Online Learning based Downlink Transmission Coordination in Ultra-Dense mmWave Heterogeneous Networks

Ryangsoo Kim[1], Yonggang Kim[1], Nam Yul Yu[1],
Seung-Jun Kim[2], and Hyuk Lim[1]

***Abstract*—In heterogeneous ultra-dense networks with millimeter wave macro cells and small cells, base stations (BSs) and mobile user equipments (UEs) perform beamforming operations to establish highly directional links. In spite of the spatial diversity achieved through directional links, as a number of BSs are densely deployed, inter-cell interference caused by concurrent directional transmissions of adjacent BSs becomes severe, resulting in downlink performance degradation in the network. However, it is very difficult to manage inter-cell interference because of the nature of the time-varying wireless fading environment, the dynamic changes in beam propagation directivity, and unpredictable UEs' locations. In this paper, we propose an online learning-based transmission coordination algorithm based on the framework of multi-armed bandits to learn the unknown characteristics of inter-BS interference and exploit learned data to derive an optimal policy for maximizing the number of successful downlink transmissions. Through numerical simulations, we verify the effectiveness of the proposed online learning-based inter-BS interference management scheme.**

***Index Terms*—Millimeter-wave wireless network, beamforming transmission, downlink coordination, affectance, online learning.**

I. Introduction

With the explosive growth in the wireless traffic demands of a variety of wireless devices in the last decade, future wireless networks are expected to support the massive connectivity requirements of a large number of devices requiring multi-gigabit data rates by utilizing limited spectrum resources. As a solution that increases the reuse per unit area of the spectrum, the ultra-dense network (UDN) has received considerable attention as one of the most promising innovations for future wireless network systems.

The UDN concept refers to the dense deployment of a number of base stations (BSs) with small-cell sizes in order to enhance the spatial-spectral efficiency in an area with high wireless service demands. However, as more BSs with small cell coverage are densely deployed, the inter-BS distance may decrease, which may increase inter-cell interference (ICI) when multiple BSs transmit their downlink packets simultaneously. As a result, the spatial-spectral efficiency is unlikely to increase even as more small cells are deployed. Therefore, it is necessary to enhance the spatial-spectral efficiency of the BSs in UDNs by applying various cutting-edge wireless communication and networking technologies.

Millimeter-wave (mmWave) small-cell deployment is considered a promising solution to this problem. The characteristics of the mmWave spectrum, such as high near-field path loss and low penetration capability, make mmWave small-cell deployment appealing for use in UDNs. In mmWave small-cell networks, the BSs and user equipments (UEs) have antenna arrays that can be integrated into small areas. They perform beamforming operations to establish highly directional links in order to enhance the spatial-spectral efficiency.

In this paper, we consider two-tier heterogeneous mmWave UDNs where small-cell BSs (SBSs) are densely deployed within a macro-cell BS (MBS) coverage area and share a single channel for downlink transmissions. In two-tier heterogeneous mmWave UDNs, the MBS provides downlink connectivity to large coverage area with high transmission power, while the SBSs with low transmission power provide multi-gigabit downlink services by exploiting broadband bandwidth capacity. This two-tier mmWave UDN architecture helps offload the MBS's traffic load to the SBSs geographically distributed over the network, thus resulting in significant improvement in spatial-spectral efficiency. Here, it is assumed that both MBS and SBS exploit antenna-array beamforming technology to establish directional downlink to their associated UEs. Although the beamforming-based downlink transmissions suppress the interference caused by neighboring BS transmissions, inter- and intra-tier interferences may occur in ultra-dense scenarios with multiple SBSs and MBS, when the BSs perform beamforming transmissions simultaneously. Because the interference among directional beams results in a significant level of packet delivery failure, managing inter-BS interference is important.

Recently, the 3GPP proposed the almost blank subframe (ABSF) method to resolve the co-channel ICI problem in LTE heterogeneous network (HetNet) environments where MBS and SBSs interfere with each other [1]. The concept of ABSF-based ICI coordination is to prohibit the channel access of MBS to a portion of the downlink subframes periodically to alleviate the inter-tier interference to the SBS's transmissions. This may enhance network downlink performance by allowing

This work was supported in part by the NRF (2017R1A2B2010478) funded by the Korea government (MSIT) and the GRI funded by the GIST in 2018. *(Corresponding author: H. Lim.)*

[1]R. Kim, Y. Kim, N. Y. Yu, and H. Lim are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea. Email: hlim@gist.ac.kr

[2]S.-J. Kim is with the Computer Science and Electrical Engineering (CSEE) Department, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA.

SBSs to transmit their downlink packets without experiencing significant interference from MBS transmissions. The ABSF method is highly effective when BSs and UEs operate in an omnidirectional mode for data transmissions. On the other hand, in mmWave HetNets where BSs and UEs establish directional links through beamforming, unless the beam directions of the UEs associated with the SBSs are aligned with the MBS's beam, the UEs can receive the downlink packets transmitted from their associated SBSs. Therefore, it is desirable to devise a new ICI coordination approach compatible with beamforming-capable mmWave HetNets where the ICI depends heavily on antenna-array beamforming directions.

In this paper, we introduce a new ICI coordination framework for beamforming-capable ultra-dense mmWave HetNets and propose a transmission coordination scheme for maximizing the number of successful downlink transmissions. In general, the network throughput performance can be evaluated by the aggregate amount of data transmitted from the BSs to UEs in a unit time interval. Therefore, the average downlink performance is approximately given by the average number of successful downlink transmissions multiplied by their average transmission link rate, which is a function of the signal to interference-plus-noise ratio (SINR) between the BSs and UEs. Here, we assume that the average downlink performance is maximized when the transmission link rates between the BSs and UEs are sufficiently high with high SINR and the transmissions are completed in a short time. In this case, the network throughput maximization can be simply formulated to maximize the number of successful downlink transmissions under the condition that the SINR of downlink transmissions should be greater than a certain high SINR threshold. Note that the SINR threshold is a controllable parameter for further optimization of the network throughput performance in practice. In order to characterize the ICI, which varies over inter-BS beam directions, we introduce the concept of inter-beam affectance, which is the amount of normalized interference per beam pair for all BSs in the network. Based on the inter-beam affectance, we propose a transmission coordination scheme that allows the SBSs to perform downlink beamforming transmissions while suppressing packet reception failures at the UEs, which are caused by the interference among directional links. In addition, because inter-beam interference also depends on the time-varying wireless fading environment and unpredictable mobile UEs' locations, it is essential to characterize the inter-beam interferences stochastically. To deal with such a problem in practice, a novel self-learning approach is needed. Thus, we propose an online learning-based transmission coordination algorithm to maximize the average network downlink performance by gradually learning and then exploiting the unknown inter-beam interference.

