Low Risk Antenna Configurations for Mobile Communication Systems: A Safe Reinforcement Learning Method

Yifang Zhang and Shaowei Wang

***Abstract*—Reinforcement learning offers an effective framework for antenna angle setting since it enables the autonomous and adaptive tuning of antenna parameters based on continuous interaction with the environment. However, due to the inherent trial-and-error nature of reinforcement learning, the agent may execute unacceptable decisions that reduce the network coverage and result in unbearable network performance degradation. In this paper, we propose a safe reinforcement learning (SRL) method to ensure that the policies made by agents can always maintain network valid coverage ratio above a threshold. The optimization task is formulated as a finite-horizon constrained Markov decision process and a confidence ball is introduced to limit the search scope within a safe range. Numerical results show that our proposed method provides an efficient and low risk scheme for antenna configuration in practical scenarios.**

***Index Terms*—Mobile networks, radio parameter optimization, safe reinforcement learning.**

I. Introduction

The fifth-generation (5G) communication system is faced with an explosive increase in mobile data traffic and massive user equipment connections, bringing about high demand for wireless network optimization which can enhance the network throughput at a low cost [1, 2]. The adjustment of antenna angles plays an important role in enhancing the service quality of the 5G network, which aims at alleviating the interference of adjacent cells so as to improve the signal-to-interference ratio of the service area. Besides, due to the advent of beamforming, focused beams directed towards specific location can be created, minimizing interference from unwanted directions [3, 4]. Traditionally, the antenna configuration is based on the experience of engineers and requires skilled personnel to manually adjust each antenna, which is time-consuming and prone to human error. Therefore, the automated antenna angle configuration is becoming increasingly essential for improving the efficiency and reliability of 5G network optimization.

The traditional optimization method, such as convex optimization method, is a feasible method for antenna angle adjustment. However, antenna angle setting involves a multitude of variables that are often nonlinear and time-varying. Additionally, the traditional methods may not efficiently adapt to changing environmental conditions based on simplified propagation models. Conversely, reinforcement learning (RL) can operate on top of high-precision simulation platforms, enabling direct learning of optimal antenna configurations from high-fidelity data [5]. By leveraging reinforcement learning algorithms, the network, acting as an agent, learns from the interactions with the environment and makes autonomous adjustments to antenna angles based on real-time feedback [6]. In [7], a fuzzy reinforcement learning method was proposed to optimize the antenna downtilt for single-tier networks, which operates in a distributed and autonomous manner without requiring prior information or human interventions. In [8], a reinforcement learning algorithm was improved by allowing different antennas to be optimized simultaneously, increasing the power coverage ratio and reducing the interference among adjacent antennas.

However, due to the inherent trial-and-error nature of reinforcement learning, the agent may make unacceptable decisions during the exploration process that results in great network performance degradation. Consequently, despite exhibiting effectiveness in simulation environments, general reinforcement learning is unsuitable for antenna configuration in real-world scenarios. During the real-time adjustment of the antenna, a low instantaneous valid coverage in the network may lead to a decline in user data rates and impact the user experience in viewing videos and images. Therefore, it is of great importance to safely tune the antenna angle, that is, to avoid making decisions that greatly reduce the network coverage during antenna angle optimization process. Safe reinforcement learning (SRL) is an alternative solution to this problem which aims at ensuring the safety and reliability of the policy [9]. Usually, the safety of strategy is achieved by incorporating cost or risk constraints into the learning process. In [10], the SRL task was formulated as a constrained Markov decision process and solved by a Lyapunov approach, which guarantees the feasibility and optimality of the algorithm under certain conditions. In [11], a constrained policy optimization method that extends trust-region policy optimization was developed to handle the constraints in SRL.

