Throughput-Constrained Antenna Sleep Management for Saving Power

Rie Tagyo, Hideaki Kinsho, Akihiro Shiozu, and Kazuhisa Yamagishi
NTT Network Service Systems Laboratories, NTT Corporation, Tokyo, Japan
Emails: {rie.tagyo, hideaki.kinsho, akihiro.shiozu, kazuhisa.yamagishi}@ntt.com

***Abstract*—In cellular networks, the rapid increase in traffic demand needs to be met by additional equipment, but the resulting increase in power consumption significantly impacts CO2 emissions. The impact on power consumption is particularly substantial due to the large number of base stations (BSs). As a result, power consumption needs to be reduced by temporarily suspending BSs with low traffic usage through BS sleep management. On the other hand, the effect of BS sleep on communication quality also needs to be considered. This paper proposes a new antenna sleep management framework to reduce power consumption while maintaining communication quality that deals with sleep on an antenna-by-antenna basis. The framework uses a two-stage deep reinforcement learning (DRL) approach that incorporates the concept of safe DRL to ensure that users' throughput is above the target value. Results indicate that for the sleep management problem of multiple antennas within a single BS, constraint violations on throughput were reduced from approximately 0 to 1.7 percent while achieving a constant power reduction close to the optimum value.**

***Index Terms*—antenna sleep management, quality of service (QoS), power reduction, safe reinforcement learning**

I. Introduction

Cellular networks provide various communication services involving vast amounts of data, and this demand is expected to increase. In response to the demand, cellular carriers must expand the network systems' capacity and ensure communication quality by installing additional base stations (BSs). However, the accompanying increase in power consumption has raised concerns about environmental impacts, such as increased CO2 emissions [1]. In particular, due to their widespread deployment, the power consumption of BSs contributes significantly to the power consumption of the overall communication system [2], [3]. Thus, cellular carriers must avoid wasting power by putting BSs to sleep or switching them off when traffic usage is low while installing many BSs. In summary, reducing power consumption while ensuring communication quality to reduce CO2 emissions is desirable.

BSs must be put to sleep or switched off, considering coverage and quality of service (QoS) metrics (e.g., throughput, delay, packet loss rate). If coverage is not ensured, connectivity is impossible, not to mention QoS. Concerning QoS, in particular, throughput is essential for providing adequate quality application services. Thus, ensuring coverage and throughput is especially critical for cellular carriers.

To ensure coverage, antenna (band) sleep management is employed, where the management is applied only to some antennas. Cellular carriers generally operate antennas across multiple frequency bands. They operate coverage antennas to avoid coverage holes and ensure connectivity and also operate capacity antennas to enhance network capacity. The coverage provided by these antennas overlaps, so coverage is ensured by only targeting the capacity antennas for control. This operation limits the potential gains in power reduction through sleep management but is essential for cellular carriers, as it prioritizes maintaining connectivity.

When user equipment (UE) previously connected to an antenna that is put to sleep re-establishes connection with another active antenna, the user's throughput potentially degrades. This degradation occurs when the UEs reconnect to an antenna with a high utilization rate. The lack of resource blocks (RBs) prevents the UE's throughput from achieving the target value preset by the carrier. In such a case, the carrier must reactivate the sleeping antennas to maintain the target throughput. Implementing this response incurs manual costs, so a sleep policy needs to be established that considers the impact on throughput. However, setting appropriate policies for antennas used in many locations under various conditions takes time and effort.

To reduce CO2 emissions, we propose an antenna sleep management method that reduces power consumption while maintaining the carriers' target throughput. Adopting a deep reinforcement learning (DRL) approach automatically sets the appropriate sleep policy for the antennas efficiently under various usage conditions. To prevent the UEs' throughput from being below the carrier's preset target value, we refer to the previous research on safe deep reinforcement learning (safe DRL) [4]. Our proposed method includes a two-stage DRL approach. In the first stage, the first DRL agent aims to reduce power consumption and improve average throughput, and in the second stage, employing the safety DRL concept, the second DRL agent minimizes constraint violations while respecting the outputs of the previous agent. We evaluate the proposed method's performance for the sleep management problem of multiple antennas within a single BS.

The remainder of this paper is organized as follows. Section II reviews related works. We describe the formulation of the throughput-constrained BS sleep management problem and the idea of our proposed two-stage DRL algorithm in Section III. In Section IV, we evaluate the performance of the proposed approach by implementing the simulation for the management scenario of multiple antennas within a single BS. Finally, we conclude this paper and mention future work in Section V.

978-3-903176-66-9

II. Related Works

This section first reviews studies on sleep management of BSs, highlighting work on the trade-off between power consumption and communication quality (e.g., delay, data rate) and then studies on the application of safe DRL.

A. BS Sleep Management

Various BS sleep management techniques have been proposed to reduce power consumption [3], [5]–[18], and these techniques are actively employed in actual network operations due to their ease of software implementation. The fundamental concept is to temporarily put to sleep or switch off BSs with low traffic utilization. Three survey papers [3], [5], [6] summarize these techniques from different perspectives: [3] focuses on energy efficiency metrics and traffic model assumptions, [5] covers sleep management in 5G environments with new wireless technologies, and [6] comprehensively surveys sleep management for energy efficiency in 5G systems, including in combination with other techniques.

This section presents studies dealing with trade-offs between power consumption and QoS. Dalal et al. [7] examined sleep management and power matching for a single BS, considering the balance between total power consumption and average delay. They theoretically analyzed how to optimize parameters within specific sleep schemes. Wu et al. [8] addressed the on/off switching problem in a macro-cell system with many small cells to minimize power consumption by using analytically calculated data rates as constraints. They proposed a distributed algorithm by sharing information among small cells. Wang and Zheng [9] theoretically analyzed the average power and delay distribution using the traffic queue model concerning typical wake-up schemes. Guo et al. [10] investigated the switching problem associated with advanced sleep modes (ASM), which have different depths and can progressively switch off more circuitry depending on such time length. They developed a closed-form expression to determine the appropriate parameter settings for ASM operation, meeting the required average delay constraints. Although these theoretical studies rely on stochastic models, the traffic handled in actual operations is more complex and does not always behave model-dependently. Thus, these methods may not achieve the expected performance levels in complex real-world environments.