The main contributions of this paper are as follows:

- We propose the inter-beam affectance, which is a new interference metric to quantify the amount of inter-beam interference in two-tier HetNets, where the BSs and UEs perform beamforming operations to establish directional links.
- We devise an inter-beam affectance-based transmission coordination framework for downlink packet delivery in ultra-dense mmWave HetNets. This metric can be applied for ICI coordination without requiring the computation of SINR in beamforming-capable mmWave HetNets.
- In order to take into account the nature of time-varying wireless fading and the unpredictable mobility of UEs, we devise an online learning-based transmission coordination algorithm based on the framework of multi-armed bandits (MABs), to gradually learn the stochastic characteristics of the inter-beam affectance and to exploit learned information to derive optimal transmission coordination.

The rest of this paper is organized as follows. In Section II, a survey of related work is presented, and then the problem statement and system model are provided in Section III. In Section IV, a transmission coordination problem that maximizes the average number of successful downlink transmissions is formulated and an online learning algorithm that gradually converges to optimal transmission policy is proposed. In Section V, numerical simulation results are presented, followed by conclusions in Section VI.

II. Related Work

We have categorized the existing machine learning (ML)-based interference management methods implemented in wireless network systems. Table I summarizes the different interference management methods along with their ML techniques and management objects.

A. Neural Network-based Approaches

In ML literature, neural networks (NNs) are well-known mathematical tools used to learn the relationships between the input and output data of systems whose models are unknown. The NN consists of numerous nodes (called neurons) positioned in input, hidden, and output layers, and the weighted connections between the nodes, extending from the input layer to the output layer. During the training phase, the weighted connections between the neurons are trained to extract unknown features from data sets, and afterwards, the output of the NN is obtained using the learned weights in multiple layers for input data.

Wijaya *et al.* [2] proposed an ICI management scheme that performs both interference cancellation at the UEs and NN-based transmit-power optimization at the BSs in multiple-input and multiple-output (MIMO) channels. If the channel state information (CSI) is known for all pairs of UEs and BSs, and if the NN is successfully trained, each BS obtains its own transmit-power independently by importing CSI as input to its dedicated NN. To boost the convergence speed in the training phase, they applied a restricted Boltzmann machine (RBM)-based pretraining phase. Adeel *et al.* [3] proposed a random NN-based power controller for uplink ICI coordination in LTE networks. They evaluated the proposed RNN algorithm with respect to four learning algorithms (gradient descent, adaptive-inertia-weight particle swarm optimization, differential evolution, and genetic algorithm) in terms of training speed, prediction accuracy, and computational complexity. In [4], the authors extended their research by integrating the RNN-based algorithm with a genetic algorithm to reduce the

TABLE I
SUMMARY OF ML-BASED ICI MANAGEMENT SCHEMES.

References	Technology	Interference type	Object	ML approach	Learning paradigms	Online learning
[2]	MIMO channel	Intra-tier	Transmit power control	Neural network	Supervised learning	No
[3], [4]	LTE-UL cognitive radio	Intra-tier	Resource block allocation transmit power control	Random neural network	Supervised learning	No
[5]	LTE HetNets	Intra-/inter-tier	Cell range extension transmit power control	Neural network	Supervised learning	No
[6]	OFDMA two-tier network	Intra-/inter-tier	Transmit power control	Q-learning	Reinforcement learning	Yes
[7]	Two-tier HetNets	Intra-/inter-tier	Cell range extension transmit power control	Q-learning	Reinforcement learning	Yes
[8]	Dense small cell networks	Intra-/inter-tier	Transmit power control	Q-learning	Reinforcement learning	Yes
[9]	Dense LTE	Intra-/inter-tier	Resource block allocation	Multi-armed bandit	Reinforcement learning	Yes
[10]	OFDMA	Intra-tier	Resource block allocation	Multi-armed bandit	Reinforcement learning	Yes

ICI further. This hybrid method performs radio resource block (RB) allocation and power control, simultaneously, where an RB corresponds to the smallest resource allocated to users, in the time, space, and frequency domains. Li *et al.* [5] proposed an NN-aided ICI coordination algorithm in LTE HetNets with mobile UEs. The proposed ANN algorithm learns the relationship between the SINR and the UE location, in order to find the optimal ABSF and cell range extension (CRE) patterns from historical data. After ANN training is completed, the ICI coordination problem is decomposed into a single-cell resource allocation problem. Thus, the computational complexity caused by information exchange among intra- and inter-tier BSs is reduced considerably.

However, most ICI management systems using the NN approach need an offline training phase to converge to a steady state. This disadvantage becomes severe as the wireless network environment becomes complicated, as in heterogeneous UDNs. Furthermore, if the environment changes dynamically, the NNs may not converge.

B. Q-learning-based Approaches

Q-learning, which is one of reinforcement learning (RL) methods, finds an optimal action policy that maximizes the long-term reward for a given Markov decision process (MDP) by exploring and exploiting the reward feedback with respect to actions in states. It gradually learns the cost of the state-action combination by continuously updating a Q-table in an iterative fashion. By controlling the learning rate and discount factor, the Q-learning based approach can manage the trade-off between exploration and exploitation. Moreover, the Q-learning algorithm can be applied without requiring a prior model for describing the system environment. Therefore, it is widely applied to ICI coordination in dynamic wireless networks in which the environment changes dynamically.

Galindo-Serrano *et al.* [6] proposed a decentralized Q-learning based interference control algorithm in an orthogonal frequency-division multiple access (OFDMA)-based two-tier HetNet consisting of a macro-cell and femto-cells. The proposed algorithm allows each BS in a femto-cell to find its optimal transmit-power with respect to its allocated RBs in order to maximize the overall femto-cell capacity. The proposed algorithm is a multi-agent system (MAS) because each femto-cell BS acts as an independent agent, without knowledge about the decisions of the other agents. The system state of each agent is designed to ensure that the SINRs at the macro-cell UEs are greater than a given threshold (for guaranteeing the quality of service (QoS) requirements of the macro UEs) and that the total transmit power does not exceed the maximum power. Simsek *et al.* [7] proposed two-stage Q-learning based ICI coordination algorithms for both the time and frequency domains in a two-tier HetNet consisting of macro-cells and pico-cells. In the first stage, each pico-cell BS selects a bias value for CRE and determines the transmit-power by considering the QoS requirement of its UE. In the second stage, the macro-cell BSs consider the pico-cell BSs' actions in the first stage, to select their transmit-power. Lu *et al.* [8] proposed a power control algorithm for coordinating time-domain ICI in dense small-cell networks while guaranteeing the QoS requirements of small-cell UEs. First, the proposed algorithm classifies interfering neighboring cells as aggressor cells if the interference power they caused is greater than a predefined threshold. Then, each aggressor cell determines its own transmit-power using the Q-learning algorithm at the ABSF.