Fundamentally speaking, safety is a bottleneck of applying reinforcement learning to antenna configuration in real application scenarios and SRL arises as an effective approach to address this issue. In this paper, we propose an SRL-based method to configure the antenna angle settings in a safe and reliable way. First, we formulate the antenna configuration problem as a constrained Markov decision process, where

Manuscript received 5 March 2024; accepted 8 April 2024. This work was partially supported by the National Natural Science Foundation of China under Grants 61931023. The associate editor coordinating the review of this letter and approving it for publication was H. J. Yang. *(Corresponding author: Shaowei Wang.)*

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: mg21230076@smail.nju.edu.cn, wangsw@nju.edu.cn).

the cost is defined as the potential degradation in network performance and subject to a constraint. Given a safe policy provided by experts, we construct two confidence sets in the reward space and cost space respectively, which are then combined as the confidence ball of exploration space. Next, the constrained Markov decision process is addressed within the confidence ball by a linear programming approach, which gives the probability distribution of the strategy. The procedure is executed iteratively until reaching the maximum episode length. Numerical results verify that our method can produce promising antenna configurations while ensuring the safety during the exploration, rendering it a suitable way for real-world antenna configuration scenarios.

II. System Model and Problem Formulation

Consider a region $\mathcal{G} \in \mathbb{R}^2$ served by a set of base stations with directional antennas. The users $\mathcal{U} = \{u_1, u_2, \cdots, u_N\}$ are distributed uniformly in the area. Note that in this paper, vectors are represented using boldface type to distinguish them from other quantities. Denote the set of antennas by $\boldsymbol{\xi} = \{\xi_1, \xi_2, \cdots, \xi_M\}$. As illustrated in Fig. 1, azimuth θ is defined as the angle between the horizontal projection of the antenna main lobe direction and the x-axis. Downtilt ϕ is defined as the angle between the antenna main lobe direction and the horizontal plane. Let $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_M\}$ and $\boldsymbol{\phi} = \{\phi_1, \phi_2, \cdots, \phi_M\}$ represent the azimuth set and downtilt set of all antennas respectively. Let θ_i and ϕ_i denote the azimuth and downtilt of antenna i respectively.

Denote $P_{i,j}$ with $1 \le i \le M, 1 \le j \le N$ as the power of reference signal received from antenna ξ_i by user u_j, which is given by

$$P_{i,j}(\boldsymbol{\theta}, \boldsymbol{\phi})|_{dBm} = P^T|_{dBm} + G_{i,j}(\boldsymbol{\theta}, \boldsymbol{\phi}) + G^F - L_{i,j}, \quad (1)$$

where P^T is the transmit power, $G_{i,j}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is the directional antenna gain and G^F is the terminal gain. $L_{i,j}$ denotes the path loss between antenna a_i and user u_j. Generally, the user is served by the antenna that provides the strongest signal, while the signals from other antennas are regarded as interference. Therefore, we define the maximal useful signal power of user u_j as

$$P_j(\boldsymbol{\theta}, \boldsymbol{\phi}) = \max_{1 \le i \le M} P_{i,j}(\theta_i, \phi_i). \quad (2)$$

The corresponding signal-to-interference ratio (SIR) can be written as follows:

$$S_j(\boldsymbol{\theta}, \boldsymbol{\phi}) = 10 \log_{10} \frac{P_j(\boldsymbol{\theta}, \boldsymbol{\phi})}{\sum_{\xi_i \in \boldsymbol{\xi}} P_{i,j}(\theta_i, \phi_i) - P_j(\boldsymbol{\theta}, \boldsymbol{\phi})}. \quad (3)$$

Define two sets $\mathcal{U}^P$ and $\mathcal{U}^S$. If the maximal useful signal power $P_j(\boldsymbol{\theta}, \boldsymbol{\phi})$ of u_j meets a predefined threshold σ^p, then $u_j \in \mathcal{U}^P$, i. e.

$$\mathcal{U}^P = \{u_j | P_j(\boldsymbol{\theta}, \boldsymbol{\phi}) > \sigma^p\}. \quad (4)$$

Similarly, if the signal-to-interference ratio $S_j(\boldsymbol{\theta}, \boldsymbol{\phi})$ of u_j exceeds a predefined threshold σ^s, then $u_j \in \mathcal{U}^S$, i. e.

$$\mathcal{U}^S = \{u_j | S_j(\boldsymbol{\theta}, \boldsymbol{\phi}) > \sigma^s\}. \quad (5)$$

Let the power coverage indicator $\gamma^p(\boldsymbol{\theta}, \boldsymbol{\phi})$ be the ratio of users that meet the power threshold σ^p, which can be defined as $\gamma^p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{|\mathcal{U}^P|}{|\mathcal{U}|}$,where $|\mathcal{U}^P|$ and $|\mathcal{U}|$ are the cardinality of the set $\mathcal{U}^P$ and $\mathcal{U}$, respectively. The capacity coverage indicator $\gamma^s(\boldsymbol{\theta}, \boldsymbol{\phi})$ is defined as $\gamma^s(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{|\mathcal{U}^S|}{|\mathcal{U}|}$,where $|\mathcal{U}^S|$ is the cardinality of the set $\mathcal{U}^S$.

