, it is challenging to achieve better performance complexity trade-offs in the emerging MORL algorithm [19], [20]. Compared to DRL, the MORL algorithms utilize multiple rewards in the form of reward vectors to maximize multiple objectives. To effectively tackle these reward vectors, the representative MORL algorithm uses the PFA strategy [24], [25] and MOO transformation to the SOO problem [34], [35]. The PFA approach utilizes the reward vectors of the selected actions to determine the optimal point among multiple objectives based on Pareto dominance and Pareto fronts. The collected data from agent-environment interactions are used to construct a Pareto set and derive the required solution for the MOO problem. This approach requires a considerable buffer memory used to generate the Pareto set. Moreover, the Pareto fronts [36] used to determine the optimal solution require significant training time in large-scale environments such as massive MIMO systems [37].

On the other hand, the transformation approaches that transform MOO problems into SOO problems employ strategies such as weighted sum [34] and constraints [35] and preference-driven approaches. Fig. 2 illustrates the process of determining the optimal policy for solving MOO problems using a MORL algorithm that uses preference weights to determine the optimal solution. In the initial stage of the MORL algorithm, a hypervolume is generated between multiple objectives through interactions between the agent and the environment, as shown in Fig. 2(a). In addition, Fig. 2(b) and (c) present the illustration of weight updates and the selection of optimal points for multiple objectives by using the relative priorities ω_1 and ω_2. This transformation strategy may achieve faster convergence compared to the PFA approach. However, it can lead to limited training efficiency due to a bias towards specific objectives and the potential for converging to sub-optimal solutions depending on the preference weight settings [38].

To this end, various methods have been suggested to effectively determine preference weights for the MORL algorithm. These methods include the utilization of uniform weights [39], random weights [40], and dynamic weights [41]. The uniform and random approaches have limitations in that the convergence of the MORL model must be verified through various experiments to train the optimal points for the MOO solution. On the other hand, the dynamic weight approach allows for dynamically determined weights to optimize the policy designed to solve the MOO problem. However, this method requires additional buffer memory to update the weights for each objective.

Therefore, we propose a novel MORL algorithm to solve multi-objective functions and trade-off problems between SE and EE. The proposed model employs a MOMDP framework and integrates a Bayesian rule-based technique for updating preference weights, a multi-objective advantage function, and a balanced reward aggregation method. The multi-objective advantage function allows for the individual evaluation of action value for each objective. Moreover, the balanced-reward aggregation method aggregates rewards considering each preference weight, ensuring a more efficient approach to action selection by the agents.

B. MOMDP-BASED TRANSFORMATION OF MOO TO SOO

The MOMDP is an extension of the MDP and deals with multiple rewards in the form of a reward vector. In addition, the MOMDP can be defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}, \mathcal{A}, \mathcal{P}(s'|s, a)$ denotes the state space, action space, and the transition probability of taking action a for state transitions from s to s', respectively. The reward vector $\mathcal{R}$ consists of the respective objective rewards for SE and EE. Thus, $\mathcal{R}$ can be expressed as $\{\mathcal{R}_o| \, \forall o \in \{1, 2 \ldots, O\}\}$, where O represents the total number of objectives. Similarly, the discount factor γ, which determines how much the agent considers long-term rewards, is defined as $\gamma \in [0, 1)$. Furthermore, we employ preference weights $\{\omega_o| \, \forall o \in \{1, 2 \ldots, O\}\}$ indicate the relative priority of each objective [42], [43].

The action space $\mathcal{A}$ consists of feasible DL transmission powers between all BSs and UEs. However, defining the action space as the set of all possible transmission powers in a multi-cell massive MIMO network, the dimensionality issue arises [21], [44].

To this end, we utilize a discretization strategy using quantization [45] of transmit power between $P_{\min}$ and $P_{\max}$ with a specific quantization level to select action a_t. The discretized action space based on quantization can be expressed as

$$\mathcal{A} = \left\{ 0, P_{\min}, P_{\min} \left(\frac{P_{\max}}{P_{\min}} \right)^{\frac{1}{|Q|-2}}, \ldots, P_{\max} \right\}, \tag{16}$$

where $|Q|$ denotes the quantization level, which indicates the degree at which the transmission power range between $P_{\min}$ and $P_{\max}$ be divided into discrete values. This discretization approach allows an increase in the power at each $|Q|$ level and effectively generates a variety of power action space between $P_{\min}$ and $P_{\max}$.

The states s_t, which facilitate the observation of various features related to the problem in (15), can be defined as

$$s_t = \{\alpha^t_{j,k}, C^t_{j,k}, a_t\}, \quad \forall j, k, \tag{17}$$

where $\alpha^t_{j,k}$, $C^t_{j,k}$, and a_t denote the channel gain, achievable channel capacity, and selected action at the time step t, respectively. These state s_t are utilized by an agent to efficiently observe the SE and EE while interacting with the DL multi-cell massive MIMO network.

The joint optimization problem in (15) is transformed and can be rewritten as

$$\begin{aligned} &\max_{a_t} \; f_{\omega_1}(\mathrm{SE}_{\mathrm{DL}}(a_t)) + f_{\omega_2}(\mathrm{EE}_{\mathrm{DL}}(a_t)), \\ &\text{s.t. } \omega_1 + \omega_2 = 1, \; \omega_1, \omega_2 \geq 0, \end{aligned} \tag{18}$$

where $f\omega_o(\mathcal{R}_o) = \omega_o \times \mathcal{R}_o$ represents the weighted sum [34]-based scalarization function for the reward vector. In addition, the ω_1 and ω_2 denote the preference weights that indicate the relative priorities of SE and EE, respectively.

Finally, the immediate reward vector obtained through the interaction between the agent and environment in a massive MIMO network at a certain time instant can be expressed as

$$r_t = \left[\overline{\mathrm{SE}}_{\mathrm{DL}}(t), \overline{\mathrm{EE}}_{\mathrm{DL}}(t) \right], \tag{19}$$

where $\overline{\mathrm{SE}}_{\mathrm{DL}}(t) = \frac{\mathrm{SE}_{\mathrm{DL}}}{L}(t)$, $\overline{\mathrm{EE}}_{\mathrm{DL}}(t) = \frac{\mathrm{EE}_{\mathrm{DL}}}{L}(t)$ denote average SE and EE, respectively. In general, the total value

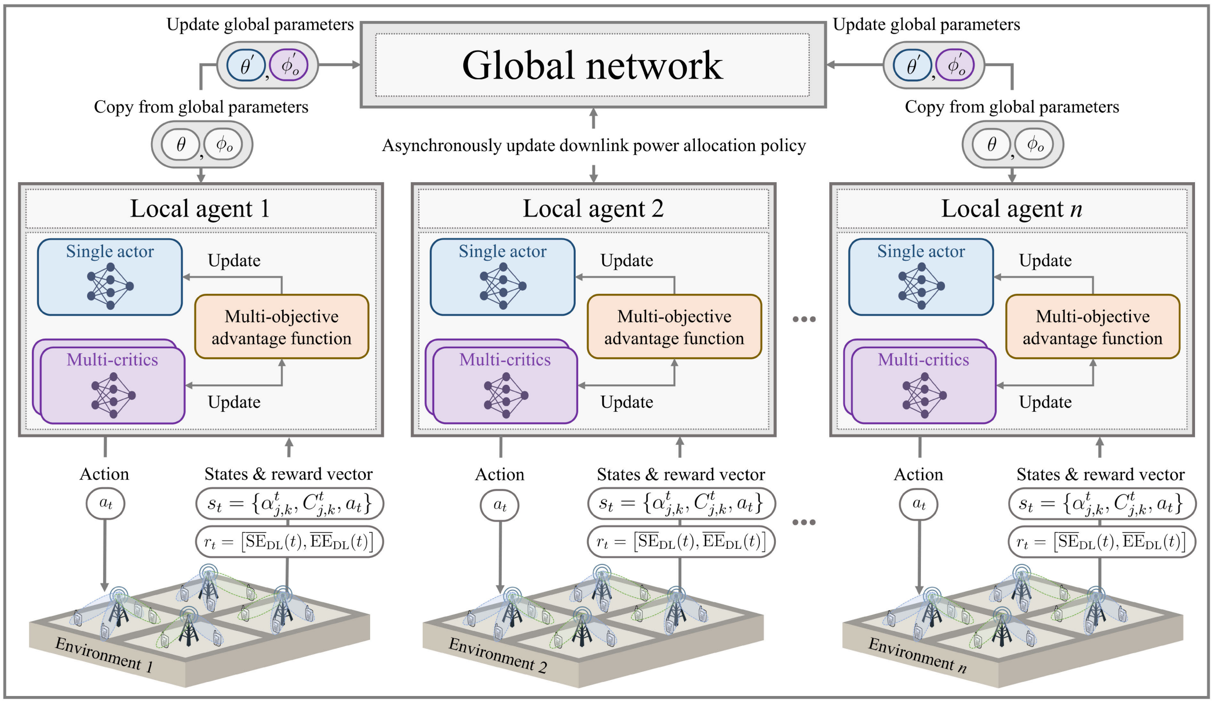


FIGURE 3. The proposed MO-A3Cs based transmit PA framework for joint SE and EE optimization in the DL multi-cell massive MIMO networks.

of SE is higher than the EE, which can lead to convergence instability and extend the training duration. Thus, we used the average SE and EE to reduce the variability of rewards and ensure smoother convergence during the model training.

