Scalable Antenna Orientation Optimization for mmWave Mobile Communication Systems

Linzhi Shen and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China.
Email: 141180100@smail.nju.edu.cn, wangsw@nju.edu.cn

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437672

***Abstract*—Tuning the azimuths and tilts of antennas in an optimal way is crucial for the mmWave mobile communication systems, nevertheless, it always yields an intractable combinatorial optimization problem due to the huge number of possible antenna configurations. In this paper, we propose a scalable reinforcement learning method to deal with this optimization task, which decomposes and distributes the policy learning process by exploiting the interactions among adjacent antennas implicitly. A distributed constraint optimization technique is introduced to coordinate the distributed learning process, which can lead the learning towards desired directions. Experiment results in large-scale scenarios demonstrate that our proposed method yields significant performance improvement in terms of signal-to-interference ratio coverage while ensuring high power coverage.**

***Index Terms*—Antenna orientation optimization, coordinated reinforcement learning, distributed constraint optimization, mmWave mobile communications.**

I. Introduction

As user demands persistently evolve, mobile network operators strive to improve spectral efficiency so as to accommodate the increasing traffic requirements. Radio network optimization serves as a crucial approach for augmenting spectral efficiency, where the power coverage and the capacity of the mobile network are optimized by adjusting parameters of base stations (BSs) [1, 2]. The power coverage refers to the area where signal is strong enough to ensure reliable communications, while the capacity relies on the signal-to-interference ratio (SIR) which is related to the strength of the reference signals and the interference ones. More specifically, the increase of transmission power may expand the power coverage of a BS, however, such a power increase also introduces more interference to the neighbouring BSs, and reduces the system capacity. Striking a balance between the two conflicting objectives is essential for radio network optimization.

The radio network optimization can be implemented in various ways, among which adjusting antenna orientation settings is of vital importance [3–5]. An advantageous set of orientation settings not only reduces zones with weak signal strength but also alleviates interference among BSs. Moreover, adjusting orientation settings is much more cost-effective compared to other approaches such as constructing new BSs or installing additional equipments.

This work was supported in part by the National Natural Science Foundation of China under Grants 61931023 and U1936202.

The fifth generation and beyond mobile communications introduce mmWave technology to acquire additional spectral resources [6], which, however, brings further challenges to the antenna orientation optimization. Specifically, mmWave signals experience more severe fading compared to signals in sub-6GHz bands due to the shorter wavelengths. To cope with the deep fading, more BSs are needed to serve a given area, which is known as network densification [7]. Moreover, high-directional antenna is adopted to concentrate power into a narrow beam, whose radiation pattern is more sensitive to the azimuth compared to the wide-beam antenna in sub-6GHz bands [8]. Therefore, antenna orientation optimization in an mmWave system usually involves adjusting the azimuths and the tilts of a large number of antennas.

Unfortunately, adjusting the azimuths and the tilts of antennas optimally always yields an intractable combinatorial optimization problem whose search space expands exponentially with the number of antennas. Brute-force search that may produce intuitive exact solution is computationally prohibitive even for a moderate number of antennas. Recently, reinforcement learning has merged as a promising approach for the antenna orientation optimization due to its remarkable adaptability across diverse network scenarios. In [9], the tilts of antennas are adjusted separately based on independent reinforcement learning. The limitation of this approach lies in the static configurations of adjacent antennas during the tuning process for a specific antenna. This overlooks the fact that the interference varies with the configurations of the adjacent antennas. In [10], a mean-field reinforcement learning approach is adopted to learn the cumulative interference from adjacent BSs, where all these BSs are treated as one virtual entity located at the center of these BSs. Such an approximation is oversimplified and may lead to imprecise interference estimation. In [11], a deep reinforcement learning method is proposed to optimize the configuration of antennas. This method is hard to implement in real-world scenarios owing to the substantial cost of training a reliable deep neural network.