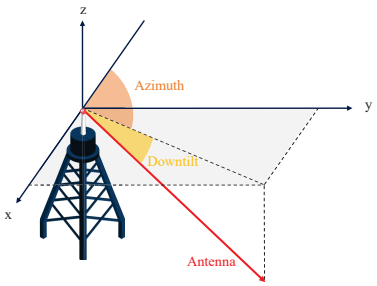

Fig. 1. Azimuth and downtilt of the antenna, where the red arrow represents the antenna.

The objective of network optimization is to maximize both the network power coverage and the capacity coverage, thus we introduce a performance indicator named valid coverage ratio written as follows:

$$\gamma^v(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{\sum_{u_j \in \mathcal{U}} \min(\mathbb{1}(P_j(\boldsymbol{\theta}, \boldsymbol{\phi}) > \sigma^p), \mathbb{1}(S_j(\boldsymbol{\theta}, \boldsymbol{\phi}) > \sigma^s))}{|\mathcal{U}|}, \quad (6)$$

where $\mathbb{1}(x)$ is an indicator expressed as

$$\mathbb{1}(x) = \begin{cases} 1 & , \text{ if } x \textit{ is } \textbf{true}, \\ 0 & , \text{ otherwise.} \end{cases} \quad (7)$$

$\gamma^v(\boldsymbol{\theta}, \boldsymbol{\phi})$ measures the ratio of users whose signal power and SIR are both above their respective thresholds.

With the azimuth set $\boldsymbol{\theta}$ and the downtilt set $\boldsymbol{\phi}$, the optimization problem can be stated as

$$\begin{aligned} \max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \quad & \gamma^v(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ \text{s.t.} \quad & \theta^L \le \theta_i \le \theta^H, \qquad 1 \le i \le M \\ & \phi^L \le \phi_i \le \phi^H, \qquad 1 \le i \le M, \end{aligned} \quad (8)$$

where θ^L and θ^H are the lower bound and upper bound of azimuth, respectively. ϕ^L and ϕ^H denote the lower bound and upper bound of downtilt, respectively.

III. Proposed Algorithm

A. Safe Reinforcement Learning Framework

Antenna adjustment is a series of operations in which the decision to adjust the antenna angle is a sequential process with a finite set of actions, and the outcome of each adjustment depends on the current state and the chosen action. Therefore, the antenna angle optimization task can be modeled as a Markov decision process. In the context of reinforcement learning, antennas are regarded as agents and iteratively interact with the environment. Due to the trial-and-error nature of the traditional reinforcement learning, it is inevitable for agents to make unsafe decisions that may cause immediate network failure. To cope with this, we introduce a safe reinforcement

learning framework with cost constraint that restricts the exploration of risky actions so as to ensure the safety during the exploration. Specifically, we consider a constrained Markov decision process denoted as $M = \langle \mathcal{S}, \mathcal{A}, r, c, P, H, \bar{C} \rangle$ during antenna tuning, where

- The state space $\mathcal{S}$ corresponds to the set of antenna locations.
- The action space $\mathcal{A} = \{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ indicates the candidate angle settings of antennas.
- The reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the network performance after executing a policy, which is defined as the valid coverage ratio $\gamma^v(\boldsymbol{\theta}, \boldsymbol{\phi})$.
- The cost $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ indicates the network performance degradation that the chosen policy yields, which is defined as $c(s,a) = e^{\frac{1-\gamma^v(\boldsymbol{\theta},\boldsymbol{\phi})}{\epsilon}}$, where ϵ is a risk threshold.
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability function with $P(s'|a,s)$ indicates the transition probability from state s to s' under action a.
- H is the number of consecutive actions performed in each episode.
- The constraint $\bar{C}$ is a scalar that specifies the maximum allowable value of the cost.