However, if a wireless network system can be described using a stochastic model with unknown parameters, it would be better to characterize the stochastic parameters directly, rather than attempting to learn the relationships among the actions and rewards with respect to the system states. Assuming that the number of parameters with unknown stochastic properties is invariant in the system, as the number of actions and states increases, the computational complexity increases exponentially owing to the increasing number of elements listed in the Q-table.

C. Multi-Armed Bandit-based Approaches

Multi-armed bandit (MAB) is used to derive optimal solutions for combinatorial optimization problems with random

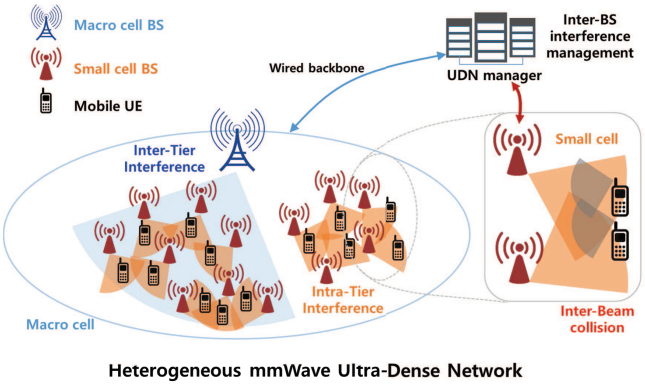


Fig. 1. Inter-cell interference scenarios in two-tier heterogeneous mmWave UDNs.

variables present in the cost function. The MAB gradually learns the characteristics of the random variables with unknown distributions, instead of learning the rewards with respect to the actions. Afterward, it exploits the learned characteristics of the cost function, in order to find an optimal policy that minimizes the system cost in an average sense. Because of its self-learning ability, MABs have been applied to a variety of wireless networking problems recently.

Feki *et al.* [9] proposed an MAB-based autonomous resource allocation algorithm to coordinate the ICI in dense LTE networks. The algorithm is divided into two phases: the cell RB selection phase and the per-user scheduling phase, where the MAB-based online learning approach is applied to the selection of optimal RBs for each cell. However, this algorithm is applicable only when the user's location and traffic load condition are static. Coucheney *et al.* [10] proposed an MAB-based frequency-time resource selection process to coordinate the ICI in an OFDMA wireless network. In general, the EXP3 (Exponential-weight algorithm for Exploration and Exploitation) in [11] is commonly used to find the optimal policy from an MAB problem. However, as the number of UEs associated with BSs increases, the computational complexity increases exponentially. To resolve such a drawback, they devised a Q-EXP3 algorithm, which allows each BS to choose RBs one by one until a predefined number of RBs are chosen, rather than enumerating all possible RB subsets.

In comparison with the NN- and Q-learning-based approaches, the main advantage of the MAB is *the ability to learn the unknown stochastic characteristics of the random variables in the system model directly*. In this paper, we propose an MAB-based online learning algorithm to coordinate the ICI in two-tier heterogeneous mmWave UDNs.

III. Motivation and System Model

A. Motivation

Consider a downlink scenario in a heterogeneous mmWave UDN consisting of MBS and a vast number of SBSs. Because both the mmWave BSs and UEs steer their antenna array phases to transmit and receive signals in specific directions—unlike conventional networks where ICI is incurred by the omni-directional transmission of the BSs—the ICI in mmWave heterogeneous UDNs depends heavily on the direction and width of the antenna array beams. Figure 1 illustrates cases of downlink ICI in a heterogeneous mmWave UDN scenario with directional beamforming. In spite of the enhanced spatial diversity through beamforming, as the density of BSs in the network increases, the intra- and inter-tier ICI becomes severe and results in network performance degradation. This emphasizes the necessity of intra- and inter-tier ICI management capable of improving network performance by alleviating interference.

In this paper, we focus our attention on centralized downlink transmission coordination as a method to manage intra- and inter-tier ICI in heterogeneous mmWave UDNs. It is possible to attempt to determine the BSs' transmissions by estimating the SINR and comparing it with the SINR threshold to ensure that the downlink packets are successfully delivered to the UEs before transmitting the packets. However, if there exist N_s interfering SBSs and N_m interfering MBSs, which are capable of performing B_s and B_m directional beams, respectively, it will be necessary to investigate cases of $B_s^{N_s} \times B_m^{N_m}$ inter-beam collisions. This incurs tremendous time overhead to measure and update all possible collision cases.

To solve this problem, we propose using an inter-beam interference matrix for all *pairs* of beams among the BSs in the network as a metric to estimate the average SINR without investigating all cases in every transmission phase. Because of the nature of time-varying wireless fading and the unpredictable mobility of UEs, it is preferable to deal with the elements of the inter-beam interference matrix as random variables, because their stochastic characteristics are unknown in practice. To capture and exploit the unknown stochastic characteristics of inter-beam interferences, we adopt the MAB framework to derive an online learning-based transmission coordination algorithm. The proposed algorithm learns an inter-beam interference matrix gradually and determines the optimal policy that allows the BSs to transmit their downlink packets only when the aggregate sum of interferences is less than the thresholds. Instead of using the power of the inter-beam interference in the decibel (dB) scale during the learning process, we propose using the inter-beam affectance, which corresponds to the amount of normalized interference. This affectance-based approach makes it easier to take into account the impact of inter-beam interference from individual BSs on successful packet reception at UEs in an average sense.