These reinforcement learning methods ignore or simplify the interactions among antennas, which inevitably results in suboptimal even undesired antenna orientation configurations. Note that the interactions among antennas are always local since signal strength diminishes with distance rapidly and becomes negligible beyond a certain range, therefore a scalable algorithm which exploits these interactions can potentially achieve a good tradeoff between computational efficiency and

optimality of antenna configurations. In this paper, we propose a coordinated reinforcement learning method to handle the antenna orientation optimization problem in a scalable and effective manner. The proposed method factors a decomposable Q-function that quantifies the power coverage and the capacity by exploiting the interactions among antennas, and distributes the learning of policy for orientation configurations of all antennas, which potentially scales up the learning to large-scale scenarios. Furthermore, the proposed method employs distributed constraint optimization techniques to coordinate the distributed learning so as to ensure the optimality of the learned policy. In addition, a max-sum algorithm is proposed to work out an approximate solution to the distributed constraint optimization problem with slight computation overhead. Numerical results show that our proposed methods can produce promising configurations of antennas in large-scale scenarios with reasonable computing load, outperforming the state-of-the-art ones by over 9% in terms of area satisfying both power coverage requirements and SIR coverage requirements.

II. System Model and Problem Formulation

Consider an urban area served by multiple BSs, each of which containing three high-directional antennas. The set of antennas is represented by $\mathbb{N}$. For each antenna $n \in \mathbb{N}$, its azimuth θ_n and tilt ϕ_n are configured from predefined sets Θ_n and Φ_n, respectively. Let $\boldsymbol{\theta} = \{\theta_n\}_{n\in\mathbb{N}}$ and $\boldsymbol{\phi} = \{\phi_n\}_{n\in\mathbb{N}}$ denote the azimuth settings and the tilt settings of all antennas, respectively. The target area is divided into regular hexagon grids with side length of 5m, as shown in Fig. 1. The center of each grid is termed as a service test point (STP) [12], and the set of STPs is represented by $\mathbb{K} = \{1, 2, ..., K\}$.

Let $p_{n,k}(\theta_n, \phi_n)$ denote the power of reference signal received by STP $k \in \mathbb{K}$ from antenna $n \in \mathbb{N}$ [13]. Note that each STP is always associated with the antenna that provides the strongest average reference signal strength. The valid signal power received by k is defined as

$$p_k(\boldsymbol{\theta}, \boldsymbol{\phi}) = \max_{n\in\mathbb{N}} p_{n,k}(\theta_n, \phi_n), \tag{1}$$

and the SIR at k is

$$\rho_k(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{p_k(\boldsymbol{\theta}, \boldsymbol{\phi})}{\sum_{n\in\mathbb{N}} p_{n,k}(\theta_n, \phi_n) - p_k(\boldsymbol{\theta}, \boldsymbol{\phi})}. \tag{2}$$

A satisfied STP means that its received valid signal power p_k and the obtained SIR ρ_k are greater than the predefined thresholds T^P and T^C, respectively.

The antenna orientation optimization problem is to maximize the proportion of the satisfied STPs. Define an index variable x_k,

$$x_k = \begin{cases} 1, & \text{STP } k \text{ is satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

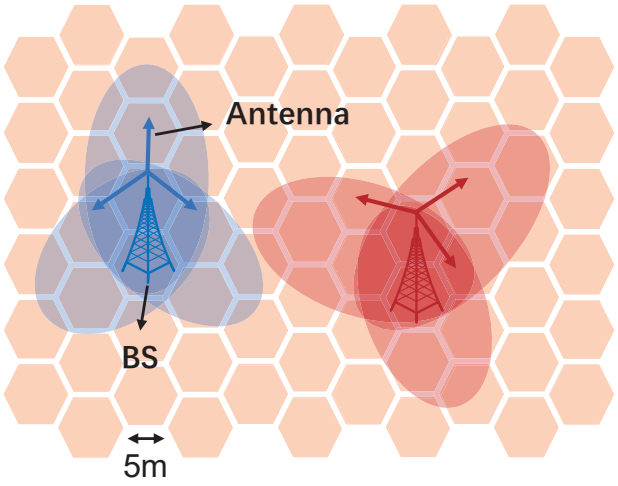


Fig. 1. Illustration of network scenario.

The optimization task can be formulated as

$$\begin{aligned} \max_{\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}} \quad & \frac{1}{K}\sum_{k=1}^{K} x_k \\ \text{s.t.} \quad C_1: \quad & x_k T^P \le p_k(\boldsymbol{\theta}, \boldsymbol{\phi}) && \forall k \in \mathbb{K} \\ C_2: \quad & x_k T^C \le \rho_k(\boldsymbol{\theta}, \boldsymbol{\phi}) && \forall k \in \mathbb{K} \\ C_3: \quad & x_k \in \{0, 1\} && \forall k \in \mathbb{K} \\ C_4: \quad & \theta_n \in \Theta_n && \forall n \in \mathbb{N} \\ C_5: \quad & \phi_n \in \Phi_n && \forall n \in \mathbb{N}, \end{aligned} \tag{4}$$

where $\boldsymbol{x} = \{x_k\}_{k\in\mathbb{K}}$. C_1 and C_2 represent the requirements of the STPs in terms of valid signal power and SIR, respectively.