In contrast, many studies have proposed data-driven DRL approaches, which do not require model assumptions. These approaches are applied by formulating the BS sleep problem as a Markov decision process (MDP) problem. Liu et al. [11] improved the DQN algorithm by adding Action-Wise Experience Replay and Adaptive Reward Scaling to enable dynamic control for non-stationary traffic. Ye and Zhang [12] proposed a deep deterministic policy gradient (DDPG)-based approach that includes traffic load prediction, focusing on power consumption, average delay, and mode switching costs. Recently, several studies on ASM management using the DRL approach have discussed the relationship between delay and power consumption [13]–[15]. Salem et al. [13] used Q-learning, and Lin et al. [14] solved the BS sleep management problem simultaneously with user association by introducing tandem learning. Malta [15] discussed the trade-off between energy reduction and QoS using SARSA by using 5G primary use case requirements. Other advanced works address BS sleep management to optimize power consumption and delay [16]–[18]. Li et al. [16] proposed a transfer Actor-Critic learning framework to enhance strategies by leveraging learned knowledge from historical periods. Wu et al. [17] focused on traffic prediction using a convolutional neural network (CNN) and long short-term memory (LSTM) while dealing with DDPG-based sleep management. Additionally, Abubakar et al. [18] introduced an improved Actor-Critic DRL for ultra-dense networks able to handle large discrete action spaces. These DRL-based BS sleep management methods considering trade-offs with QoS [16], [11], [12] minimize costs by using an objective function that is the weighted sum of total power consumption and average delay. However, they do not guarantee QoS (delay) as a constraint. Even if throughput replaces delay in these approaches, while they may reduce power consumption and minimize throughput degradation, they do not ensure a certain throughput level. Moreover, these methods focus only on average QoS, not on whether each user's QoS is sufficient.

B. Safe DRL

In real-world applications of DRL, safe DRL, which considers risks from the agent's actions, has been widely studied [19]. Safe DRL is often modeled as a constrained MDP (CMDP), where the agent maximizes rewards while satisfying safety constraints. One study introduced a safety layer that analytically solves an action correction formulation per each state [4]. The safety layer is directly added to the policy network of DRL to never violate constraints during learning. We refer to this safety layer concept to satisfy the constraint that the users' throughput is above the target value in the antenna sleep management problem.

Furthermore, some studies on application of safe DRL are reviewed. Li et al. [20] solved the charging scheduling problem for electric vehicles (EVs) using safe DRL. They formulated the problem as a CMDP and solved it using a method based on constraint policy optimization (CPO) under randomly varying EV arrival/departure times and electricity prices, with the constraint ensuring charging met the target by departure. For communication-related problems, Suzuki and Harada [21] solved the dynamic virtual network (VN) allocation problem using a multi-agent DRL approach with objective and safe agents. They minimized total maximum link and server utilization while ensuring link and server capacities were not exceeded. Liu et al. [22] solved the end-to-end resource orchestration problem for network slicing in a mobile network, handling performance and system capacity constraints using the Lagrangian relaxation method and mapping to safe actions. Although constraints need to be considered in various applications, to the best of our knowledge, no studies have highlighted the importance of safety control in the context

of conducting BS sleep management or assessing the carriers' risk that sufficient throughput is not maintained.

C. Contributions

There are challenges to antenna sleep management that reduce power consumption while maintaining the carriers' target throughput. This subsection summarizes the key points addressed in this paper.

- First, we adopt data-driven DRL approaches that do not assume a specific traffic model to make it applicable to real-world operations. This allows for data-driven, dynamic, and autonomous control.
- Previous approaches that do not treat QoS as a constraint cannot guarantee users a certain level of communication quality preset by the carriers. Moreover, these methods do not focus on each user's QoS in the control outcomes. Thus, we propose a safe antenna sleep management that minimizes violations with the constraint that the QoS, i.e., the throughput, is above the target value. Furthermore, we focus on each user's QoS in the control outcomes and evaluate the degree of throughput constraint violation.
- Existing studies do not address safe DRL in antenna sleep management or assess the risk of carriers not maintaining sufficient throughput. In the context of BS sleep management, due to the difficulty of predicting BS utilization, we apply a novel safe DRL methodology that uses a two-stage DRL approach instead of existing safe DRL techniques. We find that safe DRL works appropriately against the constraint that the throughput in sleep management achieves a target value.
- Finally, we propose a novel antenna management framework based on a two-stage DRL approach. This framework can reduce total power consumption while maintaining the target throughput set by carriers as constraints.

III. Throughput-Constrained and Power-Saving Antenna Sleep Management Methodology

We propose a two-stage DRL approach-based antenna management method that reduces total power consumption while maintaining the target throughput set by carriers as constraints. The overall architecture of the two-stage DRL approach is shown in Fig. 1. In the first stage DRL, similar to an existing approach [16], the DRL agent explores a policy that maximizes total power reduction and average throughput. In the second stage DRL, by introducing the concept of a safety layer [4] and relaxing the constraints, the DRL agent explores a policy that modifies the original actions into safe actions.

A. System Model and Problem Formulation

Before describing the proposed method for safe antenna sleep management, we define a system to be controlled. We consider a system with multiple antennas providing cellular network communication, where it is assumed that antennas can be put to sleep on a per-antenna basis. Let $\mathcal{B}$ be the set of all antennas, where $\mathcal{B}_c$ $(\in \mathcal{B})$ represents the antennas to be controlled. The total number of antennas is represented by $|\mathcal{B}|$, and the number of antennas to be controlled is denoted by $|\mathcal{B}_c|$. Note that the antennas to be controlled correspond to those for additional capacity, while the others are for coverage.

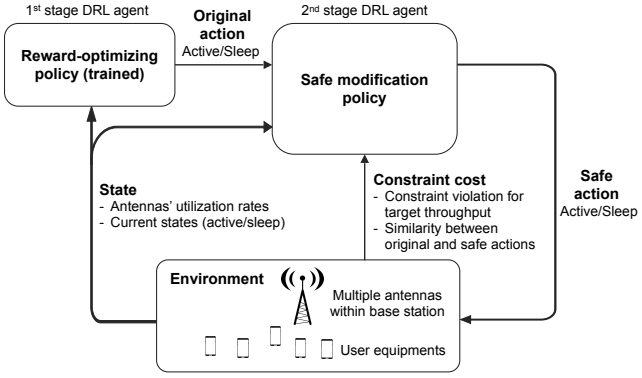

Fig. 1. Architecture of safe antenna sleep management framework with two-stage reinforcement learning approach.