B. System Model

Consider a downlink scenario in a two-tier heterogeneous UDN composed of one MBS, $(N-1)$ SBSs, and their associated UEs. Let n_M denote the MBS and $n_{S,i}$ denote the i-th SBS for $i = \{1, \cdots, N-1\}$. Further, let $\mathcal{N}_{\text{SBS}} = \{n_{S,1}, \cdots, n_{S,N-1}\}$ denote a set of SBSs, $\mathcal{N} = \mathcal{N}_{\text{SBS}} \cup \{n_M\}$ denote a set of all the BSs, and $\mathcal{U}_n$ denote a set of UEs associated with the BS $n \in \mathcal{N}$. The BSs and UEs are equipped with antenna arrays that perform a beamforming operation to establish directional links between them. We assume that a codebook-based beamforming technology is used to establish directional links. In the codebook-based

beamforming technology, the transmitter and receiver carry out sector-sweep-based beam selection operations to select beams directed toward each other from a predefined set of beams in order to maximize the received signal strength (RSS) at the UE. Assume that the BSs and UEs are capable of performing M directional beamforming. Further, we assume that the BSs are wired through a backbone network and are synchronized with each other. In order to provide the downlink beamforming information to the UDN manager attached to the backbone link, each BS broadcasts its scheduling information through the backbone network.

We consider an SINR model for describing the successful delivery of a downlink packet. Here, a downlink packet is successfully delivered to its destination UE only when the SINR at the UE is greater than a specified threshold. We consider that the signal attenuation in the mmWave band follows a close-in reference distance path loss model as follows:

$$\text{PL}(d)[\text{dB}] = \text{PL}(d_0) + 10\alpha \log_{10}\left(\frac{d}{d_0}\right) + X_n, \qquad (1)$$

where $\text{PL}(d_0) = 20\log_{10}(\frac{4\pi}{\lambda_c})$ is the path loss at the reference distance d_0, λ_c is the wavelength in meters, α is the path loss exponent, and X_n is a Gaussian random variable with zero mean and standard deviation σ in the dB scale.[1] Using the path loss model, the RSS at a UE u for the signal from the BS $n \in \mathcal{N}_{\text{SBS}} \cup \{n_M\}$ can be obtained as $P_r^{n,u} = \kappa P_t^n d_{n,u}^{-\alpha}$, where P_t^n is the transmission power of the BS n, $d_{n,u}$ is the Euclidean distance between the BS n and UE u, and $\kappa = 10^{\text{PL}(d_0)/10}$ is a scaling factor. Let $\Gamma_n(u)$ denote the SINR at $u \in \mathcal{U}_n$; $\Gamma_n(u)$ is given by

$$\Gamma_n(u) = \frac{\kappa P_t^n d_{n,u}^{-\alpha_{n,u}}}{b(\underbrace{\sum_{n' \in \{n_M\}\setminus\{n\}} \kappa P_t^{n'} d_{n',u}^{-\alpha_{n',u}}}_{\text{MBS interference}} + \underbrace{\sum_{n'' \in \mathcal{N}_{\text{SBS}}\setminus\{n\}} \kappa P_t^{n''} d_{n'',u}^{-\alpha_{n'',u}}}_{\text{SBS interference}}) + N_0 W}, \qquad (2)$$

where b is the multi-user interference (MUI) factor, $\alpha_{n,u}$ is the path loss exponent between the BS n and UE u that changes depending on the dynamics of channel environment, N_0 is the noise power spectral density, and W is the channel bandwidth. Let β_u denote the SINR threshold for successful packet reception at u. Then, u can successfully receive downlink packets only when the SINR is greater than β_u (i.e., $\Gamma_n(u) \geq \beta_u$).

IV. MAB-based Inter-cell Interference Management Scheme

A. Affectance-based transmission coordination optimization problem

In this paper, we use the affectance as a measure of normalized interference at UEs, resulting from the signals from their neighboring BSs. Under the SINR equation in (2), the affectance $v_{i,j}(u)$ from the j-th MBS to $u \in \mathcal{U}_i$ is defined as

$$v_{i,j}(u) = \min\left\{ \frac{b\frac{\kappa P_t^j}{\kappa P_t^i}\left(\frac{d_{j,u}^{-\alpha_{j,u}}}{d_{i,u}^{-\alpha_{i,u}}}\right)}{\frac{1}{\beta_u} - \frac{N_0 W}{\kappa P_t^i d_{i,u}^{-\alpha_{i,u}}}}, 1 \right\}, \text{ where } j \in \mathcal{N}\setminus\{i\}. \qquad (3)$$

Here, if the sum of affectance from all BSs in $\mathcal{N}$ is less than or equal to 1 (i.e., $\sum_{j\in\mathcal{N}\setminus\{i\}} v_{i.j}(u) \leq 1$), the SINR at u will be greater than or equal to the threshold (i.e., $\Gamma_i(u) \geq \beta_u$). Note that if b, β_u, κ, P_t^i, N_0, and W are given, the affectance depends on the locations of the UEs and BSs. The affectance in (3) includes all the intra- and inter-tier interferences.

We extend the concept of affectance to the case where the BSs and UEs perform beamforming operations for downlink transmissions. In this case, the level of interference among the downlink transmissions of the BSs depends heavily on the beams selected from the codebook for downlink transmission. Let $f_{i,u}(k)$ denote a binary beamforming boresight function for the i-th BS using the k-th beam and UE u. Here, $f_{i,u}(k) = 1$ if the k-th beam direction of the i-th BS and the beam direction of u point toward each other; otherwise, $f_{i,u}(k) = 0$. We adopt the beamforming gain model in [13], where the main lobe and side lobe gains of directional beamforming are $1-\epsilon$ and ϵ, respectively. If $f_{i,u}(k) = 1$, the beamforming gain is $(1-\epsilon)^2$; otherwise, the beamforming gain is either $(1-\epsilon)\epsilon$ or ϵ^2. Under the assumption that the the side lobe gain is sufficiently small, i.e., $0 \leq \epsilon \ll 1$, the beamforming gain is given by either 1 or 0 depending on the beam boresight function. We define an inter-BS affectance matrix (IBAM) $\mathbf{A}_{i,j} = (a_{i,j}(k,l))_{M\times M}$ to specify the inter-beam interference for all pairs of beamforming transmissions between the i-th and j-th BSs.