IV. PROPOSED MO-A3Cs TECHNIQUE FOR POWER ALLOCATION IN MASSIVE MIMO NETWORKS

In this section, we introduce the proposed MO-A3Cs model-based PA framework. This framework includes a Bayesian rule-based preference update, a multi-objective function with a reward aggregation method, and optimization of each single-actor and multi-critic network. The proposed MO-A3Cs model uses the multi-critic network to consider multi-objectives and estimate the expected value for each objective. The single actor determines the optimal power value for different objectives by aggregating the predicted values from each critic model. The proposed DL PA framework based on MO-A3Cs along with the training strategy is illustrated in Fig. 3. During the initial training process, each local agent copies key parameters from the global network with initialized preference weights. The experiences and trajectories obtained from these independent interactions ensure a diversity of training data for the global network. Unlike the conventional MORL-based MOO approach, each local agent i.e., the single-actor multi-critic networks asynchronously updates the global network and provides a range of experiences without the need for a replay buffer. Our proposed approach is inspired by the fundamental asynchronous advantage actor-critic (A3C) model, which consists of an actor network that selects actions and a critic network that evaluates the chosen actions.

In the basic A3C model, the actor and critic networks interact with each other and decide whether to take a specific action a_t from the available action space $\mathcal{A}$ at a particular s_t. Conversely, the critic evaluates the selected action a_t by the actor using the value function $V_\phi(s_t)$. The update process of the actor network is given as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{\mathrm{T}} (\mathcal{R}_t - V_\phi(s_t)) \times \nabla \log \pi_\theta(a_t|s_t), \tag{20}$$

where θ, η, t, and T denote the policy parameters of the actor, learning rate, time step, and maximum number of episodes, respectively. Moreover, the $\mathcal{R}_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$ represents the accumulated reward computed based on the discount factor γ, $V_\phi(s_t)$ is the value function for state s_t, and $\nabla \log \pi_\theta(a_t|s_t)$ denotes the gradient of the actor network. The network aims to maximize the expected reward by utilizing the advantage function, which is based on the difference between $\mathcal{R}_t$ and $V_\phi(s_t)$ for a given state s_t.

The critic aims to minimize the error between the accumulated reward from the actor and its predicted value. Hence, the value of the selected action a_t is evaluated based on

$$\phi \leftarrow \phi + \eta \frac{\partial(\mathcal{R}_t - V_\phi(s_t))^2}{\partial \phi}, \tag{21}$$

where ϕ denotes the critic network parameters and $\frac{\partial(\mathcal{R}_t - V_\phi(s_t))^2}{\partial \phi}$ is the gradient of the squared error.

The MO-A3Cs model integrates the extension of the A3C [46] along with the proposed Bayesian rule-based preference weight update, multi-objective advantage function, and balanced-reward aggregation method presented next.

A. BAYESIAN RULE-BASED PREFERENCE UPDATES

The weight update strategy ensures diverse experiences for global network training without relying on potentially biased experiences from initially collected data and helps in the joint optimization of multi-objectives. In the MORL algorithm, the preference weights are typically assumed to be uniformly and randomly initialized [41] and each objective is assigned an equal weight as follows

$$\omega_o = \frac{1}{O}, \quad \forall o = 1, 2, \ldots, O. \tag{22}$$

However, the equal preference weight initialization in solving the SE-EE trade-off problem leads to more frequent updates of the weights due to the interaction of the local agent with the distinct and independent multi-cell massive MIMO environment. Hence more training overhead is needed to find the joint optimization policy. To deal with this issue, we proposed an adaptive update mechanism using Bayesian rules with random preference initialization. The preference weight is initialized from uniform distribution as $\omega_o \sim \mathcal{U}(0, 1)$. To quantify the relative priority of the objectives, the prior probability given a particular trajectory is given by

$$p(x_o|\psi_0) = \frac{\omega_o}{\sum_{o=1}^{O} \omega_o} \forall x_o \in \{\text{SE, EE}\}, \tag{23}$$

where x_o represents the o-th objective function, and ψ_0 denotes the corresponding initial trajectory capturing the sequence of interactions between the agent and the environment. The initial trajectory at time step t is $\psi_t = (s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t, a_t, r_t)$ and prior probability derive from weight normalization process facilitates the comparison of the relative importance of each objective. The likelihood function for each objective can be expressed as

$$p(\psi_t|x_o) = \prod_{t=1}^{\mathrm{T}} p(s_t, a_t|x_o), \tag{24}$$

where the likelihood $p(\psi_t|x_o)$ represents the probability of observing the trajectory ψ_t, which consists of states s_t, actions a_t, given the objective function x_o. This metric help the agent to asses whether the trajectory ψ_t of the current state s_t sufficiently aligns with which objective function x_o. As the samples in the trajectory ψ_t change at step t, the agent adaptively adjusts the weight ω_o using the Bayesian rule and is given by [47] and [48]

$$\omega_o = p(x_o|\psi_t) = \frac{p(\psi_t|x_o)p(x_o)}{\sum_{o=1}^{O} p(\psi_t|x_o)p(x_o)}, \tag{25}$$

where the prior probability and the likelihood function are given in (23) and (24), respectively. The proposed preference weight update mechanism overcomes shortcomings of the conventional interpolation-based and buffer memory-based dynamic weights update approaches.

B. MULTI-OBJECTIVE ADVANTAGE FUNCTION AND BALANCED-REWARD AGGREGATION

The fundamental A3C model predicts the $V_\phi(s_t)$ from the current state s_t in the critic network and employs the advantage function to evaluate and update the actor-critic network. This single objective advantage function measures the difference between the $\mathcal{R}_t$ and $V_\phi(s_t)$ for a specific action a_t taken on the current state s_t. It is beneficial to evaluate the value of the chosen action a_t [46]. However, such an advantage function is optimized for training a single objective and is ineffective for solving MOO problems. To solve the MOO problem, we extend the single objective advantage function to a multi-objective advantage function. Let O be the number of objectives, then the multi-objective advantage function can be expressed as

$$G_o = \mathcal{R}_t^o - V_{\phi_o}(s_t) = \sum_{i=t}^{\infty} \gamma^{i-t} \delta_i^o, \ o \in \{1, 2, \ldots, O\}, \tag{26}$$

where $\mathcal{R}_t^o$, $V_{\phi_o}(s_t)$ indicate the multi-objective accumulated reward and value function, respectively. The extended temporal difference error (TD error) [49] is represented by $\delta_t^o = r_{t+1}^o + \gamma V_{\phi_o}(s_{t+1}) - V_{\phi_o}(s_t)$. Using G_o, the value of actions is independently assessed for each objective i.e., SE and EE. The single actor network is updated according to the proposed balanced-reward aggregation method utilizing G_o.