A policy distribution $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ specifies the probability of choosing an action a when the state is s. The value function $V_r^\pi(s)$ of the policy π given a state s can be expressed as follows:

$$V_r^\pi(s) = \mathbb{E}[\sum_{h=1}^{H} r(s_h, a_h) | s_1 = s], \tag{9}$$

where $\mathbb{E}$ represents expectation, h is the step in the current episode, $r(s_h, a_h)$ is the reward under state s_h and action a_h. The cost function $V_c^\pi(s)$ of the policy π can be defined similarly:

$$V_c^\pi(s) = \mathbb{E}[\sum_{h=1}^{H} c(s_h, a_h) | s_1 = s], \tag{10}$$

where $c(s_h, a_h)$ is the cost under state s_h and action a_h. Consequently, the safe exploration problem can then be formulated as

$$\max_{\pi} \quad V_r^\pi(s) \qquad \text{s.t.} \quad \mathrm{V}_c^\pi(\mathrm{s}) \le \bar{\mathrm{C}}. \tag{11}$$

Note that we define a policy π as a safe policy if its cumulative cost is less than the maximum allowable value $\bar{C}$, i.e. $V_c^\pi(s) \le \bar{C}$. Let $\Pi_{\text{safe}} = \{\pi : V_c^\pi(s) \le \bar{C}\}$ denote the set of safe policies. Assume that we have access to a safe baseline policy $\pi_b \in \Pi_{\text{safe}}$ that satisfies $V_c^{\pi_b}(s) \le \bar{C}$ before applying the safe reinforcement learning method to the online antenna tuning. The safe policy π_b is then utilized as the initial strategy for safe exploration of antenna configurations.

B. Safe Exploration of Angle settings

Without loss of generality, the agent interacts with the environment episodically. Let K denote the number of total episodes and $\pi_k = (\pi_{h,k})_{h=1}^H$ be the policy performed in episode k. In each step $h \in [H]$ of episode $k \in [K]$, the agent selects an action $a_{h,k}$ based on the policy $\pi_{h,k}(s_{h,k})$. For the rest of the paper, we assume that the step length H of each episode is 1, and omit the subscript h. Let $n_k(s,a)$ denote the number of times the state-action pair (s,a) was observed before episode k, which can be expressed as

$$n_k(s,a) = \sum_{k'=1}^{k-1} \mathbb{1}\{s_{k'} = s, a_{k'} = a\}. \tag{12}$$

At the beginning of the episode k, the reward of each state-action pair is estimated by

$$\hat{r}_k(s,a) = \frac{\sum_{k'=1}^{k-1} r(s,a)\mathbb{1}\{s_{k'} = s, a_{k'} = a\}}{\max\{n_k(s,a), 1\}}. \tag{13}$$

Similarly, the cost of the state-action pair (s,a) is estimated by

$$\hat{c}_k(s,a) = \frac{\sum_{k'=1}^{k-1} c(s,a)\mathbb{1}\{s_{k'} = s, a_{k'} = a\}}{\max\{n_k(s,a), 1\}}. \tag{14}$$

Inspired by the optimism in the face of uncertainty style algorithms [12], we construct confidence sets $\mathcal{R}_k$ around $\hat{r}_k$ and $\mathcal{C}_k$ around $\hat{c}_k$ respectively:

$$\mathcal{R}_k = \{\tilde{r} : |\tilde{r}_k(s,a) - \hat{r}_k(s,a)| \le \beta_k^r(s,a), \forall s,a \in \mathcal{S} \times \mathcal{A}\}, \tag{15}$$

$$\mathcal{C}_k = \{\tilde{c} : |\tilde{c}_k(s,a) - \hat{c}_k(s,a)| \le \beta_k^c(s,a), \forall s,a \in \mathcal{S} \times \mathcal{A}\}, \tag{16}$$

where

$$\beta_k^r(s,a) = \sqrt{\frac{\lambda^r}{\max\{n_k(s,a), 1\}}}, \tag{17}$$

$$\beta_k^c(s,a) = \sqrt{\frac{\lambda^c}{\max\{n_k(s,a), 1\}}}, \tag{18}$$

where λ^r and λ^c are parameters determining the size of the confidence interval. The total confidence ball is defined as $\mathcal{M}_k = \mathcal{R}_k \bigcap \mathcal{C}_k$.