Definition 1: $a_{i,j}(k,l)$ is the affectance between BSs when a UE at $u \in \mathcal{U}_i$ is served by the k-th beamforming of the i-th BS, and experiences interference from the l-th beamforming transmission of the j-th BS. It can be obtained by

$$a_{i,j}(k,l) = \mathbb{E}[v_{i,j}(u) | f_{i,u}(k) = 1, f_{j,u}(l) = 1], \qquad (4)$$

where the distribution of $v_{i,j}(u)$ depends on the distributions of $d_{i,u}$, $d_{j,u}$, $\alpha_{i,u}$, and $\alpha_{j,u}$, i.e., the distributions of the UE mobility and wireless channel characteristics. It is difficult to compute $a_{i,j}(k,l)$ from the expectation in (4) because the distribution of $v_{i,j}(u)$ is unknown and difficult to obtain. Therefore, we adopt an online learning framework to learn the IBAMs $a_{i,j}(k,l)$ through consecutive measurement and update phases. The proposed online learning-based approach makes the IBAM-based interference management framework applicable regardless of the wireless channel and UE mobility models.

It is worth noting that the IBAM-based interference management framework is capable of quantifying both intra- and inter-tier interferences. Under the two-tier heterogeneous network scenario, the MBS's transmission brings more critical interference than SBS's transmissions because the transmission power of the MBS is higher than that of the SBSs. The affectance takes into account the different transmission power

[1] According to the real-life measurement campaign for small cell mmWave channel modeling in [12], the average pathloss exponent (PLE) values with respect to the line-of-sight (LOS)- and non-line-of-sight (NLOS)-propagation under the mobile receiver scenario are 2 and 3.3, respectively, where the standard deviation σ also varies from 5.2 to 7.6 depending on the propagation condition.

levels by P_t^i and P_t^j in (3). Consequently, the proposed IBAM-based interference management framework can distinguish the amount of interference brought by heterogeneous BSs with different transmission power level.

Under the proposed IBAM concept, we formulate an optimization problem to find the optimal transmission policy for both of an MBS and SBSs. Here, the optimal transmission policy maximizes the number of successful downlink transmissions, of which the aggregate affectance is less than or equal to 1. Let $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_{N-1}, \delta_N]$ denote the transmission decision vector, where δ_i's for all $i = \{1, \cdots, N-1\}$ denote the transmission decision variable of the i-th SBS and δ_N denotes the transmission decision variable of the MBSs. For all $i = \{1, \cdots, N-1\}$, if $\delta_i = 1$, the i-th SBS is allowed to transmit its downlink packet; otherwise, it is zero. Similarly, if the MBS is allowed to transmit its downlink packet, $\delta_N = 1$; otherwise, it is zero. The optimal transmission decision vector $\boldsymbol{\delta}^*$ is obtained by solving the binary maximization problem over $\boldsymbol{\delta} \in \{0,1\}^{N\times 1}$ as follows:

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta}\in\{0,1\}^{N\times 1}}{\arg\max} \sum_{i\in\mathcal{N}} \delta_i \cdot \mathbb{I}\Bigg[\underbrace{\sum_{j\in\{n_M\}\setminus\{i\}} \delta_j \mathbf{m}_i \mathbf{A}_{i,j} \mathbf{m}_j^T}_{\text{MBS affectance}} + \underbrace{\sum_{j\prime\in\mathcal{N}_{SBS}\setminus\{i\}} \delta_{j\prime} \mathbf{m}_i \mathbf{A}_{i,j\prime} \mathbf{m}_{j\prime}^T}_{\text{SBS affectance}} \leq 1 \Bigg] \tag{5}$$

where $\mathbb{I}[x]$ is an indicator function defined as $\mathbb{I}[x] = 1$ if x holds true; otherwise, it is zero. In addition, $\mathbf{m}_i = \{0,1\}^{1\times M}$ indicates the beamforming index vector to be used for the i-th BS's transmission. If the k-th beam is scheduled to be used, the k-th element of $\mathbf{m}_i$ will be 1; otherwise, it will be zero.

The optimization in (5) attempts to find the transmission decision vector $\boldsymbol{\delta}^*$ that maximizes the function. However, $\mathbf{A}_{i,j}$ is a function of the transmission power at the BSs (P_t^i and P_t^j) and $\mathbf{m}_i$ is a function of beamforming vector. The optimization can be formulated as a higher dimensional form to find the optimal transmission power levels of BSs, their beamforming vectors, and transmission decision vectors. However, this causes an exponential increase in the computational complexity. For this reason, we focus on the scenario where the transmission power at a BS and the beamforming vector are known and have fixed values, and the multi-level transmission power selection and beamforming vector control remain as future work.

Remark 1: It is evident that if the SINR-based interference management framework is adopted, the downlink transmission coordination may achieve better network performance than the IBAM-based framework. However, the SINR-based framework requires tremendous time and computational complexity to estimate the SINR for all the UEs at every downlink transmission session, and the complexity increases exponentially with respect to the number of BSs. This means that the SINR-based interference management framework approach becomes infeasible, especially in the two-tier mmWave UDN environment. On the other hand, once the IBAMs are measured and stored, the IBAM-based framework does not need to estimate SINR for the UEs at each downlink transmission session. In fact, the IBAM is used as a metric to quantify the interference for all pairs of directional downlink transmissions in an average sense, even though there exists a certain level of accuracy loss. The low complexity brought by the expectation-based inter-beam affectance makes the IBAM-based framework appealing in the two-tier heterogeneous mmWave UDN considered in this paper.

B. Greedy algorithm and its approximation factor

The optimal transmission decision vector in (5) can be obtained by a combinatorial algorithm. One method to solve this problem is to enumerate all possible candidates. However, its complexity grows exponentially with respect to the dimension of $\boldsymbol{\delta}$, and thus it is inefficient for application in UDNs where a vast number of BSs are densely deployed. To make this problem more tractable, we propose the use of an approximation algorithm that finds a suboptimal solution in polynomial time with a guaranteed worst-case performance. The optimal solution $\boldsymbol{\delta}^*$ in (5) can be obtained by solving a multidimensional knapsack problem, which is known to be NP-complete, given as follows:

$$\begin{aligned} \text{maximize} \quad & \sum_{i\in\mathcal{N}} \delta_i \\ \text{subject to} \quad & \sum_{j\in\{n_M\}\setminus\{i\}} w_{i,j}\delta_i\delta_j + \sum_{j\prime\in\mathcal{N}_{SBS}\setminus\{i\}} w_{i,j\prime}\delta_i\delta_{j\prime} \leq b, \quad \forall i \in \mathcal{N} \\ & \delta_i \in \{0,1\}, \quad \forall i \in \mathcal{N}, \end{aligned} \tag{6}$$

where $w_{i,j}$ is a constant given by $w_{i,j} = \mathbf{m}_i \mathbf{A}_{i,j} \mathbf{m}_j^T$ for $\forall i, j \in \mathcal{N}$, and b is 1. Here, we note that the multidimensional knapsack problem described above can be transformed into a single knapsack problem according to the following proposition:

Proposition 1: Let $\sum_{j\in\mathcal{N}\setminus\{i\}} w_{i,j}\delta_i\delta_j \leq b_i, \quad \forall i \in \mathcal{N}$ denote the maximum weight constraint of the multidimensional knapsack problem. If the maximum weight capacities b_i of the multidimensional knapsack problem are equal to a constant b for all $i \in \mathcal{N}$, then the problem can be simplified into a single knapsack problem with a one-dimensional maximum weight constraint.