In the MORL algorithm, reward aggregation follows the summation of multi-objective rewards based on the scalarized function $f\omega_o$ and their relative priorities [50]. Generally, where the trade-off between objectives is not considered, all the objective rewards are summed in a single reward function and can be expressed as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{\mathrm{T}} \left(\sum_{o=1}^{O} G_o \right) \times \nabla_\theta \log \pi_\theta(a_t|s_t). \tag{27}$$

For SE and EE as objectives, the single-objective DRL model does not utilize preference weights ω_o, making it challenging to train joint optimization policies. To incorporate ω_o, a combined-reward aggregation is applied to MORL algorithms [51] to reflect the preference weights in each objective, and is written as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{\mathrm{T}} \left(\sum_{o=1}^{O} \tilde{G_o} \right) \times \nabla_\theta \log \pi_\theta(a_t|s_t), \tag{28}$$

where $\tilde{G_o} = f\omega_o(\mathcal{R}_t^o) - V_{\phi_o}(s_t)$ denotes the aggregation of the accumulated reward $\mathcal{R}_t^o$ based on the scalarized function $f\omega_o$ applied in the multi-objective advantage function. Unlike (27), this approach allows the consideration of priorities for each objective and provides a more effective way to address SE-EE trade-off problems. However, this method does not incorporate the ω_o for each objective in the value function $V_{\phi_o}(s_t)$. This means that the agent might select biased actions towards a specific objective and not consider the preferences for both SE and EE due to asymmetric updates to the value function $V_{\phi_o}(s_t)$ in each critic network. For instance, if the value function for SE is considered more important than that for the function of EE, and without preference weights ω_o, the agent focuses on prioritizing actions that maximize SE, which may potentially lead to EE degradation. To cope with this, we consider a balanced-reward aggregation given by

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{\mathrm{T}} \left(\sum_{o=1}^{O} f\omega_o(G_o) \right) \times \nabla_\theta \log \pi_\theta(a_t|s_t). \tag{29}$$

This method applies preference weights ω_o to $\mathcal{R}_t^o$ as well as $V_{\phi_o}(s_t)$ which leads to more efficient training to determine the optimal PA policy and jointly optimizing the SE and EE.

C. OPTIMIZATION OF MO-A3Cs UPDATES

This section presents the update loss functions for the single-actor and multi-critic networks for the proposed MO-A3Cs model. The update for the single actor network to decide the optimal action a_t in the MO-A3Cs model is done by minimizing the loss function given by

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{o=1}^{O} f\omega_o(G_o) \right) \times \log \pi_\theta(a_t|s_t), \tag{30}$$

where N, i, and $\pi_\theta(a_t|s_t)$ denote the number of trajectories, trajectory index, and the probability of selected a_t for state s_t according to policy π_θ, respectively.

Algorithm 1 The Proposed MO-A3Cs Model Training Procedure for Each Single-Actor Multi-Critics Thread

```
Initialize global parameters $\theta$ and $\phi_o$, $o \in \{1, 2, \ldots, O\}$.
Initialize global shared counter $\mathrm{T} = 0$.
Initialize local thread step counter $t \leftarrow 1$.
Initialize random preference weights $\omega_o$ using (23).
while $\mathrm{T} < \mathrm{T}_{\max}$ do
    Reset gradients: $d\theta \leftarrow 0$ and $d\phi_o \leftarrow 0$.
    Synchronize specific parameters: $\theta' = \theta$ and $\phi'_o = \phi_o$.
    $t_{\text{start}} = t$.
    Get state $s_t$ extracted from massive MIMO networks.
    repeat
        Perform action $a_t$ according to policy $\pi(a_t|s_t; \theta')$.
        Collect reward vector $r_t^o$ and update state to $s_{t+1}$.
        $t \leftarrow t + 1$.
        $\mathrm{T} \leftarrow \mathrm{T} + 1$.
    until terminal $s_t$ or $t - t_{\text{start}} == t_{\max}$;
    for $i \in \{t - 1, \ldots, t_{\text{start}}\}$ do
        Update cumulative rewards as $\mathcal{R}_i^o \leftarrow \mathcal{R}_i^o + \gamma \mathcal{R}_{i+1}^o$.
        Update the weights $\omega_o$ with the Bayesian rule (25).
        Compute multi-objective advantage function by (26).
        Apply the balanced-reward aggregation using (29).
        for $o \in \{1, 2, \ldots, O\}$ do
            Update multi-critic networks for $\phi'_o$ using (31).
        Update single actor network for $\theta'$ using (30).
    Asynchronous global update of $\theta$ and $\phi_o$ with $\theta'$ and $\phi'_o$.
```

The structure of the MO-A3Cs model has a separate critic network for SE and EE which leads to a more accurate estimation of the value function $V_{\phi_o}(s_t)$ for each objective. The loss function computes the difference between the expected value function and the actual reward using the multi-objective advantage function and the balanced-reward aggregation methods. Hence, it updates the action policy by considering the preference weights ω_o for each objective. The update loss function of the multi-critics can be written as

$$\mathcal{L}(\phi_o) = \frac{1}{N} \sum_{i=1}^{N} (V_{\phi_o}(s_t) - \mathcal{R}_t^o)^2, \tag{31}$$

where ϕ_o represents the parameters of the critic network for the o-th objective. Each critic network is updated independently for each objective and the value function estimation of one objective does not influence the other and vice versa. This extended multi-critic enables the estimation of the optimal value function $V_{\phi_o}(s_t)$ for each objective and facilitates a more appropriate balance between the trade-off of two objectives i.e., SE and EE.

In order to avoid premature or early convergence to sub-optimal solutions and enhance long-term convergence performance [52], action distribution entropy is utilized to encourage agents to select and explore various actions in a multi-cell massive MIMO network environment. The action distribution entropy is expressed as

$$H(\pi_\theta) = -\sum_{\mathcal{A}} \pi_\theta(a_t|s_t) \log \pi_\theta(a_t|s_t). \tag{32}$$

The larger entropy value of $H(\pi_\theta)$ enables the agent to explore the environment and search the expanded action space to collect diverse trajectories which result in effective training of MO-A3Cs. The final loss function represented in terms of action distribution entropy is used in the proposed MO-A3Cs model and is rewritten as

$$\mathcal{L}_{\text{total}} = \mathcal{L}(\theta) + \sum_{o=1}^{O} \mathcal{L}(\phi_o) + \mu H(\pi_\theta), \tag{33}$$

where $\mathcal{L}(\theta)$ and $\mathcal{L}(\phi_o)$ denote the single actor loss function, and multi-critic loss function, respectively, while μ with value ranges between 0 and 1 is a weight used for regularizing the action distribution entropy, and is set to 0.001.

The training procedure of the MO-A3Cs model for each thread is given in Algorithm 1. The algorithm initializes by synchronizing the key parameters between the global network and local agents followed by the preference weight initialization. For each time instant of the local thread, random preference weights are assigned to each local agent based on a distinct environment. The collected trajectories from each agent are leveraged and the global network is updated asynchronously. This training strategy benefits from the independent evaluation of SE and EE by the multi-critic network, distinct from existing MORL algorithms. This evaluation directs the joint optimization policy updates through the proposed balanced reward aggregation function.

Ultimately, by integrating a MARL-based training strategy and the proposed innovative MORL algorithm.

V. SIMULATION RESULTS AND ANALYSES

In this section, we present the simulation setup, and the performance of the proposed MO-A3Cs-based PA in comparison with the other benchmarks in DL multi-cell massive MIMO networks.

A. SIMULATION PARAMETERS AND HYPERPARAMETERS OF MO-A3Cs ARCHITECTURE

In the simulation setup, we consider 16 square cells each with an area of 250 m × 250 m, and on a BS is deployed per cell. All the UEs are equipped with a single antenna and are randomly and uniformly distributed in each cell. The minimum distance between the BS and the UE is set to 25 m. The channel gain at a distance of 1 km is -148.1 dB, and the path loss exponent is set to 3.76. The noise power and noise figure of each BS are set to -94 dBm and 7 dBm, respectively. The parameters considered in the simulation setup are listed in Table 2.

The simulation setup is implemented in Google Colab. To enable the asynchronous training of multiple local agents at various instances of the multi-cell environments, a multiprocessing package is employed. Furthermore, the DL multi-cell massive MIMO network environment is constructed using the OpenAI Gym toolkit, which allows the definition of the MOMDP components and facilitates the design of custom reinforcement learning environments. The proposed MO-A3Cs model is implemented in Python version 3.10.12 and PyTorch version 2.1.0.

TABLE 2. System parameters of the DL multi-cell massive MIMO setup with the circuit power consumption model [3], [13].