Let $x_i^{(t)} \in \{0,1\}$ $(i \in \mathcal{B})$ be the active/sleep state of the i-th antenna at time t (1 indicates active, 0 indicates sleep), and we define the state of all antennas $\mathcal{B}$ as $\boldsymbol{x}_t = [x_1^{(t)}, x_2^{(t)}, \cdots, x_{|\mathcal{B}|}^{(t)}]$.The traffic utilization, such as the RB utilization rate, of the i-th antenna at time t is $\rho_i^{(t)}$ and that of all antennas as $\boldsymbol{\rho}_t = [\rho_1^{(t)}, \rho_2^{(t)}, \cdots, \rho_{|\mathcal{B}|}^{(t)}]$, where the utilization rate of the antenna in sleep mode is zero.

We use the general power consumption model for BSs consisting of constant power and adaptive power proportional to BS's utilization, adopted in other studies, including [16], for the power consumption of the antennas. The total power consumed by all antennas at time t is represented as follows.

$$P_{\text{total}}(t) = \sum_{i \in \mathcal{B}} x_i^{(t)} \left[(1-\alpha_i) P_i + \alpha_i \rho_i^{(t)} P_i \right], \qquad (1)$$

where $\alpha_i \in (0,1)$ is the constant power consumption percentage for the i-th antenna, and P_i is the maximum power consumption of the i-th antenna when it is fully utilized.

Let $\mathcal{N}$ be the set of UEs in the coverage area provided by antennas $\mathcal{B}$. The throughput of the n-th UE at time t is denoted as $q_n^{(t)}$ and the condition we aim to satisfy is shown below,

$$q_n^{(t)} \geq q_n^* \quad (\forall n \in \mathcal{N}), \qquad (2)$$

where the throughput values for all UEs are above their respective target values q_n^* $(n \in \mathcal{N})$. Note that it is assumed that each UE can achieve the target value if the antennas provide sufficient capacity. In other words, the assumption is that RB and power allocations appropriately work to ensure that each UE's throughput achieves its target value.

We explain the problem formulation and present the power reduction rate and average throughput as the components of the objective function. Defining the power consumption at time t when all antennas are in an active state as $\tilde{P}(t)$, the power reduction rate at time t, denoted as $P_r(t)$, is defined as $P_r(t) = 1 - P_{\text{total}}/\tilde{P}(t)$. It equals zero when all antennas are in an active state. The average throughput at time t is denoted as $Q(t)$ and

is given by $Q(t) = \sum_{n\in\mathcal{N}} q_n^{(t)}$. The problem to be solved is as follows,

$$\begin{aligned} \max_{\boldsymbol{x}_t} \quad & w_1 P_r(t) + (1-w_1)Q(t) \\ s.t. \quad & q_n^{(t)} \geq q_n^*, \quad \forall n \in \mathcal{N}, \end{aligned} \tag{3}$$

where w_1 represents the weight parameter indicating the importance of the power reduction rate and average throughput. The original problem (3) is divided into two MDP problems: $\mathcal{M}_1$ and $\mathcal{M}_2$. Then, these problems are solved using a DRL methodology as a two-stage DRL approach.

B. The First Stage: Reward-Optimizing Problem

The problem in the first stage denoted as $\mathcal{M}_1$, optimizes the power reduction rate and average throughput. It is defined as an MDP with a 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states; $\mathcal{A}$ is the set of actions; $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function; $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function; γ is a discount factor.

State $\mathcal{S}$: The state at time t is defined as $s_t = (\boldsymbol{\rho}_t, \boldsymbol{x}_{t-1})$, containing two types of information: the utilization of all antennas and the active/sleep states of all antennas before control execution.

Action $\mathcal{A}$: The original action executed at time t after observing the state is defined as $a_t = \{a_i^{(t)} | i \in \mathcal{B}_c, a_i^{(t)} \in \{0, 1\}\}$. This means if the i-th antenna is an active state, $a_i^{(t)} = 1$; if it is a sleep state, $a_i^{(t)} = 0$.

Reward R: The reward function $R(s_t, a_t)$ is denoted as r_t, which is formulated as $r_t = w_1 P_r(s_t) + (1 - w_1)Q(s_t)$. The objective is to explore a policy π that maximizes the total discounted return,

$$\max_{\pi} \quad \mathbb{E}_{\tau\sim\pi}\left[\sum_{t\in\mathcal{T}} \gamma^t r_t\right], \tag{4}$$

where $\mathcal{T}$ denotes the entire control time and τ represents the trajectory of states and actions over $\mathcal{T}$. $\mathbb{E}(*)$ denotes the expected value, and in this problem, it calculates the expected value of discounted rewards obtained through the trajectory under policy π. Moreover, the transition probability denoted as P, is influenced by the randomness of UEs reconnecting when the active/sleep state of the antennas changes. In this real-world scenario, the transition probabilities are considered unknown.

C. The Second Stage: Safe Modification Problem

In the second stage, the agent explores a policy that modifies the original actions into safe actions, where it minimizes constraint violations and the similarity between the actions to be exploited by the second DRL and the original actions received from the first agent. Minimizing constraint violations is intended to reduce the number of users having throughput under the target values, and minimizing the similarity is designed to respect the outputs of the previous agent.

The problem in the second stage denoted as $\mathcal{M}_2$ modifies the original actions into safe actions. It is defined as another MDP with $(\mathcal{S}', \mathcal{A}', P', C, \gamma')$, where $\mathcal{S}'$ is the set of states; $\mathcal{A}'$ is the set of actions; $P' : \mathcal{S}' \times \mathcal{A}' \times \mathcal{S}' \to [0, 1]$ is the transition probability function; $C : \mathcal{S}' \times \mathcal{A}' \times \mathcal{S}' \to \mathbb{R}$ is the cost function; γ' is a discount factor.

State $\mathcal{S}'$: The state at time t is represented as $s'_t = (s_t, a_t) = (\boldsymbol{\rho}(t), \boldsymbol{x}(t), a_t)$, which includes the action outputted by the first stage DRL, in addition to the same information as the state s_t.

Action $\mathcal{A}'$: The safe action at time t is defined as $a'_t = \{a_i'^{(t)} | i \in \mathcal{B}_c, a_i'^{(t)} \in \{0, 1\}\}$. This is the same notation as the original action in the first stage.