III. Scalable Antenna Orientation Optimization

A. Cooperative Markov Game Model

Observe that the orientation configuration policies of individual antennas are interdependent due to the potential for overlapping coverage areas, which leads to mutual interference. We adopt a finite-horizon cooperative Markov game (CMG) model to capture the dependence, where multiple agents are involved and they work together to maximize a shared reward function [14]. The CMG components are defined accordingly:

Agent: Each antenna $n \in \mathbb{N}$.

State: The observation history $\boldsymbol{h}$ of the target area, which is obtained by all STPs.

Action: One possible orientation setting of the antenna, i.e., $\psi_n = (\theta_n, \phi_n)$. The action set of agent n is represented by $\Psi_n = \{(\theta_n, \phi_n) | \theta_n \in \Theta_n \wedge \phi_n \in \Phi_n\}$ and the action set of all agents is $\boldsymbol{\Psi} = \prod_{n\in\mathbb{N}} \Psi_n$.

Reward: The proportion of the satisfied STPs, i.e., $R = \frac{1}{K}\sum_{k\in\mathbb{K}} x_k$.

A policy maps the observations of all antennas to a configuration $\boldsymbol{\psi} = \{\psi_n\}_{n\in\mathbb{N}}$. Our aim is to find the optimal policy that maximizes the expected reward over a finite horizon T. Q-learning methods can work out the optimal policy by iteratively estimating a Q-function which quantifies the

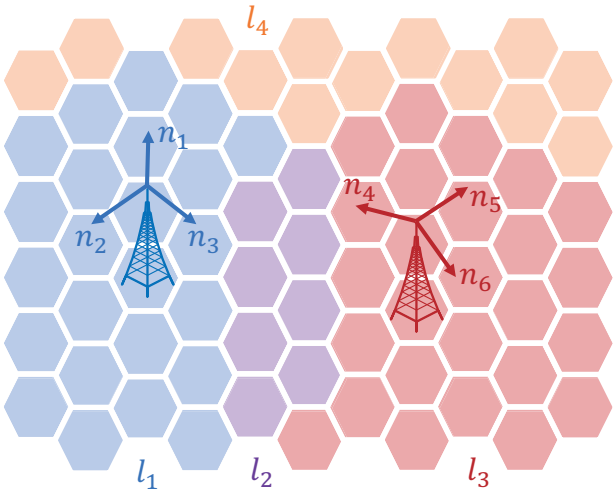


Fig. 2. Decomposition for target area.

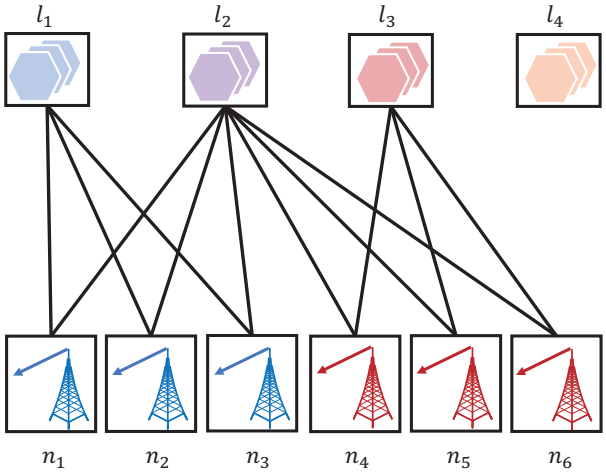


Fig. 3. Factor graph.

expected cumulative reward from obtained observations [15]. Specifically, the Q-function $Q(\boldsymbol{h}, \boldsymbol{\psi})$ is represented by the expected reward of action $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ with observation history $\boldsymbol{h}$ and behaving optimally from then on. The optimal policy $\boldsymbol{\pi}$ can be derived by setting

$$\boldsymbol{\pi}(\boldsymbol{h}) = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{h}, \boldsymbol{\psi}). \tag{5}$$