To further ensure the safety of the exploration, we modify the cost $\hat{c}_k(s,a)$ as

$$\bar{c}_k(s,a) = \hat{c}_k(s,a) + \beta_k^c(s,a), \tag{19}$$

which adds penalties of the less observed (s,a) pairs and inhibits their explorations. However, such a modification may prevent necessary explorations to learn optimal policies, thus we also modify the reward $\hat{r}_k(s,a)$ by adding a term to stimulate the exploration:

$$\bar{r}_k(s,a) = \hat{r}_k(s,a) + \frac{\beta_k^r(s,a)}{\bar{C} - \bar{C}_b}, \tag{20}$$

where $\bar{C}_b$ is the cost of the safe baseline policy π_b.

The safe exploration problem (11) can then be rewritten as

$$\max_{\pi} \quad V_{\bar{r}}^\pi(s) \qquad \text{s.t.} \quad \mathrm{V}_{\bar{\mathrm{c}}}^\pi(\mathrm{s}) \le \bar{\mathrm{C}}, \tag{21}$$

which can be reformulated to a linear programming problem. Firstly, let $w_k(s,a)$ be the occupancy measure indicating the

Algorithm 1 Safe exploration of antenna tuning

Input: $r, c, \pi_b, \bar{C}$
Initialization: $n_k(s,a) = 0, \forall s \in \mathcal{S}, a \in \mathcal{A}$
for $k = 1 : K$ **do**
  Estimate the $\hat{r}_k$ and $\hat{c}_k$ by Eq. (13) and Eq. (14)
  Calculate the reward confidence ball $\mathcal{R}_k$ by Eq. (15)
  Calculate the cost confidence ball $\mathcal{C}_k$ by Eq. (16)
  $\mathcal{M}_k = \mathcal{R}_k \cap \mathcal{C}_k$
  if $k == 1$ **then**
    Select the policy $\pi_k = \pi_b$
  else
    Solve the linear programming problem by Eq. (25)
    Select the policy π_k by Eq. (26)
  end if
  Select action $a \sim \pi_k(s, \cdot)$
  Incur the reward $r_k(s,a)$ and cost $c_k(s,a)$
  Update the counts: $n_k(s,a) \leftarrow n_k(s,a) + 1$
end for

probability of choosing action a under the policy π at the episode k, which is defined as

$$w_k(s,a) = \mathbb{E}[\mathbb{1}\{a_k = a\}|s_k = s, \pi] = \mathbb{P}(a_k = a|s_k = s, \pi). \tag{22}$$

Then the value function $V^{\pi}_{\bar{r}}(s)$ and cost function $V^{\pi}_{\bar{c}}(s)$ can be expressed by the occupancy measure as

$$V^{\pi}_{\bar{r}}(s) = \sum_{s,a} w^{\pi}_k(s,a)\bar{r}(s,a), \tag{23}$$

$$V^{\pi}_{\bar{c}}(s) = \sum_{s,a} w^{\pi}_k(s,a)\bar{c}(s,a). \tag{24}$$

The problem (21) can be rewritten as a linear programming accordingly:

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \sum_{s,a} w_k(s,a)\bar{r}(s,a) \\ \text{s.t.} \quad & C_1 : \sum_{s,a} w_k(s,a)\bar{c}(s,a) \le \bar{C} \\ & C_2 : \sum_{s,a} w_k(s,a) = 1, \forall s \in \mathcal{S} \\ & C_3 : w_k(s,a) \ge 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \end{aligned} \tag{25}$$

where the constraint C_1 guarantees that the cost function is less than the threshold $\bar{C}$. C_2 indicates the sum of probability is 1, and C_3 means that the occupancy measure should be greater than 0. After working out the optimal solution of (21), the policy for the antenna tuning problem can be computed as

$$\pi_k(s,a) = \frac{w^{\pi}_k(s,a)}{\sum_{\hat{a}\in\mathcal{A}} w^{\pi}_k(s,\hat{a})}. \tag{26}$$

Then an action is selected by policy $\pi_k(s,a)$, resulting in the reward $r_k(s,a)$ and the cost $c_k(s,a)$. The counting indicator is also updated by $n_k(s,a) = n_k(s,a) + 1$. Next, a new confidence ball $\mathcal{M}_{k+1}$ can be calculated encompassing the state-action pair (s,a), wherein the linear programming problem is solved to derive the subsequent policy. This process is repeated until reaching the episode length K, and the whole procedure is shown in Algorithm 1.