Cost C: The cost function $C(s'_t, a'_t) = C(s_t, a_t, a'_t)$ is denoted as c_t. This cost consists of two components. The first is the cost for constraint violations, and the second is the cost indicating the similarity between the action a from the previous stage and the safe action a' to be outputted. The cost is formulated as

$$c_t = w_2 C_v(s'_t, a'_t) + (1 - w_2)C_s(a_t, a'_t), \tag{5}$$

where w_2 represents the weight parameter that indicates the magnitude of the costs associated with constraint violations and the similarity. The objective is to explore a policy π',

$$\min_{\pi'} \quad \mathbb{E}_{\tau'\sim\pi'}\left[\sum_{t\in\mathcal{T}} \gamma'^t c_t\right]. \tag{6}$$

The two costs are explained in detail below. The cost for constraint violation is denoted as C_v, which is expressed as follows,

$$C_v(s'_t, a'_t) = \tanh(\eta|\mathcal{N}_v|). \tag{7}$$

Here, $|\mathcal{N}_v|$ represents the number of UEs that violate the constraint of not achieving their respective target throughput as follows,

$$|\mathcal{N}_v| = \sum_{n\in\mathcal{N}} \mathbb{I}\left[q_n(t) < q_n^*\right], \tag{8}$$

and η represents the scaling parameter of the hyperbolic tangent function. The reason for using a smooth (differentiable) function, not a step function for the cost function, is to make learning easier. The cost indicating the similarity between a_t and a'_t is denoted as C_s, which is represented using the difference in the number of active antennas, denoted as $\Delta_t = \sum_{i\in\mathcal{B}_c}(a_i'^{(t)} - a_i^{(t)})$, as follows.

$$C_s(a_t, a'_t) = \begin{cases} \dfrac{\exp(\Delta_t) - 1}{\exp(|\mathcal{B}_c|) - 1} & \Delta_t \geq 0 \\ 1 & \Delta_t < 0 \end{cases}. \tag{9}$$

This function is adjusted to take values between 0 and 1. In the case of a constraint violation, the safe action a'_t requires more active antennas than the original action a_t. Thus, when reducing antennas, that is $\Delta_t < 0$, the cost is set to 1, whereas when increasing antennas, that is $\Delta_t \geq 0$, a function that exponentially increases is used to minimize constraint violations while adding as few antennas as possible.

Algorithm 1 Training Procedure of Two-Stage DRL

Initialize first policy networks π_θ, replay buffer D
for each episode **do**
  Initialize state s_t from the environment
  for each step **do**
    Sample action $a_t \sim \pi_\theta(s_t)$
    Take action a_t, observe new state s_{t+1}, reward r_t
    Store transition (s_t, a_t, r_t, s_{t+1}) in replay buffer D
    if D is sufficiently large **then**
      Sample a batch from D
      Update first policy network π_θ
end training first policy network π_θ
Initialize second policy network π'_ϕ, replay buffer D'
for each episode **do**
  Initialize state s_t from the environment
  for each step **do**
    Get greedy action $a_t = \pi_\theta(s_t)$
    $s'_t = (s_t, a_t)$
    Sample safe action $a'_t \sim \pi'_\phi(s'_t)$
    Take safe action a'_t, observe new state s_{t+1}
    Calculate c_t using (5) based on s'_t, a'_t
    Get greedy action $a_{t+1} = \pi_\theta(s_{t+1})$
    $s'_{t+1} = (s_{t+1}, a_{t+1})$
    Store transition $(s'_t, a'c_t, s'_{t+1})$ in replay buffer D'
    if D' is sufficiently large **then**
      Sample a batch from D'
      Update second policy network π'_ϕ
end training second policy network π'_ϕ
return π_θ, π'_ϕ

D. Training Procedure and Computational Cost

The parameters of the deep neural networks for the policies π and π' are denoted by θ and ϕ, respectively, with the networks represented as $\pi_\theta(s_t)$ and $\pi'_\phi(s'_t)$. The first policy network, $\pi_\theta(s_t)$, is trained, and subsequently, the second policy network, $\pi_\phi(s'_t)$, is trained using the first trained network, $\pi_\theta(s_t)$, as described in Algorithm 1.

The two-stage approach involves two DRL processes, approximately doubling the training time compared to a single DRL. However, once the training phase is completed, the inference time itself is very short, making the approach feasible for near real-time applications.

IV. Numerical Analysis

This section evaluates our proposed method by solving the sleep management problem for multiple antennas within a single BS. This evaluation aims to clarify whether our method can reduce power consumption while ensuring throughput by minimizing constraint violations to values close to zero, compared to baseline methods. Furthermore, we evaluate the performance of the proposed method by using different weight parameters in the cost function from (5) and discuss how to set this parameter to achieve an appropriate trade-off that reduces power consumption and constraint violations.

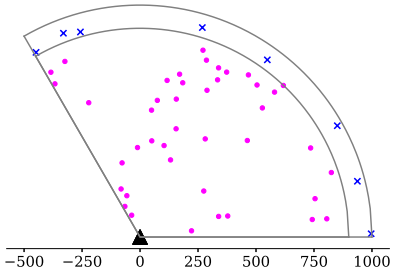

Fig. 2. The covered area of the antennas within a single base station in the simulation.

A. Simulation Settings and Conditions

We describe the simulation settings and conditions for the sleep management problem of multiple antennas within a single BS. These are three-sector antennas facing the same direction, with overlapping coverage areas depicted as sector-shaped regions shown in Fig. 2. The black triangular plot represents the BS, including all antennas. There is one coverage antenna and $|\mathcal{B}_c|$ capacity antennas. The coverage antenna operates on the 800 MHz band and covers an area with a radius of 1000 meters. The capacity antennas operate at the 1.5 GHz band, each covering an area with a radius of 900 meters. The x-shaped cross and circle plots each represent a single UE: those with the blue x-shaped cross plots are always served by the coverage antenna, while those with the magenta circle plots are served by either the coverage antenna or one of the capacity antennas. In this scenario, it is assumed that the coverage antenna has a broader coverage area, and each capacity antenna is assigned a unique frequency within the 1.5 GHz band, preventing signal interference among them.