In principle, $Q(\boldsymbol{h}, \boldsymbol{\psi})$ can be directly estimated by standard Q-learning update rule:

$$\begin{aligned} Q(\boldsymbol{h}^t, \boldsymbol{\psi}^t) =& (1-\alpha) Q(\boldsymbol{h}^t, \boldsymbol{\psi}^t) \\ &+ \alpha [R^t + \gamma \max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi})], \end{aligned} \tag{6}$$

where α is the learning rate, R^t is the immediate reward of configuring $\boldsymbol{\psi}^t$, and γ is the discount factor which is set to 1 for a finite horizon. This centralized method theoretically yields an optimal policy, but is practically infeasible since the computation and storage overheads expand exponentially with the number of agents. Alternatively, the agents can ignore the actions and rewards of the other agents and simultaneously learn their own Q-functions solely based on their local observations and rewards, which is known as independent reinforcement learning. This method is distributed and can lead to significant savings in computation and storage while eliminating the communication overhead during learning and execution, but the agents lack coordination in this method, which may result in suboptimal even undesired policies. Therefore, developing a method that simultaneously attains scalability and optimality is imperative for effective antenna orientation optimization.

B. Coordinated Reinforcement Learning

Note that an mmWave antenna can only cover a limited range since the signal strength diminishes rapidly with increasing transmission distance. Therefore, we can divide the target area into multiple subregions and evaluate each subregion separately. When focusing on a subregion, only the antennas located in or near the subregion needs to be jointly considered, and the antennas situated far from the subregion can be disregarded due to their negligible impact on the subregion. The reward R can be rewritten as a sum of local rewards,

$$R = \sum_{l \in L} R_l, \tag{7}$$

where R_l denotes the local reward of subregion l, which is given by the proportion of the satisfied STPs in l to the total STPs in the target area, and is only associated with the antennas in or near the subregion l. For instance, the network in Fig. 1 can be partitioned into 4 subregion, namely l_1, l_2, l_3, l_4, as shown in Fig. 2. In this case, l_1 only relies on (n_1, n_2, n_3), while l_2 is associated with all the 6 antennas, l_3 only relies on (n_4, n_5, n_6), and l_4 is independent of the antennas. The reward can be expressed as $R = R_{l_1} + R_{l_2} + R_{l_3} + R_{l_4}$.

Inspired by the structured interactions among antennas, we present a coordinated reinforcement learning (CRL) method to cope with the antenna orientation optimization, which decomposes and distributes the policy learning process, and coordinates the distributed learning process to ensure the coverage and capacity performance. Define that antennas associated with subregion l collectively form a group l, and L is the set of groups. The Q-function $Q(\boldsymbol{h}, \boldsymbol{\psi})$ is represented by a sum of local Q-functions based on the groups,

$$Q(\boldsymbol{h}, \boldsymbol{\psi}) = \sum_{l \in L} Q_l(\boldsymbol{h}_l, \boldsymbol{\psi}_l), \tag{8}$$

where $\boldsymbol{h}_l$ is the observation history of group l, $\boldsymbol{\psi}_l$ is the set of antenna orientation configurations for group l, and $Q_l(\boldsymbol{h}_l, \boldsymbol{\psi}_l)$ is the expected reward for the agents in group l by doing action $\boldsymbol{\psi}_l$ at history $\boldsymbol{h}_l$ and behaving optimally from then on with respect to maximizing $Q(\boldsymbol{h}, \boldsymbol{\psi})$. The decomposability of $Q(\boldsymbol{h}, \boldsymbol{\psi})$ is proven in Appendix.

With the decomposition in (8), the Q-learning update rule

can be rewritten as

$$\sum_{l\in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^t) = (1-\alpha)\sum_{l\in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^t) + \alpha[\sum_{l\in L} R_l^t + \gamma \max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi})]. \quad (9)$$

The discounted future reward $\max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi})$ cannot be directly expressed as the sum of local discounted future rewards since it relies on the action that maximizes the reward with respect to the whole area. Define that

$$\boldsymbol{\psi}^* = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi}), \quad (10)$$

$$\max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi}) = Q(\boldsymbol{h}^{t+1}, \boldsymbol{\psi}^*) = \sum_{l\in L} Q_l(\boldsymbol{h}_l^{t+1}, \boldsymbol{\psi}_l^*). \quad (11)$$

Subsequently, all terms in (9) can be decomposed and the update rule for each group l is written as

$$Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^t) = (1-\alpha) Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^t) + \alpha[R_l^t + \gamma Q_l(\boldsymbol{h}_l^{t+1}, \boldsymbol{\psi}_l^*)]. \quad (12)$$