Proof: Under the assumption that $b_i = b$ for all $i \in \mathcal{N}$, the maximum weight capacity constraint in (6) can be rewritten as $\max_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}\setminus\{i\}} w_{i,j}\delta_i\delta_j \leq b$, where $w_{i,j}$ is a constant. Let $\mathcal{A}_{\boldsymbol{\delta}} = \{i | \delta_i = 1, i \in \mathcal{N}\}$ denote the index set for $\delta_i = 1$ and $\mathbb{1}_N \in \mathbb{R}^N$ denote an all-ones column vector. Let us assume that an element $\delta_k \in \boldsymbol{\delta}$ is 0, and that $D_a(\mathcal{A}_{\boldsymbol{\delta}}) = \max_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}\setminus\{i\}} w_{i,j}\delta_i\delta_j$ is a set function of the maximum weight constraint. When δ_k is changed from 0 to 1, the variation of the set function with respect to δ_k is always greater than or equal to 0. In other words, for all $\mathcal{A}_{\boldsymbol{\delta}_1} \subseteq \mathcal{A}_{\mathbb{1}_N}$ and for all $\delta_k \in \mathcal{A}_{\mathbb{1}_N} \setminus \mathcal{A}_{\boldsymbol{\delta}_1}$, if $\mathcal{A}_{\boldsymbol{\delta}_2} = \mathcal{A}_{\boldsymbol{\delta}_1} \cup \mathcal{A}_{\delta_k}$, then $D_a(\mathcal{A}_{\boldsymbol{\delta}_2})$ is always greater than or equal to $D_a(\mathcal{A}_{\boldsymbol{\delta}_1})$. Because the maximum weight constraint function $D(\mathcal{A}_{\boldsymbol{\delta}})$ is a nondecreasing function with respect to $\boldsymbol{\delta}$, it is possible to

transform the multidimensional knapsack problem in (6) into a single knapsack problem as follows:

$$\begin{aligned} \text{maximize} \quad & \sum_{i\in\mathcal{N}} \delta_i \\ \text{subject to} \quad & \max_{i\in\mathcal{N}} \left(\sum_{j\in\{n_M\}\setminus\{i\}} w_{i,j}\delta_i\delta_j + \sum_{j\prime\in\mathcal{N}_{SBS}\setminus\{i\}} w_{i,j\prime}\delta_i\delta_{j\prime} \right) \le b \\ & \delta_i \in \{0,1\}, \quad \forall i \in \mathcal{N}. \end{aligned} \tag{7}$$

■

Based on the above proposition, we devise a greedy algorithm that iteratively finds the suboptimal solution by setting one element of $\boldsymbol{\delta}$ to 1. First, the algorithm enumerates all feasible $\boldsymbol{\delta}$'s with $|\mathcal{A}_{\boldsymbol{\delta}}| = 3$ and starts with the $\boldsymbol{\delta}_g^*$ that gives the minimum weight capacity, from among the feasible $\boldsymbol{\delta}$'s, i.e.,$\boldsymbol{\delta}_g^* = \arg\min_{\boldsymbol{\delta}\in\{\boldsymbol{\delta}|\boldsymbol{\delta}\in\{0,1\}^{N\times 1},|\mathcal{A}_{\boldsymbol{\delta}}|=3\}} D_a(\mathcal{A}_{\boldsymbol{\delta}})$. Then, the algorithm picks one element $\delta_k \in \mathcal{A}_{\mathbb{1}_N} \setminus \mathcal{A}_{\boldsymbol{\delta}_g^*}$, minimizing the increment of the maximum weight capacity as follows:

$$k = \underset{k\in\mathcal{A}_{\mathbb{1}_N}\setminus\mathcal{A}_{\boldsymbol{\delta}_g^*}}{\arg\min} \; D_a(\mathcal{A}_{\boldsymbol{\delta}_g^*} \cup \{k\}). \tag{8}$$

This algorithm terminates when the maximum weight capacity exceeds 1.

As the greedy algorithm may fail to find the global optimal solution $\boldsymbol{\delta}^*$, it is better to check the worst-case performance of the proposed greedy algorithm to guarantee its effectiveness, by deriving its approximation factor. If ϵ denotes the approximation factor, the worst-case performance of the greedy algorithm will be bounded by $f(\boldsymbol{\delta}_g^*) \ge \epsilon f(\boldsymbol{\delta}^*)$, where $f(\cdot)$ is the objective function of a maximum optimization problem. If the approximation factor is a constant value, the greedy algorithm can find the suboptimal solution in polynomial time. Based on the above proposition, the optimization problem in (5) can be transformed into a submodular maximization problem (SMP), which achieves the $1-e^{-1}$ approximation factor. The detailed procedure to derive the approximation factor of the proposed greedy algorithm is described in Appendix A.

C. Policy design

In our downlink transmission coordination problem, $\mathbf{A}_{i,j}$ is unknown in advance, and the online learning process estimates $\mathbf{A}_{i,j}$. Let t denote the iteration index representing a decision period for online learning. The proposed MAB-based transmission coordination algorithm updates the estimated affectance $\hat{\mathbf{A}}_{i,j}$ for each decision period and finds the optimal solution that maximizes the cost function in (4) with $\hat{\mathbf{A}}_{i,j}$. The detailed procedure is described in Algorithm 1. This algorithm is inspired by the linear rewards (LLR) approach in [14], which exploits the learned information from the operation of each action to make decisions about the next action under the assumption that the same random variable can be observed from the operation of different actions.