The number of UEs, described as N, is generated in accordance with a uniform distribution $U(N_{\min}, N_{\max})$, and each UE is randomly located within the coverage area in each episode. Note that each episode contains several control steps and that the number and position of UEs do not change in each step. The throughput of a UE is calculated using Shannon's capacity, which represents the maximum achievable data rate for a channel under specified bandwidth and signal-to-noise ratio (SNR) conditions.

$$q_n = B \log(1 + \mathrm{SNR}). \tag{10}$$

To calculate the SNR, we use the free-space path loss (FSPL) model, which accounts for distance attenuation effects on signal strength as described in [23], and includes thermal noise for a single RB band. Here, B represents the channel bandwidth in hertz, corresponding to the number of allocated RBs. This formula assumes a noise-limited channel with no interference.

The operational maximum power consumption P_i $(\forall i \in \mathcal{B})$ for each antenna is set at 100 Watts, $\alpha_i = 0.4$ $(\forall i \in \mathcal{B})$, and the antennas are 30 meters high. The range of these values was determined by referring to existing studies [16], [17]; however, whereas those studies used the BS basis, we

TABLE I
SYSTEM PARAMETERS

Parameters	Value
$P_i,\ \forall i \in \mathcal{B}$	100 [W]
$\alpha_i,\ \forall i \in \mathcal{B}$	0.4
Frequency band of coverage / capacity antenna(s)	800 [MHz] / 1.5 [GHz]
Covered area's radius of coverage / capacity antenna(s)	1000 [m] / 900 [m]
BS hight	30 [m]
Number of each antenna's RBs	100
Total transmission power of each antenna	0.1 [W]
FSPL	$20\log_{10}(d) + 20\log_{10}(f)$ $+20\log_{10}(4\pi/c)$ $[dB]$
Noise	-121.45 [dBm]
w_1, w_2	0.7, 0.7
η	0.5

TABLE II
HYPERPARAMETERS FOR SAC-DISCRETE

Hyperparameters	Value
Layers	3 fully connected layers
Hidden size	256
Discount factor γ, γ'	0.9
Batch size	256
Replay buffer size	100,000
Learning rate	0.0003
Soft update parameter	0.01
Target entropy	$0.98 \times (-\log(1/\lvert\mathcal{B}_c\rvert))$
Episodes	20,000
Steps	20

determined our values considering the antenna basis. Each antenna is configured with 20 MHz bandwidth, that is 100 RBs, and the total transmission power of each antenna is set to 0.1 Watts. The other parameters for the objectives are set to $w_1 = w_2 = 0.7$ and $\eta = 0.5$. These system parameters are summarized in Table I.

We use the Soft Actor-Critic for discrete action algorithm (SAC-Discrete) [24] for both stages of the two-stage DRL process, as it enhances exploration diversity by maximizing policy entropy, leading to improved learning stability and performance in complex environments. The critic networks are trained using the techniques of double Q-learning [25] and soft target updates. The policy network and the two critic networks each consist of three linear layers. The first two layers use the Rectified Linear Units (ReLU) activation function. The policy network employs a softmax function for the final layer, whereas the critic networks use no activation function. The hyperparameters for SAC-Discrete training are summarized in Table II.

The evaluation process conducts K episodes, with each episode consisting of a single step varying the number of UEs and their location. In each episode, the single-step control is executed, and the outcomes are evaluated. The evaluation conditions are set with the number of antennas to be controlled, $|\mathcal{B}_c|$, at 2, 3, and 4. The target throughput is set to the same value for all UEs. It is denoted as q^*, and the uniform distribution of UEs, $U(N_{\text{min}}, N_{\text{max}})$, are described as pairs $\{q^*, U(N_{\text{min}}, N_{\text{max}})\}$, and the conditions are set to $\{1\text{Mbps}, U(5, 300)\}$, $\{3\text{Mbps}, U(5, 150)\}$, and $\{5\text{Mbps}, U(5, 100)\}$. The maximum number of UEs for each target throughput is different because each antenna has a fixed bandwidth, i.e., the number of RBs. Thus, increasing the target throughput reduces the number of UEs that can be accommodated. The simulation is conducted on a server with an NVIDIA A100 GPU and Xeon Gold 6326 CPU.

B. Performance Metrics

We evaluate our proposed method using four performance metrics. These are described below. Note that in this evaluation, the metrics omit time t as a variable since each episode consists of a single step.

The first metric is the proportion of episodes with constraint violations across K, denoted as V_K. An episode with a constraint violation means at least one UE whose throughput does not meet the target value. The second metric is the average power consumption rate, denoted as $\langle P_c \rangle_K$, calculated by averaging the power consumption rates from each K episode. The power consumption rate, denoted as P_c, is formulated as $P_c = 1 - P_r(t) = P_{\text{total}}(t)/\tilde{P}(t)$. This metric can be reduced as the number of antennas increases in this evaluation setting. Since the power consumption values in the real environment depend on the specifications of each machine that makes up the BSs, we focus on understanding the impact in terms of percentages rather than absolute values. The third metric is $\langle Q \rangle_K$, calculated by using the average throughput for UEs, $Q(t)$, in each episode and then averaging these values across K episodes. The last metric is denoted as $\langle \Delta q \rangle_K$, which is represented by calculating the standard deviation of the UE throughputs in each episode as $\Delta q = \sqrt{\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}(q_i^{(t)} - Q(t))^2}$, and then averaging these values across K episodes.

C. Baseline Methods

We compare four other methods with our proposed method (described as "SafeRO").

All Active (AA): This method does not execute sleep management, i.e., it operates all antennas in an active state.

Reward-Optimizing policy (RO): This is the control by the reward-optimizing agent in the first stage of SafeRO, which does not consider throughput constraints. The same hyperparameters as those used in SafeRO are employed.

RO + Manual Operation (RO+MO): After executing control with RO, if there are any constraint violations, that is, if the throughput for any UE is below the target value, the operation to increase the number of active antennas by one is performed.

Exhaustive Search (ES): By conducting an exhaustive search, this method executes the action that maximizes the reward among those actions that result in zero constraint violations. This method shows the optimal solutions. Note that this involves exploring by executing all actions, which is not feasible in actual control.