The local contribution $Q_l(\boldsymbol{h}_l^{t+1}, \boldsymbol{\psi}_l^*)$ may be lower than the maximum value of the local Q-function $\max_{\boldsymbol{\psi}_l} Q_l(\boldsymbol{h}_l^{t+1}, \boldsymbol{\psi}_l)$ because $Q_l(\boldsymbol{h}_l^{t+1}, \boldsymbol{\psi}_l)$ is unaware of the dependencies among groups. Distributed constraint optimization (DCOP) techniques can be employed to compute $\boldsymbol{\psi}_l^*$, which will be discussed later.

Using the update rule in (12), the proposed method decentralizes the learning of the Q-function among groups. Assume that each group has a representative agent which learns Q_l on behalf of the group. The representative agent can be chosen arbitrarily from the group. The entire learning process is as follows. During each learning episode t, the agents in group l receive their observations and transmit the observations to the representative agent of their group after executing actions $\boldsymbol{\psi}_l^t$, and the representative agent receives the group reward R_l^t. Then the representative agent computes the next best action $\boldsymbol{\psi}_l^*$ for the updated observation history $\boldsymbol{h}_l^{t+1}$ by DCOP techniques, and updates the local Q-function Q_l using rule (12). Finally, the action $\boldsymbol{\psi}_l^*$ is distributed to the agents in l.

During execution, action selections are computed in a distributed manner since the Q-function is denoted by the local Q-functions of the representative agents. Note that the local Q-function $Q_l(\boldsymbol{h}_l, \boldsymbol{\psi}_l)$ depends on the observation history of l, which scales exponentially with the horizon. A fixed-size observation window is adopted to cope with a large horizon, which helps reduce the overhead of computing $Q_l(\boldsymbol{h}_l, \boldsymbol{\psi}_l)$ by constraining the horizon to a reasonable range.

C. Joint Action Selection by DCOP

The proposed method requires computing the optimal joint action that maximizes the Q-function. This problem can be formulated as a DCOP, which is defined by a set of variables $\boldsymbol{\psi} = \{\psi_n\}_{n\in\mathbb{N}}$ and a set of functions $Q = \{Q_l\}_{l\in L}$, where ψ_n represents the action of agent n, and Q_l represents the Q-function of group l [16]. The history $\boldsymbol{h}$ can be ignored in the following discussion since it is fixed for every computation, i.e., we can denote $Q_l(\boldsymbol{h}, \boldsymbol{\psi}_l)$ by $Q_l(\boldsymbol{\psi}_l)$. The goal is to find the optimal joint action $\boldsymbol{\psi}^*$ such that $\boldsymbol{\psi}^* = \arg\max_{\boldsymbol{\psi}} \sum Q_l(\boldsymbol{\psi}_l)$. This problem can be depicted as a bipartite factor graph by generating a node for each variable and each function, and linking a function node to a variable node if the function relies on the variable. For instance, the problem arising from Fig. 2 can be depicted as a factor graph consisting of 6 variable nodes, 4 function nodes and 12 links, as shown in Fig. 3.

Variable elimination is a standard method for obtaining the optimal solution for DCOPs, but its computation cost increases exponentially with the induced width of the factor graph [17]. In this paper, we investigate the max-sum algorithm for an approximate solution of the DCOP, which only requires limited computation overhead and can trade off the quality and efficiency of computing joint actions [18].

The max-sum algorithm operates directly on the factor graph by passing messages between variable nodes and function nodes. The messages are defined as follows:

Message from variable node n to function node l:

$$q_{n\to l}(\psi_n) = \sum_{g\in\mathbb{F}_n\setminus l} r_{g\to n}(\psi_n) + c_{n,l}, \quad (13)$$

where $\mathbb{F}_n$ is the set of function nodes which are connected to variable node n, and $c_{n,l}$ is a normalizing constant that prevents the messages from increasing infinitely in cyclic graphs, and is chosen such that

$$\sum_{\psi_n\in\Psi_n} q_{n\to l}(\psi_n) = 0. \quad (14)$$

Message from function node l to variable node n:

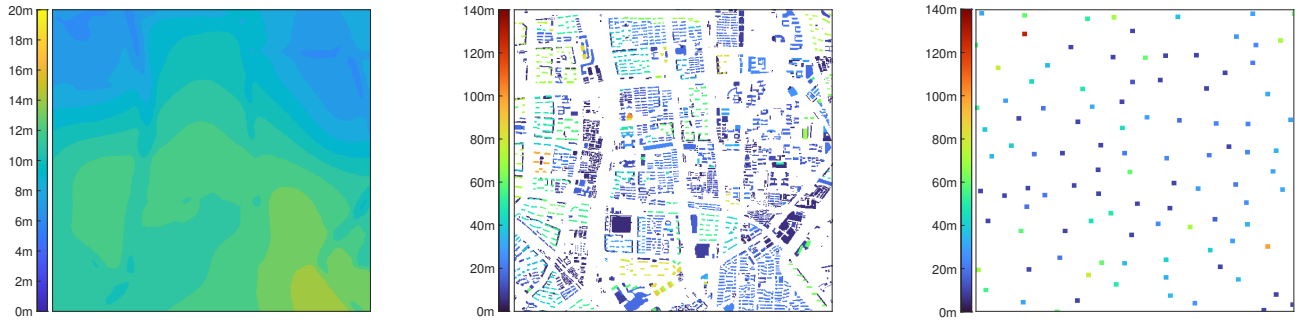
$$r_{l\to n}(\psi_n) = \max_{\boldsymbol{\psi}_l\setminus\psi_n} [Q_l(\boldsymbol{\psi}_l) + \sum_{g\in\mathbb{V}_l\setminus n} q_{g\to l}(\psi_g)], \quad (15)$$

where $\mathbb{V}_l$ is the set of variable nodes which are connected to function node l, and $\boldsymbol{\psi}_l\setminus\psi_n = \{\psi_g | g \in \mathbb{V}_l\setminus n\}$. Here variable node n denotes agent n and function node l denotes the representative agent of group l that hosts the corresponding local Q-function.

During the execution of the max-sum algorithm, the value $\sum_{g\in\mathcal{F}_n} r_{g\to n}(\psi_n)$, which is derived from incoming messages of agent n, approximates the exact reward of action ψ_n with other agents acting optimally. Therefore, the computation cost of the max-sum algorithm can be restricted by controlling the number of rounds of passing messages, which will trade off the quality and efficiency of the action selection. In addition, the max-sum algorithm is inherently distributed and scalable. The messages scale linearly with the maximum number of actions of each agent, and the number of messages varies linearly with the number of agents and groups. The computational complexity scales exponentially with the group size, which is obviously lower than the number of agents.

IV. Numerical Results

Consider a 3km×3km urban area whose terrain information and building layout are shown in Fig. 4(a) and Fig. 4(b), respectively. The mobile communication network in the area



(a) Terrain information. (b) Building layout. (c) BS locations and heights.

Fig. 4. Map information.

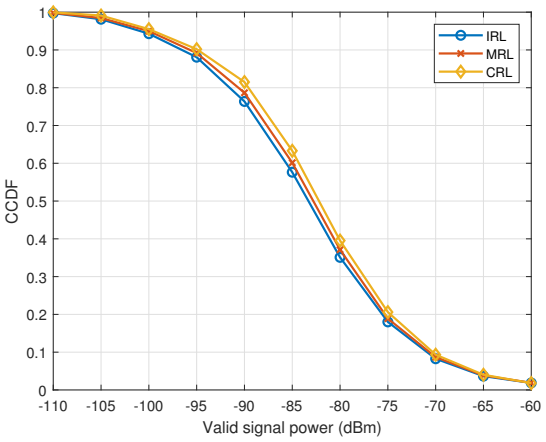


Fig. 5. Complementary cumulative distribution function of valid signal power.

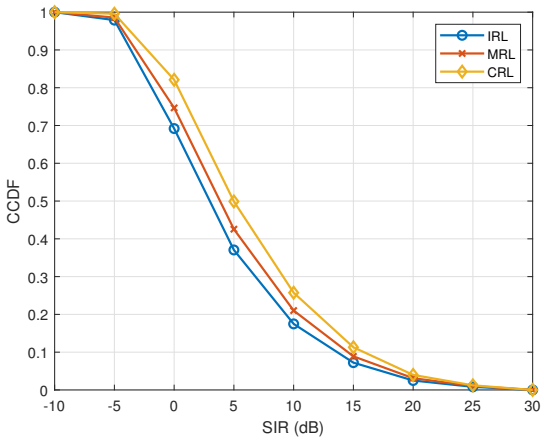


Fig. 6. Complementary cumulative distribution function of signal-to-interference ratio.