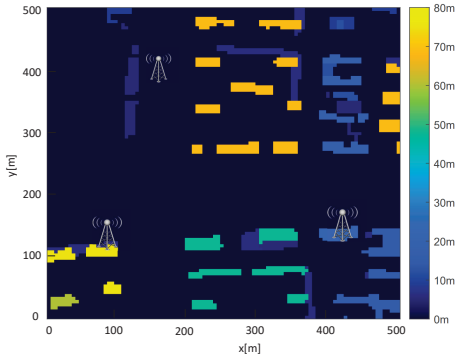

Fig. 2. Building layout and BS locations.

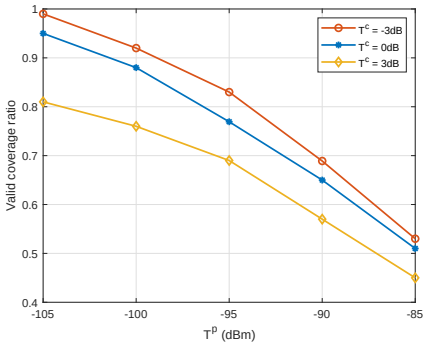

Fig. 3. Valid coverage ratio under different power thresholds and SIR thresholds.

IV. Numerical Experiment

Consider a geographical region of size 500m×500m with 3 base stations, as shown in Fig. 2. The transmit power P^T is set to 15.2dBm, the carrier frequency is 3.5GHz and the bandwidth is 100MHz. The range of electronic azimuth and downtilt are $[0°, 360°]$ and $[0°, 20°]$ with the interval of 5° and 2°, respectively. Note that using a continuous action space approach could result in an overly fine-grained control that may not align with the real world discrete antenna angle settings. Therefore, we adopt the discrete action space which provides a more practical and effective solution for the task. Additionally, We adopt the COST 231 Final model as the path loss model. Users are distributed randomly in the area with the constant height of 1.5m and the terminal gain of 0dBm. The episode length K is 400.

First, we investigate the performance of our proposal under different power and SIR thresholds, as shown in Fig. 3. It can be seen that when the SIR threshold T^c remains constant, the valid coverage ratio decreases as the power threshold T^p increases. Similarly, the valid coverage ratio decreases as the SIR threshold T^c increases while the power threshold T^p is kept constant. Additionally, it is noteworthy that the curve for higher SIR thresholds exhibits a slower decline compared to that of lower SIR thresholds, since the users meeting higher SIR thresholds are more likely to have stronger received signal strength under the same power threshold. Consequently, as the power threshold T^p increases, the number of users meeting the higher SIR threshold that become invalid is comparatively

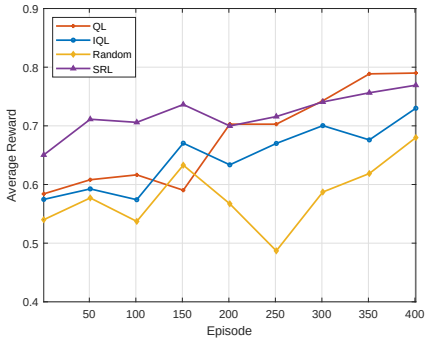

Fig. 4. Reward of different algorithms.

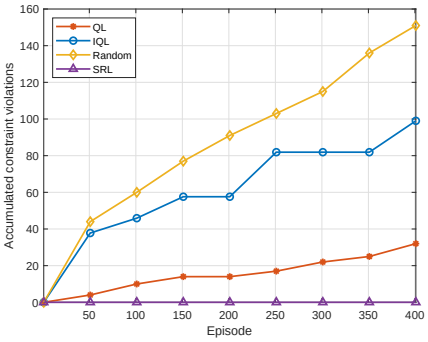

Fig. 5. Accumulated constraint violations of different algorithms.

smaller than those meeting the lower SIR threshold.