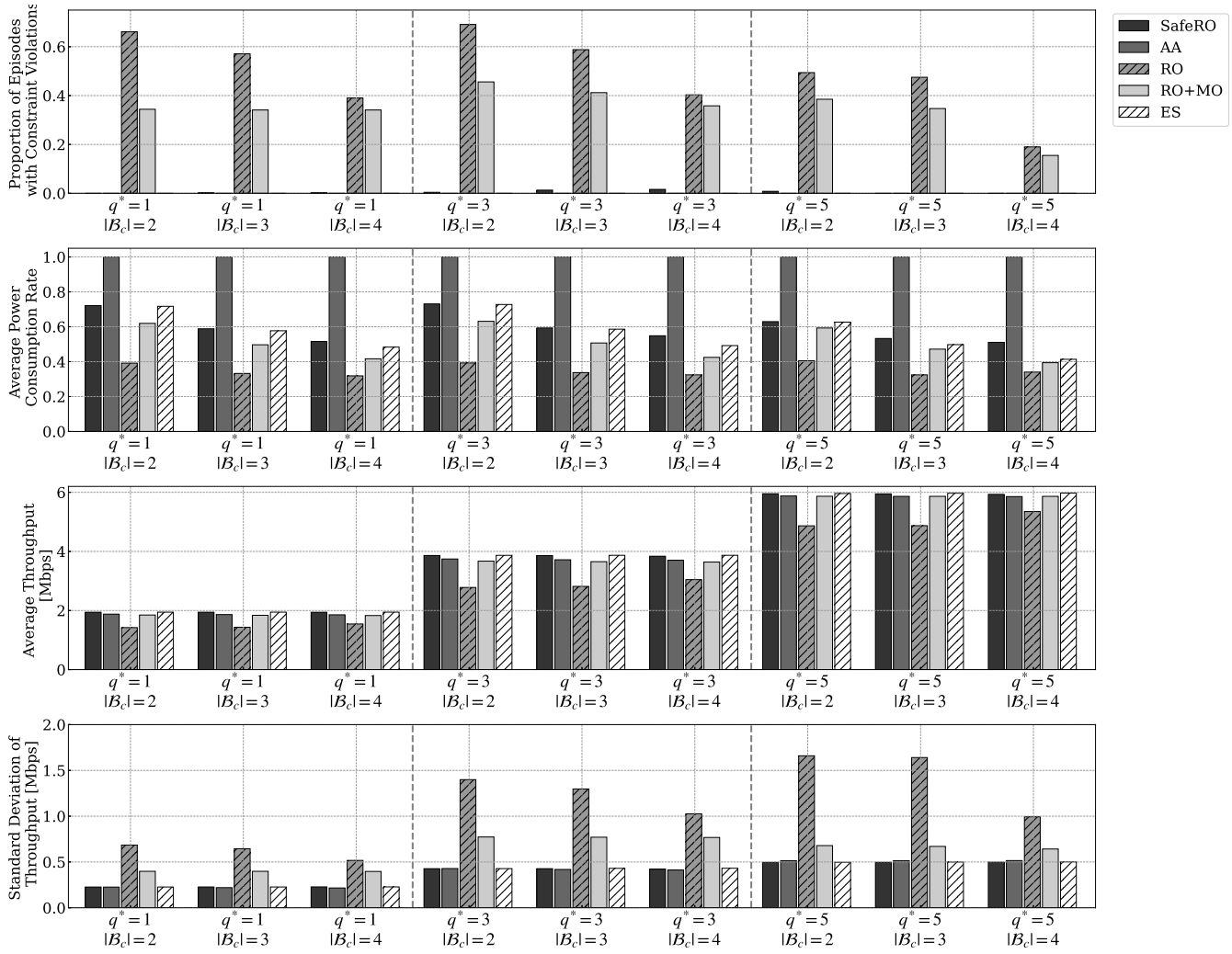

Fig. 3. Results of the comparative evaluation under the conditions that $|\mathcal{B}_c| = 2, 3, 4$, and the target throughput set to 1, 3, and 5 [Mbps].

D. Comparative Evaluation Results and Discussion

The evaluation results from 1000 episodes with varying numbers of UEs and their location are represented in Fig. 3. The evaluations are conducted for nine combinations of conditions, with $|\mathcal{B}_c| = 2, 3, 4$ antennas and target throughputs of 1, 3, and 5 Mbps, respectively. The results are analyzed using the four metrics.

First, the results for the proportion of episodes with constraint violations, V_{1000}, and the average power consumption rate, $\langle P_c \rangle_{1000}$, are presented in the first and second rows in Fig. 3. For AA, as all antennas are active, naturally, V_{1000} is 0, and $\langle P_c \rangle_{1000}$ is 1. In contrast, ES selects actions that result in zero constraint violations and then optimizes for reward; hence, V_{1000} is 0, and $\langle P_c \rangle_{1000}$ represents the optimal minimum power consumption achievable without any violations. Since RO prioritizes power reduction, $\langle P_c \rangle_{1000}$ is the lowest rate, which indicates that RO reduces power more effectively than AA. On the other hand, V_{1000} is the highest due to reducing power too much. The total number of RBs decreases when the number of active antennas is reduced for power saving. Hence, the throughput of more than 20 to 60 percent of UEs is below the target value, depending on the conditions. RO+MO adds one more active antenna if there is a constraint violation from the results of control by RO. Although RO+MO consumes more power than RO, that is, $\langle P_c \rangle_{1000}$ increase, it partially improves the violation episodes, i.e., V_{1000} is reduced. However, V_{1000} is still about 20 percent, which indicates that more than two active antennas need to be added for some episodes of the RO results. Furthermore, the number of lacking RBs is higher when the target throughput is 5 Mbps than when it is 1 Mbps, as more RBs must be allocated to the UEs. Thus, the higher the target throughput, the smaller the difference between V_{1000} for RO and RO+MO. In other words, the higher the target the throughput, the smaller the improvement in V_{1000} when one active antenna is added. SafeRO has fewer violation episodes than RO and RO+MO, with V_{1000} around 0 to 1.7 percent in all conditions. As for the power consumption, SafeRO is almost always close to the optimal value for ES, although in some cases, it consumes slightly more power than