In Algorithm 1, the initial learning process is performed for each element in $\hat{\mathbf{A}}_{i,j}$, so that every inter-beam affectance is updated at least once. In lines 3–9, an arbitrary binary transmission decision vector is chosen, and the estimated affectance $\hat{\mathbf{A}}_{i,j}$ is observed and updated. Based on $\hat{\mathbf{A}}_{i,j}$, an optimal transmission policy is determined as described in (9), in line 13. Let $\mathbf{m}_i(t)$ denote the beamforming index vector of the i-th BS at the t-th iteration. In (9), $\hat{w}_{i,j}(t) = \mathbf{m}_i(t)\hat{\mathbf{A}}_{i,j}\mathbf{m}_j(t)^T$ and $\eta_{i,j}(t) = \mathbf{m}_i(t)\boldsymbol{\eta}_{i,j}\mathbf{m}_j(t)^T$, where $\boldsymbol{\eta}_{i,j} = (\eta_{i,j}(k,l))_{M\times M}$ is the linear reward matrix that controls the tradeoff between exploration and exploitation. It can be expressed by $\eta_{i,j}(k,l) = \sqrt{\frac{(N(N-1)+1)\ln t}{c_{i,j}(k,l)}}$ where $c_{i,j}(k,l)$ is the observation time up to the current iteration for $a_{i,j}(k,l)$.

Algorithm 1 Proposed online learning approach

1: **while** $\min c_{i,j}(k,l) = 0$ **do**
2: $t := t+1$;
3: Play an arm $\boldsymbol{\delta}$ such that at least one unexplored element $\hat{a}_{i,j}(k,l)$ is observed, i.e., $\delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j = 1$.
4: Measure the instantaneous affectance $v'_{i,j}(u)$
5: $\hat{a}_{i,j}(k,l) = \frac{c_{i,j}(k,l)\times\hat{a}_{i,j}(k,l)+v'_{i,j}(u)\times\delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}{c_{i,j}(k,l)+\delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}$;
6: $c_{i,j}(k,l) = c_{i,j}(k,l) + \delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j$;
7: **end while**
8: // Main loop
9: **while** 1 **do**
10: $t := t+1$;
11: Play any arm $\boldsymbol{\delta} \in \{0,1\}^{N\times 1}$, which solves the following problem:

$$\max_{\boldsymbol{\delta}\in\{0,1\}^{N\times 1}} \sum_{i\in\mathcal{N}} \delta_i \cdot \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}} (\hat{w}_{i,j}(t) - \eta_{i,j}(t))\delta_j \le 1 \right]; \tag{9}$$

12: Measure the instantaneous affectance $v'_{i,j}(u)$.
13: $\hat{a}_{i,j}(k,l) = \frac{c_{i,j}(k,l)\times\hat{a}_{i,j}(k,l)+v'_{i,j}(u)\times\delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}{c_{i,j}(k,l)+\delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}$;
14: $c_{i,j}(k,l) = c_{i,j}(k,l) + \delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j$;
15: **end while**

Note that the proposed learning algorithm follows a *contextual* learning strategy, in which the learning rate varies over the learning situation. When the exploration is more important than exploitation, the algorithm operates on highly explorative behavior and vice versa. In Algorithm 1, the linear reward matrix is used as a contextual learning control component that determines the learning behavior. In (9), $\eta_{i,j}(k,l)$ is inversely proportional to the observation time $c_{i,j}(k,l)$. This indicates that as more observations are performed, the elements of the linear reward matrix decrease. When the linear reward matrix becomes an all 0's matrix, the algorithm finds the solution based on the learned data without any exploration behavior. This implies that after a sufficient observation process, the learning rate of the proposed algorithm may become 0. Therefore, the learning rate of the proposed algorithm is contextually determined by the linear reward matrix.

The proposed algorithm iteratively finds a global optimal binary transmission decision vector that maximizes the number of successful downlink transmissions in an average sense. We assume that the UE observes the instantaneous affectance

$v'_{i,j}(u)$ and reports it to the BS. [2] Afterward, the accumulated mean affectance $\hat{\mathbf{A}}_{i,j}$ is updated as follows:

$$\hat{a}_{i,j}(k,l) = \frac{c_{i,j}(k,l) \times \hat{a}_{i,j}(k,l) + v'_{i,j}(u) \times \delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}{c_{i,j}(k,l) + \delta_i(\mathbf{m}_i(t))_k(\mathbf{m}_j(t))_l\delta_j}. \tag{10}$$

As the estimated affectance $\hat{\mathbf{A}}_{i,j}$ is updated over time, it gradually converges to the *actual* affectance $\mathbf{A}_{i,j}$. The proposed algorithm requires two storage units of size $N(N-1) \times M^2$ to store $\hat{\mathbf{A}}_{i,j}$ and $\mathbf{C}_{i,j} = (c_{i,j}(k,l))_{M\times M}$.

D. Regret analysis

To verify the feasibility of the solution to the proposed online learning algorithm, it is necessary to analyze the *regret*, which is the difference accumulated between the maximum rewards obtained by the optimal decision and those obtained by the proposed MAB over time. The *regret* after T iterations is given by

$$\mathcal{R}(T) = \sum_{t=1}^{T} g(\boldsymbol{\delta}^*(t)) - g(\boldsymbol{\delta}(t)), \tag{11}$$

where $g(\boldsymbol{\delta}^*(t)) = \sum_{i\in\mathcal{N}} \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}} w_{i,j}(t)\delta_j^*(t) \le 1\right]\delta_i^*(t)$ is the aggregate number of successful downlink transmissions by the optimal transmission coordination $\boldsymbol{\delta}^*(t) = [\delta_1^*(t), \cdots, \delta_N^*(t)]$ and t is the iteration index. If the *regret* of the proposed online learning algorithm grows logarithmically over time, the online learning algorithm finds the optimal solution exponentially more often as time passes [18].

The *regret* analysis is performed by deriving an upper bound for the regret, in terms of the number of iterations. Let $C_{NO}(T)$ denote the number of times, where a non-optimal transmission decision vector is selected for the first T iterations. To derive the upper bound of $C_{NO}(T)$, we define $C_{i,j,k,l}(T)$ as a counter for $a_{i,j}(k,l)$. Once the online learning algorithm selects a non-optimal transmission decision vector, the index $\{i,j\} \subseteq \mathcal{A}_{\boldsymbol{\delta}}$ such that $\{i,j\} = \arg\min_{\{i,j\}\in\mathcal{A}_{\boldsymbol{\delta}}} c_{i,j}(k,l)$ is selected, and the corresponding counter is increased by 1. Then, it is evident that when the non-optimal transmission decision vector is selected, only one counter will increment its value. As a result, the following equation must hold:

$$C_{NO}(T) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{k=1}^{M}\sum_{l=1}^{M} C_{i,j,k,l}(T).$$

Then, the upper bound of the *regret* is given by

$$\mathcal{R}(T) = \Delta_{\max} \times \left(\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{k=1}^{M}\sum_{l=1}^{M} C_{i,j,k,l}(T)\right), \tag{12}$$

[2]As described in (3), the affectance can be derived by measuring RSS of the served BS and interfering BSs. Under the assumption that all the BSs periodically broadcast their beacon messages toward all available directions, the UEs are able to measure the RSS of the beacon packets transmitted by its neighboring BSs. After successfully receiving the downlink packet, the UEs compute the affectance through the measured RSS and then report the affectance to their served BSs through the acknowledgement packet.