System parameters	Network setup
Number of cells (L)	16
Number of UEs per cell (K)	[5, 10]
Number of transmit antennas (M)	[20, 100]
Pilot reuse factor (τ_p)	4
Coherence block length (τ_c)	200
Bandwidth (B)	20 MHz
Power for BS antennas (p_{BS})	0.4 W
Power for BS local oscillator (p_{LO})	0.2 W
Power per UE (p_{UE})	0.2 W
Power for backhaul traffic (p_{BT})	0.25 W/(Gbit/s)
Power for data encoding (p_{ENC})	0.1 W/(Gbit/s)
Power for data decoding (p_{DEC})	0.8 W/(Gbit/s)
BS computation efficiency (L_{BS})	75 Gflops/W
UL transmit power (p^{UL})	0.1 W
Fixed BS power (P_{FIX})	10 W
Fixed power of signal process (P_{SP})	0.1 W
Minimum transmission power ($P_{\min}$)	5 dBm
Maximum transmission power ($P_{\max}$)	38 dBm

The proposed MO-A3Cs model comprises four fully connected layers, including two hidden layers and an input and output layer. The state space size is utilized as an input to the first layer, and the size of both the first and second hidden layers is 128 and uses a ReLU activation function. The single actor network outputs the probability of possible action a_t given the state s_t using the softmax function and the two critic networks are used. The hyperparameters of each model used in the training and performance evaluation are given in Table 3.

TABLE 3. Hyperparameters of the utilized DRL and MORL models.

Hyperparameters	SE-DQN	EE-DQN	PQN	MO-A3Cs
Learning rate	0.001	0.001	0.001	0.001
Discount factor	0.98	0.98	0.98	0.98
Batch size	64	64	64	64
Update interval	10	10	10	N/A
Number of agents	N/A	N/A	N/A	16
Optimizer	Adam	Adam	Adam	Adam
Maximum training steps	100,000	100,000	100,000	100,000
Warm-up steps	10,000	10,000	30,000	N/A
Replay buffer size	50,000	50,000	100,000	N/A
Initial ϵ	1.0	1.0	1.0	N/A
Final ϵ	0.01	0.01	0.01	N/A
$\epsilon-$decay	0.995	0.995	0.995	N/A
Loss function	Huber	Huber	MSE	Eqn. (33)

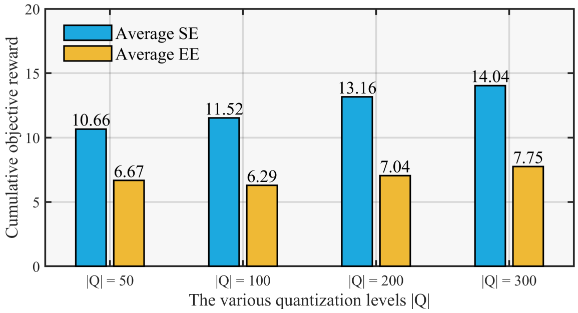


FIGURE 4. The average cumulative multi-objective reward achieved by the proposed MO-A3Cs model for various quantization levels $|Q|$.

B. BENCHMARK METHODS

The performance of the proposed PA scheme is compared with the existing benchmarks, including iterative algorithm-based PA methods, conventional DRL models, and MORL model-based PA techniques.

1) ALGORITHMIC APPROACHES

The considered benchmark algorithms are 1) an equal PA method [53] which allocates equal DL transmission power, and 2) the Dinkelbach algorithm-based PA [54]. Typically, the Dinkelbach algorithm addresses the fractional programming problem [55]. For Dinkelbach algorithm, the problem in (15) is transformed into $\max_x \frac{f(x)}{g(x)} \simeq \max_x \frac{SE}{EE}$, where $x = p_{j,k}$ indicates the DL transmission power. This fractional programming problem exhibits non-linear and non-convex characteristics. In the Dinkelbach algorithm, the problem is transformed into a sub-problem of the form $\max_x [f(x) - \kappa g(x)]$ and then solved in an iterative manner based on an arbitrary scalar value κ updated in each iteration. This value is updated until the optimal solution x^* is obtained at each stage. Moreover, the iterative process is conducted by gradually adjusting the transmission power from $\{0 < p_{j,k} \leq P_{\max}\}$. The Dinkelbach-based DL PA method optimizes the $p_{j,k}$ until the ratio of SE to EE in the transformed sub-problem becomes less than the predefined parameter ζ.

TABLE 4. Impact of quantization levels $|Q|$ on training efficiency the proposed MO-A3Cs in terms of total training time.

Evaluation metrics	$\|Q\| = 50$	$\|Q\| = 100$	$\|Q\| = 200$	$\|Q\| = 300$	$\|Q\| = 500$
Convergence training steps	19,144 / 100,000	28,421 / 100,000	42,345 / 100,000	47,451 / 100,000	66,753 / 100,000
Convergence time (minutes)	11.21	34.19	58.43	71.16	203.54
Total training time (minutes)	60.30	94.24	141.44	191.02	282.42

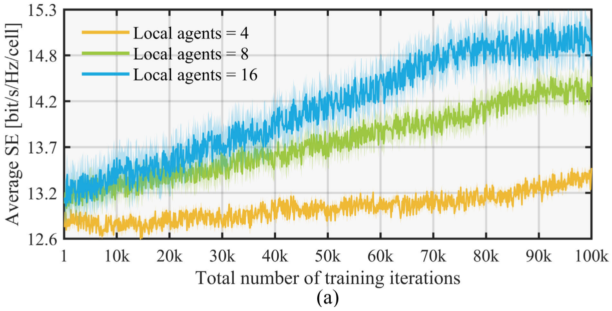


FIGURE 5. Training efficiency analysis of MO-A3Cs with varying local agents up to the maximum training iterations; (a) average SE, (b) average EE.

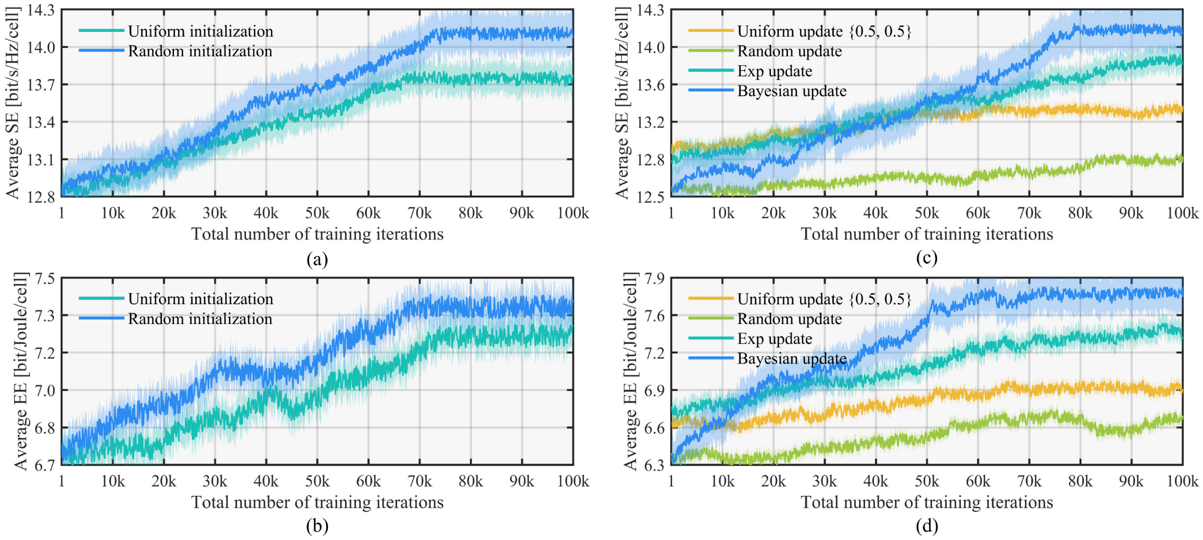


FIGURE 6. Training efficiency of MO-A3Cs as a function of weight initialization and preference update techniques: (a) Average SE reward and (b) Average EE reward for different initialization methods; (c) Average SE and (d) Average EE reward with various preference update mechanism.

Considering the computational complexity and accuracy of the Dinkelbach algorithm, the value ζ is set to 0.001.

2) REINFORCEMENT LEARNING APPROACHES

For a fair comparison, the DRL and MORL techniques such as SE-DQN [13], EE-DQN [56], and the PFA-based DQN (PQN) [57] are considered as a benchmark, where the SE-DQN and EE-DQN follow a single objective optimization and maximize SE and EE, respectively. However, the PQN model deals with the multi-objective problem to jointly optimize SE and EE. During training, these models use the ϵ-greedy algorithm, and an exploration and exploitation strategy [58]. The value of $\epsilon \in (0, 1)$ and the remaining hyperparameters for each model are set according to Table 3. In the DRL models, the state space of SE-DQN includes channel gain and DL user rate, while the EE-DQN model state space consists of power consumption and computed EE. Similarly, the same state space as PQN is considered by the proposed MO-A3Cs.