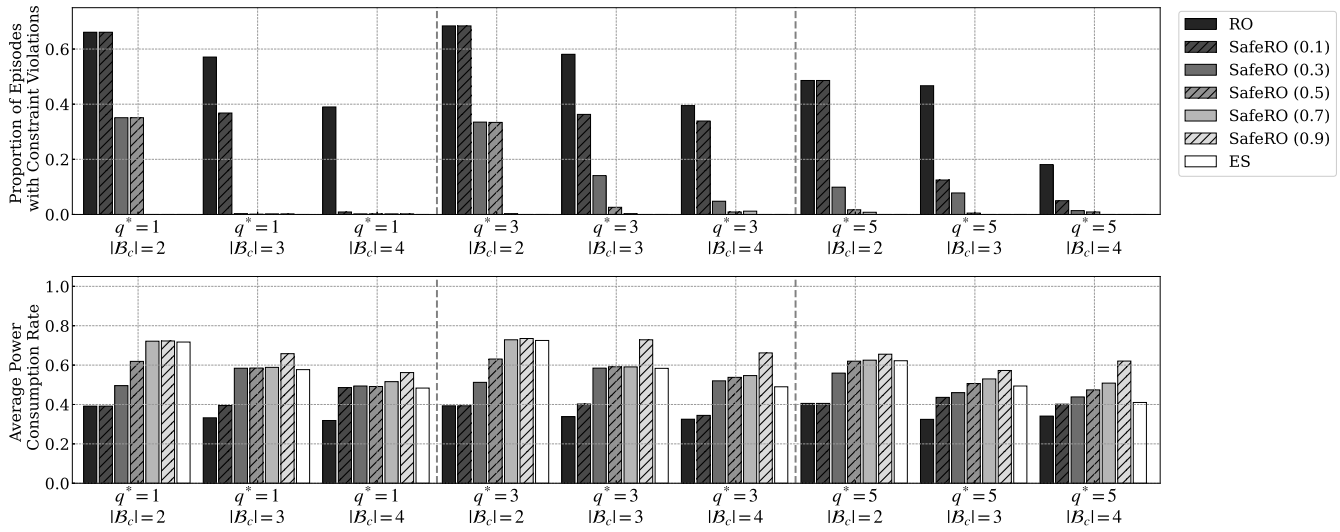

Fig. 4. Results of the parameter analysis for w_2 under the conditions that $|\mathcal{B}_c| = 2, 3, 4$, and the target throughput set to 1, 3, and 5 [Mbps].

ES. This result indicates that SafeRO appropriately increased the active antennas in the violation episode on the basis of the results of the first stage, RO. Although this increases the power consumption, the fact that it is close to the optimal value of the ES indicates that SafeRO appropriately controls without consuming power unnecessarily.

Second, the third and fourth rows in Fig. 3 present the results for the K-episode average of the mean throughput, $\langle Q \rangle_{1000}$, and the K-episode average of the standard deviation of throughput, $\langle \Delta q \rangle_{1000}$. The $\langle Q \rangle_{1000}$ and $\langle \Delta q \rangle_{1000}$ results are almost identical for AA, SafeRO, and ES. This slight difference is due to the different associations between the UE and the antenna in the evaluation environment of each method. The throughput changes when the association between the UE and the antenna changes because the frequency bands of the coverage and capacity antenna differ. For RO, $\langle Q \rangle_{1000}$ has the lowest value, and $\langle \Delta q \rangle_{1000}$ has the highest value. This corresponds to the high number of violating episodes and results from RO prioritizing power reduction. Moreover, this result indicates that more than a certain number of UEs with constraint violations are included in the violation episodes. RO prioritizes power reduction without considering constraints, so only in RO are some conditions where even the average value of mean throughput, $\langle Q \rangle_{1000}$, does not achieve the target throughput. Similarly, but not as much as RO, RO+MO has a slightly lower $\langle Q \rangle_{1000}$ and a higher $\langle \Delta q \rangle_{1000}$ because it includes a certain number of UEs with constraint violations. The results of $\langle Q \rangle_{1000}$ are approximately 2, 4, and 6 Mbps when the target throughput is 1, 3, and 5 Mbps. The reason for this is that in the setting of the network environment treated in this simulation, there is an average capacity of about 2 Mbps for the allocations of a single RB.

All SafeRO results are close to the ES results, which indicate optimal values. This suggests that the proposed SafeRO has been appropriately trained and executed with control rules in accordance with our intentions to reduce power consumption while avoiding the constraint violations that the UEs' throughput is below the target value. The proposed method also significantly reduces constraint violations compared to RO, which is a constraint-insensitive method based on existing research. This indicates that the constraints were adhered to during the second stage DRL. On the other hand, the power consumption was comparable to that of ES, which indicates that the proposed method not only satisfied the constraints but also considered solutions that contributed to the reduction in power consumption obtained by the first stage DRL.

E. Parameter Analysis

We compare the proposed methods using different w_2, a crucial parameter defining the weight of adherence to throughput constraint; the larger it is, the stricter the decision on the constraint. RO and ES are used as benchmarks. Our models are trained with w_2 set to $0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. Note that w_2 is set to 0.7 in the previous section IV-D.

Figure 4 shows the evaluation results of 1000 episodes with varying numbers of UEs and their location, using the same nine evaluation conditions and two critical evaluation metrics (V_{1000} and $\langle P_c \rangle_{1000}$) as in the comparative evaluation. For each evaluation metric, a smaller w_2 of safeRO tends to bring the results closer to RO, while a larger w_2 tends to reduce constraint violations more. Figure 4 shows that when the number of controlled antennas is two, the number of violating episodes approaches zero at w_2 above 0.5, whereas when the number of controlled antennas is four, the number of violating episodes is almost zero with w_2 of 0.3. This indicates that w_2 needs to be set larger when there are fewer target antennas than when there are more. This is because the cost of adding one antenna is higher when the number of controlled antennas is small due to the design of the cost of similarity, as shown in Fig. 5. The results are similar in most conditions when w_2 is more significant than 0.5. However, under conditions with four

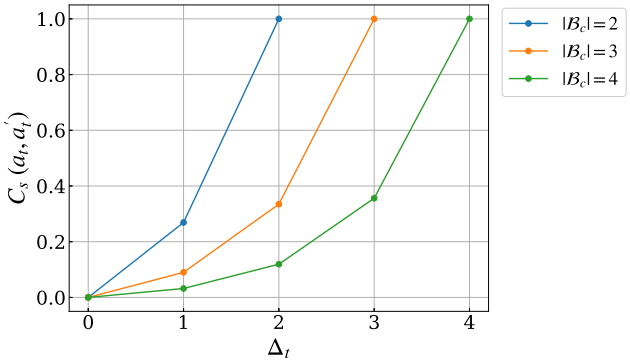

Fig. 5. The cost indicating the similarity between the original action and the modified safe action in the number of controlled antennas set to 2, 3, and 4.

controlled antennas and a target throughput of 5 Mbps, $w2$ of 0.3 is optimal, and above 0.5, the power consumption is larger than in the ES. Thus, set w_2 needs to be set appropriately in accordance with the number of controlled antennas and conditions when using the proposed method.