Then, we compare the network performance of our proposed method with the following three methods: the IQL method developed in [7], the QL method proposed in [8] and the random search method. The power threshold is set to -95dBm, the SIR threshold is 0dB and the cost constraint $\bar{C}$ is e. As shown in Fig. 4, at the end of the episode, SRL method yields a reward of 0.77, which is observed to be the second-best performing algorithm. The QL algorithm achieves the highest reward of 0.79 due to its capability of searching for solutions within a larger action space. The performance of the SRL is slightly inferior to that of the QL because the SRL explores only within the confines of the confidence ball, resulting in a conservative policy execution. Therefore, while the SRL has the potential to enhance the safety and reliability of reinforcement learning algorithms, it may also come at a slight cost in terms of performance.

Fig. 5 shows the accumulated constraint violations of different algorithms, which is defined as the accumulated number of times that the valid coverage ratio falls below 0.6. When the coverage ratio is less than 0.6, nearly half of users may experience dropped calls, poor call quality, and difficulty in maintaining a stable connection. The slow data speeds and intermittent connectivity can frustrate users trying to browse the internet or use other data-intensive applications. As can be seen from the figure, the SRL algorithm exhibits zero violations of constraints during the exploration process. Note that the number of accumulated constraint violations of the QL algorithm reaches 31 by episode 400, signifying severe network performance degradation during the exploration. A coverage rate below 60% implies that nearly 40% of users within the area cannot use their mobile phones for activities such as image browsing, significantly impacting user experience. Therefore, even if the valid coverage rate of the QL algorithm surpasses that of the SRL algorithm in the final stage of reinforcement learning, the substantial decrease in temporary coverage caused by the exploration phase of the QL algorithm can severely compromise the user experience at that moment. Consequently, based on the results from Fig. 4 and Fig. 5, it can be inferred that the SRL algorithm is a more feasible approach for practical antenna configuration scenarios.

V. Conclusion

In this paper, we investigated the reliable antenna angle tuning for mobile communication systems. We proposed a safe reinforcement learning based antenna angle adjustment algorithm aiming at maximizing the valid coverage ratio while avoiding significant degradation of network performance during the exploration. The antenna configuration task is formulated as a finite-horizon constrained Markov decision process, which is solved by a linear programming approach within the scope of a confidence ball. Numerical results show that our proposal can produce promising antenna configurations while ensuring safety during the exploration process, providing an effective and safe way for antenna tuning in practical scenarios.

References

[1] I. Ismath *et al.*, "Deep contextual bandits for fast neighbor-aided initial access in mmwave cell-free networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2752–2756, Dec. 2021.

[2] L. Shen and S. Wang, "Scalable antenna orientation optimization for mmwave mobile communication systems," in *Proc. IEEE GLOBECOM'23*, Kuala Lumpur, Malaysia, Dec. 2023.

[3] Z. Lin *et al.*, "Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6345–6360, Sept. 2021.

[4] ——, "Refracting ris-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, Aug. 2022.

[5] S. Marco *et al.*, "Cellular network capacity and coverage enhancement with MDT data and deep reinforcement learning," *Comput. Commun.*, vol. 195, no. 1, pp. 403–415, Sept. 2022.

[6] L. Shen, Y. Zhang, and S. Wang, "Codebook based antenna configuration: A new network planning paradigm for mmwave mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10 368–10 379, Mar. 2023.

[7] V. Buenestado *et al.*, "Self-tuning of remote electrical tilts based on call traces for coverage and capacity optimization in LTE," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4315–4326, May 2017.

[8] N. Islam and A. Mitschele, "Reinforcement learning strategies for self-organized coverage and capacity optimization," in *Proc. IEEE WCNC'12*, Paris, France, Apr. 2012.

[9] J. Garcia and F. Fernandez, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, Mar. 2015.

[10] Y. Chow *et al.*, "A Lyapunov-based approach to safe reinforcement learning," in *Proc. NeurIPS'18*, Montreal, Canada, Dec. 2018.

[11] J. Achiam *et al.*, "Constrained policy optimization," in *Proc. ICML'17*, Sydney, Australia, Aug. 2017.

[12] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Proc. NeurIPS'08*, Vancouver, Canada, Dec. 2008.