C. PERFORMANCE COMPARISON AND ANALYSES

In this section, we evaluate the performance of the proposed MO-A3Cs-based PA in terms of training efficiency, SE, EE, and the trade-off between SE and EE in the DL multi-cell massive MIMO network.

1) TRAINING PERFORMANCE EVALUATION

Fig. 4 illustrates the training performance of the MO-A3Cs model at various quantization levels $|Q|$. Fig. 4 shows that at $|Q| = 50$ the model achieved the lowest average cumulative

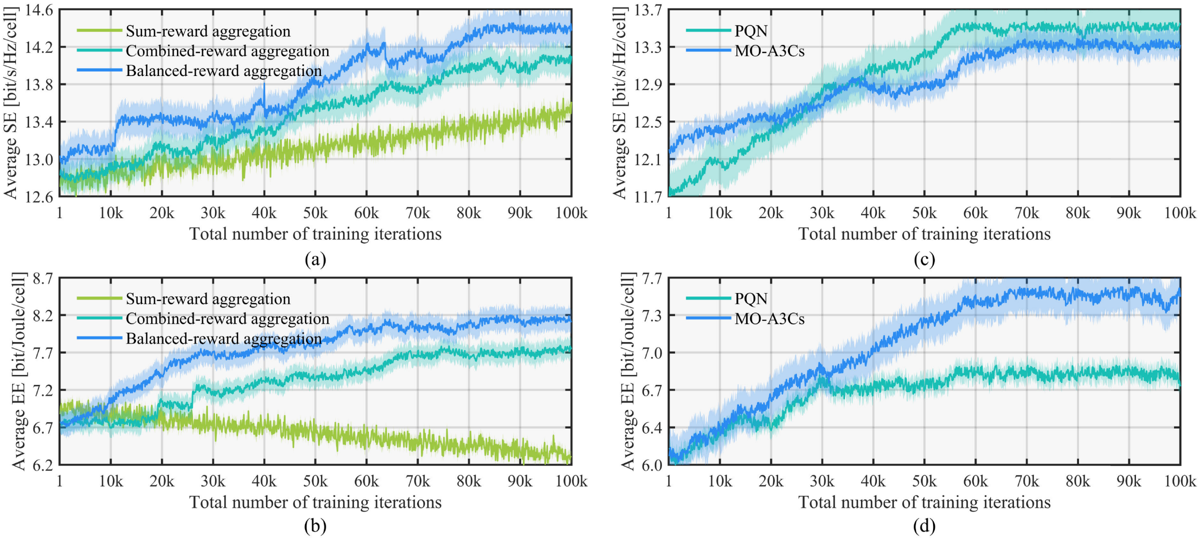


FIGURE 7. Training efficiency analysis; (a) the average SE reward for various aggregation methods, (b) the average EE reward for various aggregation methods, (c) the average SE reward comparison of the MORL models, (d) the average EE reward comparison of the MORL models.

rewards for SE and EE i.e., 10.66 and 6.67, respectively. However, $|Q|$ = 100 results in a 0.86 increase in SE, but a 0.38 decrease in EE compared to $|Q| = 50$. Further, increasing to $|Q|$ = 200 results in higher SE and EE achieving values of 13.16 and 7.04, respectively. Furthermore, within the range of utilized $|Q|$ levels, the proposed model achieves the maximum SE and EE rewards of 14.04 and 7.75, respectively, for $|Q| = 300$.

Table 4 shows the training efficiency of the MO-A3Cs model for different values of $|Q|$. It is clear from Table 4 that for $|Q|$ = 50, the fastest convergence is achieved in 11.21 minutes. In contrast, for $|Q| = 500$ the proposed model took 203.54 minutes to converge which indicates a significant increase in training time and duration. In addition, varying $|Q|$ from 200 to 300 requires 5,106 more steps, and 19,302 more steps are required for convergence when changing $|Q|$ from 300 to 500. The experiment demonstrates that the most balanced setting between performance and training complexity for the proposed MO-A3Cs model is setting $|Q|$ = 300. Table 4 concludes that using high quantization levels $|Q|$ can lead to exponential increases in the discretized action space [21], [44] and hence leads to high training time.

The MO-A3Cs architecture allows each local agent to interact independently within a DL multi-cell massive MIMO network, collecting diverse trajectories and training the global network. Fig. 5 analyzes the impact of the number of local agents on the training performance of the proposed MO-A3Cs model. It can be observed in Fig. 5(a), that 4 agents have limited diversity in the collected samples which results in relatively lower training performance. In contrast, employing 16 agents provides an enhanced average SE reward, outperforming 4 and 8 agents by 9.48% and 3.37%. Similarly, Fig. 5(b) depicts that a higher average EE reward is achieved by 16 agents compared to the 4 and 8 agents and outperforms the counterpart by 7.97% and 3.83%, respectively.

Fig. 6(a) and (b) present the impact of uniform and random initialization methods in average SE and EE rewards in the MO-A3Cs model training, respectively. The simulation results demonstrate that utilizing the random method in the proposed MO-A3Cs model achieves 1.57% and 1.55% higher SE and EE rewards than the uniform initialization and is more effective in the proposed Bayesian rule-based preference update mechanism. Similarly, Fig. 6(c) and (d) present the average objective rewards of the MO-A3Cs model for various preference weight update methods. It is worth noting that the uniform update method sets both preference weights to 0.5 while the random update method adjusts preference weights randomly between 0 and 1. The exp update method adjust preference weights based on $\omega_o(t) = \exp(-\upsilon_o \times t)$, where, υ_o is the weight decrease rate. The simulation results show that the proposed Bayesian rule-based preference weight update technique outperforms other methods in terms of objective rewards. Fig. 6(c) shows that the proposed Bayesian rule provides improved SE reward of 0.55%, 1.58%, and 5.95% compared to the exp, uniform, and random methods, respectively. Similarly, Fig. 6(d) depicts that the proposed Bayesian rule provides improved EE reward of 3.16%, 7.88%, and 12.75% compared to the exp, uniform, and random methods, respectively. Conventional update methods in the MORL algorithm tend to prioritize SE over EE due to relatively higher values, leading to a decreased priority for EE. On the other hand, our proposed Bayesian rule-based update method adaptively adjusts the weights based on trajectories obtained from an interaction between the agents and the environment.

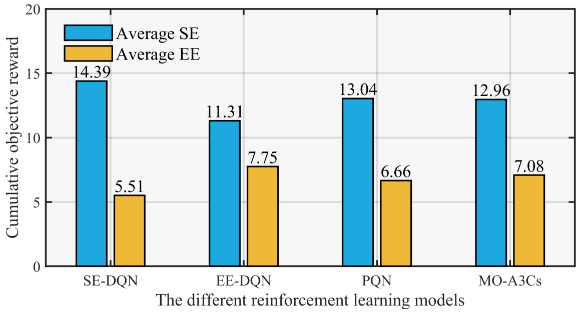


FIGURE 8. Comparative analysis of average cumulative multi-objective reward achieved by the various DRL and MORL models training.

TABLE 5. Comparison of trade-offs optimization in DL transmit PA methods across varying numbers of transmit antennas. (M = [20, 100]).

PA methods	Optimal point	Average SE	Average EE
Equal	(21.41, 4.61)	24.90	3.96
Dinkelbach	(25.08, 5.16)	29.16	4.42
SE-DQN	(28.16, 4.38)	29.75	3.91
EE-DQN	(19.87, 6.72)	28.78	4.55
PQN	(25.84, 4.89)	29.70	4.15
MO-A3Cs	(25.69, 5.26)	29.60	4.37

TABLE 6. Comparison of trade-offs optimization in DL transmit PA methods across varying maximum transmit power. ($P_{\max}$ = [20, 60]).