V. Conclusion

This paper proposed an antenna sleep management method that reduces power consumption while maintaining the target throughput, aiming to reduce CO2 emissions. Our method uses a two-stage deep reinforcement learning (DRL) approach to automatically search for and implement the appropriate rules. Comparative evaluations showed that our method significantly reduced throughput constraint violations and minimized power consumption compared to baseline methods. Parameter analysis of the cost function weights clarified the relationship between the number of controlled antennas and the parameters, and the setting of the parameters was discussed. We need to evaluate the proposed method over more extended control periods in environments with varying traffic utilization, and to extend the method to scenarios involving multiple base stations (BSs). The practical applicability of our approach also needs to be further examined by evaluating the time required for training and control execution, as well as through simulations that closely resemble real-world deployments or in actual environments.

References

[1] R. Friedrich, S. Hoffmann, T. Lampe and S. Ullrich, "Putting Sustainability at the Top of the Telco Agenda," bcg.com. https://www.bcg.com/publications/2021/building-sustainable-telecommunications-companies. (accessed May 9, 2024).

[2] O. Alamu, A. Gbenga-Ilori, M. Adelabu, A. Imoize and O. Ladipo, "Energy efficiency techniques in ultra-dense wireless heterogeneous networks: An overview and outlook," Engineering Science and Technology, an International Journal, vol. 23, no. 6, pp. 1308-1326, 2020.

[3] J. Wu, Y. Zhang, M. Zukerman and E. K. -N. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey," in IEEE Communications Surveys & Tutorials, vol. 17, no. 2, pp. 803-826, Secondquarter 2015.

[4] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru and Y. Tassa "Safe exploration in continuous action spaces," arXiv preprint arXiv:1801.08757, 2018.

[5] M. Feng, S. Mao and T. Jiang, "Base Station ON-OFF Switching in 5G Wireless Networks: Approaches and Challenges," in IEEE Wireless Communications, vol. 24, no. 4, pp. 46-54, Aug. 2017.

[6] D. López-Pérez, A. De Domenico, N. Piovesan, G. Xinli, H. Bao, S. Qitao and M. Debbah, "A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning," in IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 653-697, Firstquarter 2022.

[7] J. Wu, S. Zhou and Z. Niu, "Traffic-Aware Base Station Sleeping Control and Power Matching for Energy-Delay Tradeoffs in Green Cellular Networks," in IEEE Transactions on Wireless Communications, vol. 12, no. 8, pp. 4196-4209, August 2013

[8] Q. Wang and J. Zheng, "A Distributed base station On/Off Control Mechanism for energy efficiency of small cell networks," 2015 IEEE International Conference on Communications (ICC), 2015, pp. 3317-3322.

[9] X. Guo, Z. Niu, S. Zhou and P. R. Kumar, "Delay-Constrained Energy-Optimal Base Station Sleeping Control," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 5, pp. 1073-1085, May 2016.

[10] D. Renga, Z. Umar and M. Meo, "Trading Off Delay and Energy Saving Through Advanced Sleep Modes in 5G RANs," in IEEE Transactions on Wireless Communications, vol. 22, no. 11, pp. 7172-7184, Nov. 2023.

[11] J. Liu, B. Krishnamachari, S. Zhou and Z. Niu, "DeepNap: Data-Driven Base Station Sleeping Operations Through Deep Reinforcement Learning," in IEEE Internet of Things Journal, vol. 5, no. 6, pp. 4273-4282, Dec. 2018.

[12] J. Ye and Y. A. Zhang, "DRAG: Deep Reinforcement Learning Based Base Station Activation in Heterogeneous Networks," in IEEE Transactions on Mobile Computing, vol. 19, no. 9, pp. 2076-2087, Sept. 2020.

[13] F. E. Salem, T. Chahed, Z. Altman and A. Gati, "Traffic-aware Advanced Sleep Modes management in 5G networks," 2019 IEEE Wireless Communications and Networking Conference (WCNC), 2019, pp. 1-6.

[14] S. Lin et al., "DADEs: 5G Dual-Adaptive Delay-aware and Energy-saving System with Tandem Learning," 2022 IEEE Global Communications Conference (GLOBECOM), 2022, pp. 1-6.

[15] S. Malta, P. Pinto and M. Fernández-Veiga, "Using Reinforcement Learning to Reduce Energy Consumption of Ultra-Dense Networks With 5G Use Cases Requirements," in IEEE Access, vol. 11, pp. 5417-5428, 2023.

[16] R. Li, Z. Zhao, X. Chen, J. Palicot and H. Zhang, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," in IEEE Transactions on Wireless Communications, vol. 13, no. 4, pp. 2000-2011, April 2014.

[17] Q. Wu, X. Chen, Z. Zhou, L. Chen and J. Zhang, "Deep Reinforcement Learning With Spatio-Temporal Traffic Forecasting for Data-Driven Base Station Sleep Control," in IEEE/ACM Transactions on Networking, vol. 29, no. 2, pp. 935-948, April 2021.

[18] A. I. Abubakar, M. S. Mollel and N. Ramzan, "FAMAC: A Federated Assisted Modified Actor-Critic Framework for Secured Energy Saving in 5G and Beyond Networks," arXiv preprint arXiv:2311.14509, 2023.

[19] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang and A. KnollShangding, "A Review of Safe Reinforcement Learning: Methods, Theory and Applications," arXiv preprint arXiv:2205.10330, 2022.

[20] H. Li, Z. Wan and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," in IEEE Transactions on Smart Grid, vol. 11, no. 3, pp. 2427-2439, 2019.

[21] A. Suzuki and S. Harada, "Safe Multi-Agent Deep Reinforcement Learning for Dynamic Virtual Network Allocation," 2020 IEEE Global Communications Conference (GLOBECOM), 2020, pp. 1-7.

[22] Q. Liu, N. Choi and T. Han, "Constraint-Aware Deep Reinforcement Learning for End-to-End Resource Orchestration in Mobile Networks," 2021 IEEE 29th International Conference on Network Protocols (ICNP), 2021, pp. 1-11.

[23] A. Goldsmith, "Wireless Communications," Cambridge, U.K.: Cambridge Univ. Press, 2005.

[24] P. Christodoulou, "Soft actor-critic for discrete action settings," arXiv preprint arXiv:1910.07207, 2019.

[25] Hasselt, H. V, "Double Q-learning. In Advances in Neural Information Processing Systems, pp. 2613–2621, 2010.