consists of 103 BSs, as depicted in Fig. 4(c). The transmission power of each antenna is 15.4dBm. Line-of-sight (LOS) and none-line-of-sight (NLOS) cases are discussed separately as follows:

$$L_{n,k} = \begin{cases} 11.85 + 47.69\lg(d_{n,k}) + 5.83\lg(h_n) \\ \quad + B_k - 6.55\lg(d_{n,k})\lg(h_n) \quad , \text{ if LOS}, \\ 16.24 + 47.69\lg(d_{n,k}) + 5.83\lg(h_n) \\ \quad + B_k - 6.55\lg(d_{n,k})\lg(h_n) \quad , \text{ if NLOS}, \end{cases} \tag{16}$$

where $d_{n,k}$ denotes the distance between antenna n and STP k, and h_n denotes the height of n. The building penetration loss B_k is given by

$$B_k = \begin{cases} 14\text{dB}, & \text{if } k \text{ is indoor}, \\ 0\text{dB}, & \text{if } k \text{ is outdoor}. \end{cases} \tag{17}$$

The power threshold T^P and the SIR threshold T^C are -95dBm and 0dB, respectively. The orientation setting of each antenna is chosen from a predefined codebook so that unnecessary explorations for similar or unpromising settings are reduced, and each codebook is defined by the surrounding propagation environment of the corresponding BS [19, 20]. The proposed coordinated reinforcement learning (CRL) method is compared to the state-of-the-art scalable methods, including the independent reinforcement learning (IRL) [9] and the mean-field reinforcement learning (MRL) [10].

Fig. 5 shows the complementary cumulative distribution function (CCDF) of the valid signal power received by the STPs. Observe that over 85% of the STPs receive signals that exceed the power threshold for all the methods. It suggests that most of the STPs can receive strong enough signals in a densely deployed network even if the antenna orientation settings are not optimized.

Fig. 6 shows the CCDF of SIR at STPs. The proportions of the STPs that exceed the SIR threshold T^C are 69.19%, 74.67% and 82.10%, for the IRL, the MRL and the CRL, respectively. The CRL works out a higher proportion of the STPs that obtain SIR exceeding T^C compared to the IRL and the MRL, since the IRL ignores the interactions among antennas while the MRL simplifies these interactions.

The proportions of satisfied STPs of different methods are

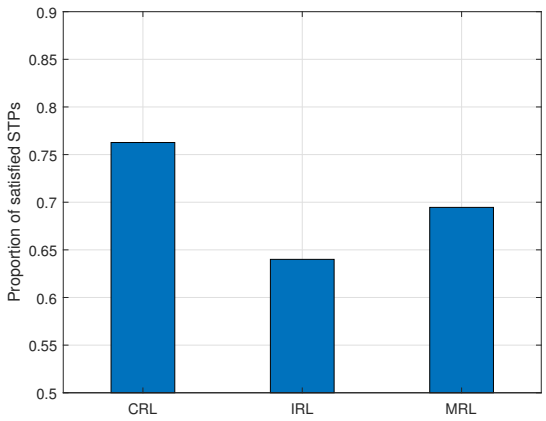


Fig. 7. Network performance in terms of satisfied STPs.

illustrated in Fig. 7. The proportion of the satisfied STPs is 76.27% for the CRL, which outperforms the IRL and the MRL by over 9%. The CRL method achieves a better tradeoff between scalability and optimality compared to the state-of-the-art ones by exploiting the interactions among antennas.

V. Conclusion

In this paper, we proposed a scalable antenna orientation optimization method based on coordinated reinforcement learning to maximize the satisfied service test points in the target area. The proposed method decomposes and distributes the policy learning by exploiting the structured interactions among antennas, and the distributed learning is coordinated through joint antenna configuration derived from existing observations using distributed constraint optimization techniques, thereby ensuring the optimality of the learning. In addition, a max-sum algorithm is used to compute the joint configuration in practical execution so as to trade off computation efficiency and solution quality. Numerical results show that our proposed scheme can produce promising antenna configurations with limited computing resources in large-scale networks, which outperforms the state-of-the-art ones by over 9% in terms of satisfied service test points.