PA methods	Optimal point	Average SE	Average EE
Equal	(19.77, 4.04)	22.87	3.37
Dinkelbach	(23.55, 4.56)	27.20	3.80
SE-DQN	(24.49, 4.18)	28.01	3.38
EE-DQN	(23.52, 4.69)	26.82	3.90
PQN	(24.36, 4.30)	27.80	3.59
MO-A3Cs	(24.22, 4.63)	27.74	3.71

Fig. 7(a) and (b) illustrate the training performance of various aggregation methods such as sum-reward, combined-reward, and the proposed balanced-reward. The simulation results reveal that the sum-reward method mainly focuses on increasing SE during model training and overlooks the trade-off between SE and EE. In contrast, the combined-reward method, which reflects preference weights provides enhanced performance with 3.24% and 11.10% improvement for EE and SE compared to the sum-reward method. Furthermore, our proposed balanced-reward method outperformed other aggregation methods with an improved EE and SE of 5.23% and 17.92% compared to the sum-reward method and 1.93% and 6.13% compared to the combined-reward method. Fig. 7(c) and (d) present the training performance of the proposed MO-A3Cs model in comparison to the PQN model, a representative method of the MORL algorithm. The simulation results indicate that the average SE reward of MO-A3Cs was approximately 2.26% lower while its EE reward was about 7.56% higher than the PQN model. This concludes that the proposed MO-A3Cs model achieves more effective joint optimization of the average SE and EE rewards compared to the PQN model. Furthermore, the PQN model converges rapidly to a sub-optimal solution for average EE reward due to reliance on samples from replay buffers, while the MO-A3Cs model employs a multi-agent strategy that enhances sampling efficiency without utilizing buffer memory.

Fig. 8 depicts the average cumulative rewards achieved for each SE and EE during the training process of the utilized DRL and MORL models. The SE-DQN aims to maximize SE only and achieve the highest average cumulative reward of 14.39 for SE. Similarly, the EE-DQN focuses only on EE maximization and achieves an EE reward of 7.75. However, these single-objective models tend to maximize one target reward at the expense of other objectives. In contrast, the proposed MO-A3Cs achieved average cumulative SE and EE rewards of 12.96 and 7.08, respectively. Furthermore, The difference between the average cumulative SE and EE rewards is 5.88 for MO-A3Cs, 8.88 for SE-DQN, 3.56 for EE-DQN, and 6.38 for the PQN model.

2) SPECTRAL AND ENERGY EFFICIENCY EVALUATION

Table 5 demonstrates the trade-off optimization performance for each PA method with varying numbers of transmit antennas from 20 to 100. The equal PA technique shows the lowest performance across all metrics, as it does not undertake efficient DL power control. In addition, the optimal points for the SE-DQN and EE-DQN methods were recorded as (28.16, 4.38) and (19.87, 6.72), respectively. These methods can maximize specific objectives while sacrificing other objectives. In contrast, the Dinkelbach method achieved points of (25.08, 5.16). Moreover, the proposed MO-A3Cs model-based PA technique recorded the most balanced point (25.69, 5.26), while the PQN showed a slightly higher SE value of (25.84, 4.89) than the MO-A3Cs. However, a notable difference is observed in EE values. Regarding average SE, the PQN method approximates the performance of SE-DQN with a value of 29.70, while the MO-A3Cs record a slightly lower at 29.60. For average EE, the MO-A3Cs achieve an improved value of 4.37, representing a 0.22 enhancement over PQN. This suggests that the MO-A3Cs model-based DL transmission PA technique optimizes most effectively at a trade-off problem between SE and EE.

Table 6 presents the SE-EE trade-off optimization performance with the number of antennas fixed at 40, while the maximum transmission power ranges from 20 to 60 dBm. These changes in transmission power constraints directly impact the action space and the policy performance of the utilized models. The simulation results demonstrate that the proposed MO-A3Cs method achieves the most efficient optimal points at (24.22, 4.63). In contrast, the PQN appears to closely approximate the performance of the SE-DQN. Moreover, with its adopted PFA approach, the PQN method requires diverse samples to generate the Pareto set and approximate the Pareto front, especially with changes in key parameters such as $P_{\max}$. This implies a need for expanded buffer memory and training duration, in contrast to our proposed MO-A3Cs method.

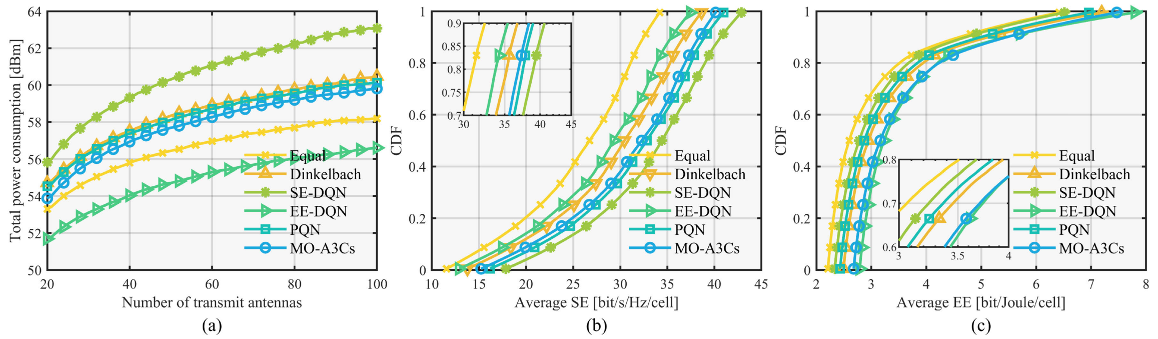


FIGURE 9. Performance analysis of the proposed PA technique in multi-cell massive MIMO networks; (a) Total power consumption vs. numbers of antennas per BS, (b) CDF of average SE, and (c) CDF of EE.

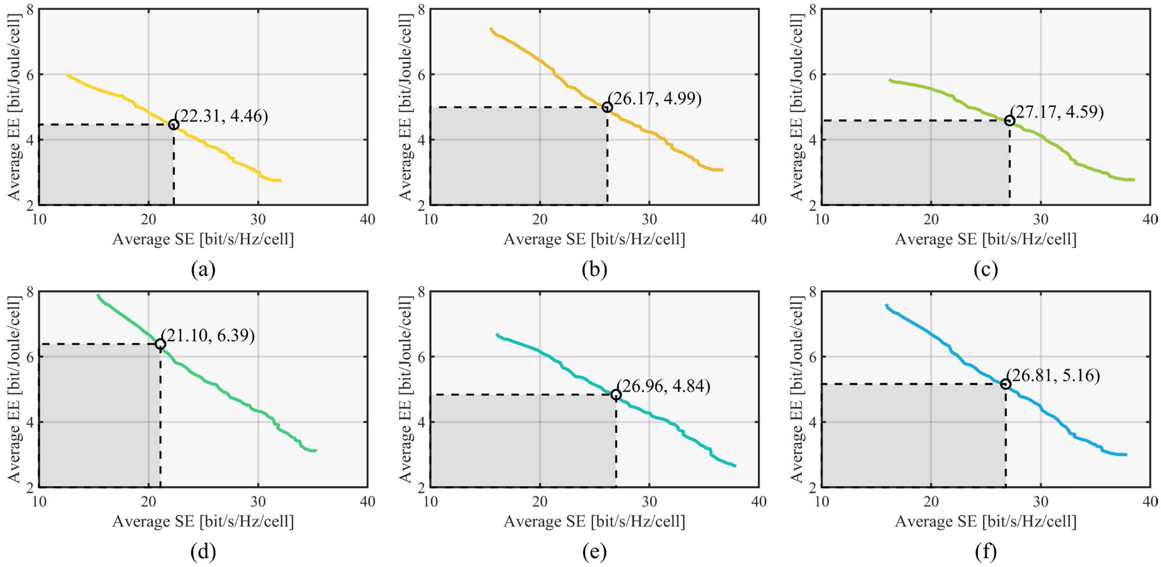


FIGURE 10. SE and EE trade-off for different PA techniques in massive MIMO networks with $L = 16$, $K = 10$, and M ranges between 20 to 80: (a) Equal, (b) Dinkelbach, (c) SE-DQN, (d) EE-DQN, (e) PQN, and (f) MO-A3Cs.

Fig. 9 presents the performance of the proposed MO-A3Cs model-based PA in terms of total power consumption, average SE, and EE. Fig. 9(a) illustrates the total power consumption of the proposed MO-A3Cs in comparison with the benchmark PA techniques while settings $L = 16$, $K = 10$, and varying transmit antennas M between 20 and 100. It can be observed that total power consumption increases with increases in the number of antennas. In addition, the average power consumption for the SE-DQN, EE-DQN, Dinkelbach, and PQN methods across this range of antennas are 60.37 dBm, 54.76 dBm, 58.32 dBm, and 58.13 dBm, respectively. Furthermore, the proposed MO-A3Cs-based PA showed an average consumption of 57.68 dBm, which is about 4.66% lower than the SE-DQN and 5.07% higher than the EE-DQN. It also consumed approximately 0.78% and 1.11% less power than the PQN and Dinkelbach, respectively.

Fig. 9(b) and (c) present the cumulative distribution function (CDF) of the SE and EE for various PA techniques under the settings. In Fig. 9(b) depicts that the SE performance of the proposed MO-A3Cs-based PA technique is approximately 2.19% lower than the PQN method. However, compared to EE-DQN and the Dinkelbach techniques, MO-A3Cs achieve performance gains of 9.07% and 5.39%, respectively. The enhanced SE performance of the PQN compared to the MO-A3Cs can be attributed to differences in their training policies, leading to different optimal DL transmission powers. Similarly, Fig. 9(c) indicates that the EE performance of the MO-A3Cs is close to the EE-DQN method. However, in terms of average EE both PQN and Dinkelbach perform 8.79% and 6.32% lower than the proposed MO-A3Cs, respectively. It can be concluded from Fig. 9(b) and (c) that the proposed MO-A3Cs-based PA technique provides the optimized SE and EE performance while ensuring the trade-off between them in multi-cell massive MIMO networks.

Fig. 10 illustrates the trade-off curves for various algorithms and the proposed MO-A3Cs-based PA methods for the different numbers of transmit antennas. It can be observed

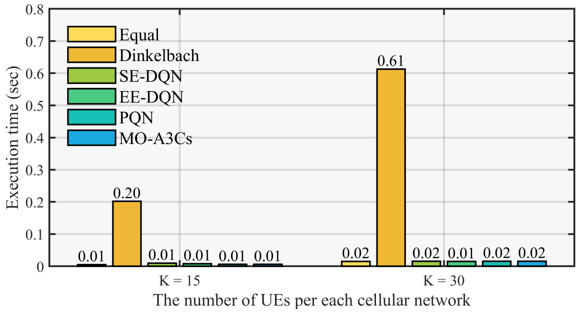


FIGURE 11. Execution time comparison of each DL PA for different numbers of UEs and $L = 16$, and $M = 40$ including pre-trained models.

that the equal PA technique resulted in the lowest performance for both objectives with the achievable trade-off point of (22.31, 4.46) for SE and EE. The SE-EE trade-off points for SE-DQN and EE-DQN are (27.17, 4.59) and (21.10, 6.39), respectively, which indicates that maximizing one performance metric comes at the expense of other metrics. In contrast, Dinkelbach's optimal points are determined to be (26.17, 4.99), showing a more balanced solution between SE and EE compared to SE-DQN and EE-DQN. Moreover, the PQN achieves an optimal trade-off point between SE and EE of (26.96, 4.84) with an improvement in SE by 0.79 and a decrease in EE by 0.15 over the Dinkelbach-based PA. Finally, the proposed PA method achieves an optimal trade-off point of (26.81, 5.16) between SE and EE which has a slightly lower SE of 0.15 than the PQN model but exhibits the highest EE performance among all the PA techniques considering MOO. In addition, considering the power consumption of the MO-A3Cs model-based PA method as presented in Fig. 9(a), it is observed that it consumes the least DL transmission power among the PA benchmarks for achieving MOO, while also achieving the SE-EE trade-off most efficiently.

Fig. 11 presents the execution time of the pre-trained proposed MO-A3Cs model-based PA method as a function of K in comparison with the other PA techniques. Fig. 11 depicts that the execution time for the Dinkelbach-based PA method increases exponentially with the number of UEs whereas the DRL, MORL-based, and equal PA techniques have less computational complexity even for a higher number of UEs.

These simulation results show that the proposed MO-A3Cs-based PA framework provides reduced computational costs compared to the iterative algorithms while ensuring robust and joint optimization of SE and EE in dynamically changing DL multi-cell massive MIMO networks.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a transmit PA technique based on the MO-A3Cs model to achieve the SE-EE trade-off in DL multi-cell massive MIMO networks. The proposed MO-A3Cs model learns the optimal joint policies to optimize the SE and EE by integrating the MARL-based training strategy with the proposed MORL algorithm. Unlike deep learning and iterative algorithms, trial-and-error-based reinforcement learning maximizes the rewards, takes optimal action through real-time interactions with the environment, and ensures adaptability and robustness in various network scenarios.

Comprehensive simulation results demonstrate that our proposed MO-A3Cs-based PA method optimizes the SE-EE trade-off more effectively and outperforms the conventional MORL algorithm with the PFA approach in terms of joint SE-EE optimization in a dynamic environment. Furthermore, the proposed approach is capable of achieving an optimal EE of 5.16 and SE of 26.81 in multi-cell massive MIMO networks with less computational complexity compared to the iterative algorithms.

In spite of the gains achieved by the proposed approach, there remains a challenge in considering the continuous action space. This work considers a discretized action space, which may not yield the optimal DL transmit power that might be achievable in a continuous action space. Nevertheless, the MORL algorithm is easily scalable and addresses issues related to training strategies with replay memory usage. There is a need to study the training patterns of action space and policies within heterogeneous network scenarios with different cell characteristics. Moreover, to enhance the practical applicability of our proposed architecture, we need to validate it in real-world scenarios. To this end, we also plan to implement a software-defined radio (SDR)-based testbed. This will enable us to validate our architecture in a setting that closely resembles real-world multi-cell massive MIMO networks. As for future work, we aim to extend the proposed MO-A3Cs framework to a continuous action space and address a broader spectrum of complex MOO challenges in 5G network resource allocation.

REFERENCES

[1] F. A. P. de Figueiredo, "An overview of massive MIMO for 5G and 6G," *IEEE Latin Amer. Trans.*, vol. 20, no. 6, pp. 931–940, Jun. 2022.

[2] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

[3] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[4] P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: A survey," *IEEE Access*, vol. 5, pp. 11727–11758, 2017.

[5] Z. Liu, W. Du, and D. Sun, "Energy and spectral efficiency tradeoff for massive MIMO systems with transmit antenna selection," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4453–4457, May 2017.

[6] A. Salh, L. Audah, Q. Abdullah, N. Abdullah, N. S. M. Shah, and A. Saif, "Trade-off energy and spectral efficiency with multi-objective optimization problem in 5G massive MIMO system," in *Proc. 1st Int. Conf. Emerg. Smart Technol. Appl. (eSmarTA)*, Aug. 2021, pp. 1–6.

[7] X. Tian, Y. Huang, S. Verma, M. Jin, U. Ghosh, K. M. Rabie, and D.-T. Do, "Power allocation scheme for maximizing spectral efficiency and energy efficiency tradeoff for uplink NOMA systems in B5G/6G," *Phys. Commun.*, vol. 43, Dec. 2020, Art. no. 101227.

[8] W.-Y. Chen, P.-Y. Hsieh, and B.-S. Chen, "Multi-objective power minimization design for energy efficiency in multicell multiuser MIMO beamforming system," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 31–45, Mar. 2020.

[9] X. Wang, Y. Wang, W. Ni, R. Sun, and S. Meng, "Sum rate analysis and power allocation for massive MIMO systems with mismatch channel," *IEEE Access*, vol. 6, pp. 16997–17009, 2018.

[10] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep learning power allocation in massive MIMO," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput. (ACSSC)*, Oct. 2018, pp. 1257–1261.

[11] R. H. Y. Perdana, T.-V. Nguyen, and B. An, "Deep learning-based power allocation in massive MIMO systems with SLNR and SINR criterions," in *Proc. 12th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Aug. 2021, pp. 87–92.

[12] M. Zhang and M. Chen, "Power allocation in multi-cell system using distributed deep neural network algorithm," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2019, pp. 1–4.

[13] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[14] A. Anzaldo and G. Andrade, "Training effect on AI-based resource allocation in small-cell networks," in *Proc. IEEE Latin-American Conf. Commun. (LATINCOM)*, Nov. 2021, pp. 1–6.

[15] S. Zhang, L. Li, J. Yin, W. Liang, X. Li, W. Chen, and Z. Han, "A dynamic power allocation scheme in power-domain NOMA using actor-critic reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 719–723.

[16] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, "Downlink power control for cell-free massive MIMO with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6772–6777, Jun. 2022.

[17] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[18] Y. Yang, F. Li, X. Zhang, Z. Liu, and K. Y. Chan, "Dynamic power allocation in cellular network based on multi-agent double deep reinforcement learning," *Comput. Netw.*, vol. 217, Nov. 2022, Art. no. 109342.