Appendix

Proposition 1. The Q-function $Q(\boldsymbol{h}^t, \boldsymbol{\psi}^t)$ is decomposable for any t in $[0, T]$, i.e.,

$$Q(\boldsymbol{h}^t, \boldsymbol{\psi}^t) = \sum_{l \in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^t).$$

Proof. The proposition holds for $t = T$ since

$$Q(\boldsymbol{h}^T, \boldsymbol{\psi}^T) = R^T = \sum_{l \in L} R_l^T = \sum_{l \in L} Q(\boldsymbol{h}_l^T, \boldsymbol{\psi}_l^T).$$

Assume that the statement holds for t where $1 \le t \le T$. Define $\max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^t, \boldsymbol{\psi}) = \max_{\boldsymbol{\psi}} \sum_{l \in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l) = \sum_{l \in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^*)$, then,

$$\begin{aligned} Q(\boldsymbol{h}^{t-1}, \boldsymbol{\psi}^{t-1}) &= R^{t-1} + \gamma \max_{\boldsymbol{\psi}} Q(\boldsymbol{h}^t, \boldsymbol{\psi}) \\ &= \sum_{l \in L} R_l^{t-1} + \gamma \sum_{l \in L} Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^*) \\ &= \sum_{l \in L} [R_l^{t-1} + \gamma Q_l(\boldsymbol{h}_l^t, \boldsymbol{\psi}_l^*)] \\ &= \sum_{l \in L} Q_l(\boldsymbol{h}_l^{t-1}, \boldsymbol{\psi}_l^{t-1}). \end{aligned}$$

□

References

[1] A. Engels *et al.*, "Autonomous self-optimization of coverage and capacity in LTE cellular networks," *IEEE Tran. Veh. Technol.*, vol. 62, no. 5, pp. 1989–2004, Jun. 2013.

[2] S. Wang, W. Zhao, and C. Wang, "Budgeted cell planning for cellular networks with small cells," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4797–4806, Oct. 2015.

[3] B. Partov, D. Leith, and R. Razavi, "Utility fair optimization of antenna tilt angles in LTE networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 1, pp. 175–185, Feb. 2015.

[4] S. Berger *et al.*, "Joint downlink and uplink tilt-based self-organization of coverage and capacity under sparse system knowledge," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2259–2273, Apr. 2016.

[5] S. Wang and C. Ran, "Rethinking cellular network planning and optimization," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 118–125, Apr. 2016.

[6] W. Hong, K. Baek, and S. Ko, "Millimeter-wave 5G antennas for smartphones: Overview and experimental demonstration," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6250–6261, Dec. 2017.

[7] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov. 2014.

[8] S. Razavizadeh, M. Ahn, and I. Lee, "Three-dimensional beamforming: A new enabling technology for 5G wireless networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 94–101, Oct. 2014.

[9] 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.

[10] E. Balevi and J. Andrews, "Online antenna tuning in heterogeneous cellular networks with deep reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1113–1124, Dec. 2019.

[11] R. Dreifuerst *et al.*, "Optimizing coverage and capacity in cellular networks using machine learning," in *Proc. IEEE ICASSP'21*, Toronto, ON, Canada, Jun. 2021.

[12] S. Hurley, "Planning effective cellular mobile radio networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 243–253, Mar. 2002.

[13] L. Shen and S. Wang, "Monte Carlo Tree Search for network planning for next generation mobile communication networks," in *Proc. IEEE GLOBECOM'21*, Madrid, Spain, Dec. 2021.

[14] K. Son *et al.*, "Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proc. ICML'19*, Long Beach, CA, USA, Jun. 2019.

[15] R. Sutton and A. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[16] A. Petcu and B. Faltings, "A scalable method for multiagent constraint optimization," in *Proc. IJCAI'05*, Edinburgh, Scotland, UK, Jul. 2005.

[17] C. Guestrin, M. Lagoudakis, and R. Parr, "Coordinated reinforcement learning," in *Proc. ICML'02*, Sydney, Australia, Jul. 2002.

[18] A. Farinelli *et al.*, "Decentralised coordination of low-power embedded devices using the max-sum algorithm," in *Proc. AAMAS'08*, Estoril, Portugal, May 2008.

[19] L. Shen and S. Wang, "An efficient codebook based radio parameter optimization method for mobile networks," in *Proc. IEEE GLOBECOM'22*, Rio de Janeiro, Brazil, Dec. 2022.

[20] 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, Aug. 2023.