[19] M. Rahmani, M. Bashar, M. J. Dehghani, P. Xiao, R. Tafazolli, and M. Debbah, "Deep reinforcement learning-based power allocation in uplink cell-free massive MIMO," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 459–464.

[20] R. Chen, F. Sun, L. Chen, K. Li, L. Wu, J. Wang, and Y. Yang, "Adaptive multi-objective reinforcement learning for Pareto frontier approximation: A case study of resource allocation network in massive MIMO," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 1631–1635.

[21] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dyn. Syst.*, vol. 13, nos. 1–2, pp. 41–77, 2003.

[22] T. Van Chien, E. Bjömson, and E. G. Larsson, "Joint pilot sequence design and power control for max-min fairness in uplink massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[23] H. T. Dao and S. Kim, "Disjoint pilot power and data power allocation in multi-cell multi-user massive MIMO systems," *IEEE Access*, vol. 6, pp. 66513–66521, 2018.

[24] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Mach. Learn.*, vol. 84, nos. 1–2, pp. 51–80, Jul. 2011.

[25] R. Yang, X. Sun, and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14636–14647.

[26] A. Al-hubaishi, N. Noordin, A. Sali, S. Subramaniam, and A. Mohammed Mansoor, "An efficient pilot assignment scheme for addressing pilot contamination in multicell massive MIMO systems," *Electronics*, vol. 8, no. 4, p. 372, Mar. 2019.

[27] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[28] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[29] B. Saleeb, M. Shehata, H. Mostafa, and Y. Fahmy, "Performance evaluation of RZF precoding in multi-user MIMO systems," in *Proc. IEEE 62nd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2019, pp. 1207–1210.

[30] H. Liu, H. Deng, Y. Yi, Z. Zhu, G. Liu, and J. Zhang, "Energy efficiency optimization based on power allocation in massive MIMO downlink systems," *Symmetry*, vol. 14, no. 6, pp. 1145–1160, Jun. 2022.

[31] J. Zhang, H. Deng, Y. Li, Z. Zhu, G. Liu, and H. Liu, "Energy efficiency optimization of massive MIMO system with uplink multi-cell based on imperfect CSI with power control," *Symmetry*, vol. 14, no. 4, pp. 780–795, Apr. 2022.

[32] H. Yang and T. L. Marzetta, "Total energy efficiency of cellular large scale antenna system multiple access mobile networks," in *Proc. IEEE Online Conf. Green Commun. (OnlineGreenComm)*, Oct. 2013, pp. 27–32.

[33] O. Amin, E. Bedeer, M. H. Ahmed, and O. A. Dobre, "Energy efficiency–spectral efficiency tradeoff: A multiobjective optimization approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1975–1981, Apr. 2016.

[34] T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim, "A multi-objective deep reinforcement learning framework," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103915.

[35] S. Huang, A. Abdolmaleki, G. Vezzani, P. Brakel, D. J. Mankowitz, M. Neunert, S. Bohez, Y. Tassa, N. Heess, M. Riedmiller, and R. Hadsell, "A constrained multi-objective reinforcement learning framework," in *Proc. Conf. Robot Learn.*, 2022, pp. 883–893.

[36] P. Vamplew, J. Yearwood, R. Dazeley, and A. Berry, "On the limitations of Scalarisation for multi-objective reinforcement learning of Pareto fronts," in *Proc. 21st Austa. Joint Conf. Artif. Intell. (AJCAI)*. Heidelberg, Germany: Springer, 2008, pp. 372–378.

[37] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10607–10616.

[38] C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 3, pp. 385–398, Mar. 2015.

[39] E. Friedman and F. Fontaine, "Generalizing across multi-objective reward functions in deep reinforcement learning," 2018, *arXiv:1809.06364*.

[40] T. Basaklar, S. Gumussoy, and U. Ogras, "PD-MORL: Preference-driven multi-objective reinforcement learning algorithm," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=zS9sRyaPFlJ

[41] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 11–20.

[42] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers, "A practical guide to multi-objective reinforcement learning and planning," *Auton. Agents Multi-Agent Syst.*, vol. 36, no. 1, p. 26, Apr. 2022.

[43] K. Van Moffaert, M. M. Drugan, and A. Nowé, "Scalarized multi-objective reinforcement learning: Novel design techniques," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn. (ADPRL)*, Apr. 2013, pp. 191–199.

[44] J. Lee and J. So, "Reinforcement learning-based joint user pairing and power allocation in MIMO-NOMA systems," *Sensors*, vol. 20, no. 24, p. 7094, Dec. 2020.

[45] K. I. Ahmed and E. Hossain, "A deep Q-learning method for downlink power allocation in multi-cell networks," 2019, *arXiv:1904.13032*.

[46] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[47] J. V. Stone, *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sheffield, U.K.: Sebtel Press, 2013.

[48] N. Vlassis, M. Ghavamzadeh, S. Mannor, and P. Poupart, "Bayesian reinforcement learning," *Reinf. Learn., State-Art*, vol. 12, no. 4, pp. 359–386, 2012.

[49] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, Aug. 1988.

[50] H. Lu, D. Herman, and Y. Yu, "Multi-objective reinforcement learning: Convexity, stationarity and Pareto optimality," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=TjEzIsyEsQ6

[51] R. Pasunuru and M. Bansal, "Multi-reward reinforced summarization with saliency and entailment," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2018, pp. 646–653.

[52] J. Schulman, X. Chen, and P. Abbeel, "Equivalence between policy gradients and soft Q-learning," 2017, *arXiv:1704.06440*.

[53] Q. Zhang, S. Jin, M. McKay, D. Morales-Jimenez, and H. Zhu, "Power allocation schemes for multicell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5941–5955, Nov. 2015.

[54] G. Yu, L. Xu, D. Feng, Z. Zhang, G. Y. Li, and H. Zhang, ‘‘Energy efficiency tradeoff in interference channels,’’ *IEEE Access*, vol. 4, pp. 4495–4508, 2016.

[55] S. Schaible, ‘‘Fractional programming: Applications and algorithms,’’ *Eur. J. Oper. Res.*, vol. 7, no. 2, pp. 111–120, Jun. 1981.

[56] H. Li, H. Gao, T. Lv, and Y. Lu, ‘‘Deep Q-learning based dynamic resource allocation for self-powered ultra-dense networks,’’ in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[57] S. Sharma and W. Yoon, ‘‘Multiobjective reinforcement learning based energy consumption in C-RAN enabled massive MIMO,’’ *Radioengineering*, vol. 31, no. 1, pp. 155–163, Apr. 2022.

[58] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

YOUNGWOO OH (Student Member, IEEE) received the B.S. degree from the Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea, in 2022. He is currently pursuing the M.S. degree with the Smart Networking Laboratory (SNL), Chosun University. His current research interests include massive MIMO, 5G and beyond 5G networks, deep reinforcement learning-based optimization, the Internet of Things, radio frequency sensing, and experiment-driven evaluation of wireless networks.

ARIF ULLAH (Member, IEEE) received the M.S. degree in electrical engineering from COMSATS University, Islamabad, Pakistan, and the Ph.D. degree in electronic and communication engineering from the Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology, Topi, Pakistan. He was a Graduate Research Assistant with the GIK Institute and a Research Fellow with the Smart Networking Laboratory (SNL), Chosun University, South Korea, from November 2021 to April 2022. He is currently an Assistant Professor with the Department of Computer Engineering, College of IT Convergence, Chosun University. His research interests include modeling and performance analysis of cellular networks using stochastic geometry, millimeter wave and terahertz communication, vehicular networks, the Internet of Things, 5G and beyond 5G networks, and machine learning-aided wireless communication.

WOOYEOL CHOI (Member, IEEE) received the B.S. degree from the Department of Computer Science and Engineering, Pusan National University, Busan, Republic of Korea, in 2008, and the M.S. and Ph.D. degrees from the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, in 2010 and 2015, respectively. From 2015 to 2017, he was a Senior Research Scientist with the Korea Institute of Ocean Science and Technology (KIOST), Ansan, Republic of Korea. From 2017 to 2018, he was a Senior Researcher with the Korea Aerospace Research Institute (KARI), Daejeon, Republic of Korea. He is currently an Associate Professor with the Department of Computer Engineering, Chosun University, Gwangju. His research interests include cross layer-protocol design, learning-based resource optimization, and experiment-driven evaluation of wireless networks.