TABLE II
SIMULATION PARAMETERS FOR MMWAVE HETEROGENEOUS NETWORK.

Parameter	Value
MBS Tx power [15]	40 dBm (10 W)
MBS Tx range	1000 m
SBS Tx power [15]	20 dBm (100 mW)
SBS Tx range	100 m
Bandwidth [16]	1200 MHz
Noise power [16]	-134 dBm/MHz
path loss exponent [12]	2
Number of Tx and Rx beam sectors	4 (90° per beam)

where $\Delta_{\max} = \max_{t=1,\cdots,T} g(\boldsymbol{\delta}^*(t)) - \min_{\boldsymbol{\delta}\in\mathcal{F}} g(\boldsymbol{\delta}(t))$. The upper bound of the *regret* function $\mathcal{R}(T)$ is derived from the upper bound of the counter $C_i(T)$. It is given by

$$\mathbb{E}\left[C_{i,j,k,l}(T)\right] \le \frac{(N(N-1)+1)\ln T}{\zeta_{\min}^2} + 1 + \frac{\pi^2}{3}N, \tag{13}$$

where $\zeta_{\min}^2$ is a constant less than or equal to 1. The detailed derivation of the upper bound for $\mathbb{E}\left[C_{i,j,k,l}(T)\right]$ is described in Appendix B.

The above equations show that the upper bound for the *regret* of the proposed online learning algorithm increases logarithmically. This implies that as time goes, the proposed algorithm finds an optimal solution[3] more frequently where the optimal solution maximizes the *expected* reward [17]. To verify this, we derive the reward discrepancy between the optimal solution and the solution given by the proposed algorithm. Based on (11), $R(T) - R(T-1)$ becomes the reward discrepancy between the optimal solution and the solution given by the proposed algorithm at the T-th decision period, i.e., $R(T) - R(T-1) = g(\boldsymbol{\delta}^*(T)) - g(\boldsymbol{\delta}(T))$. As T approaches infinity, the reward discrepancy between $g(\boldsymbol{\delta}^*(T))$ and $g(\boldsymbol{\delta}(T))$ converges to 0 as follows:

$$\lim_{T\to\infty} R(T) - R(T-1) \le$$
$$\lim_{T\to\infty} \ln\frac{T}{T-1}\left(\Delta_{\max} N(N-1)M^2\frac{(N(N-1)+1)}{\zeta_{\min}^2}\right) = 0. \tag{14}$$

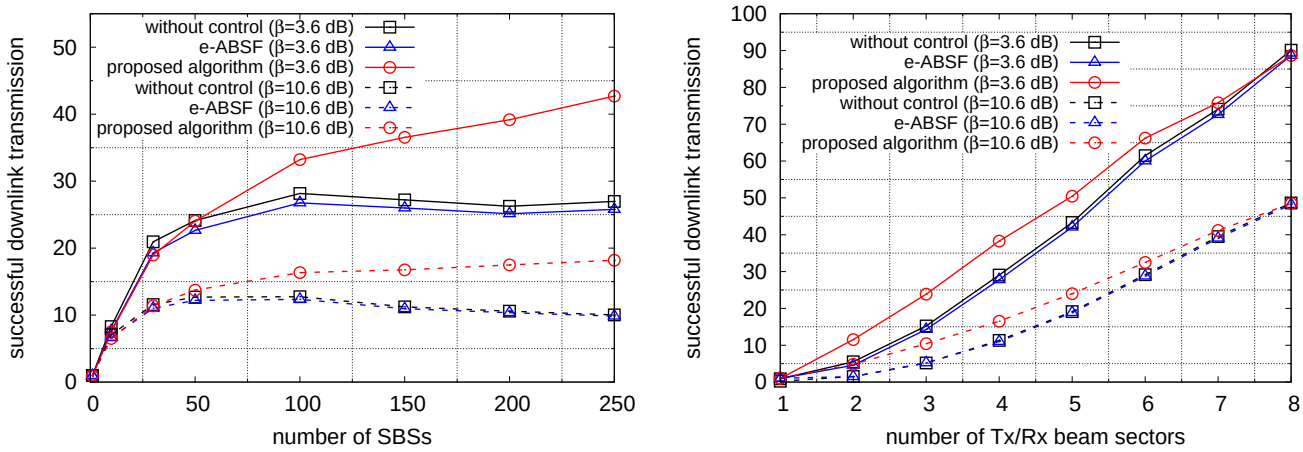
This implies that the proposed online learning algorithm finds the optimal solution by exploiting the learned information rather than exploring more information as time passes [18]. This is an interesting behavior of the *contextual* learning strategy where the learning rate varies over the learning situation.

V. PERFORMANCE EVALUATION

A. Simulation environment

We performed numerical simulations to evaluate the efficiency of the proposed transmission coordination algorithm. In the simulation, the SBSs were randomly deployed in the network over an area of $\mathcal{A} = 500 \times 500\text{m}^2$ and their UEs were uniformly distributed over the transmission range. In addition, a single MBS was deployed in the center of the network area. It was assumed that the BSs always have data

[3]Here, the optimal solution represents the solution given by the optimization problem in (5) where all the stochastic characteristics are assumed to be known in advance.



(a) (b)

Fig. 2. Aggregate number of successful downlink transmissions with respect to a) the number of SBSs and b) the number of Tx/Rx beam sectors.

packets to transmit to their UEs. For comparison, a naïve method without coordination control and enhanced-ABSF (e-ABSF) method were considered. The naïve method without coordination control allows all BSs to always perform their downlink transmissions. In addition, the e-ABSF method is an enhanced version of the ASBF method [1] by taking into account the directional inter-tier interference, and it allows the SBSs to perform their downlink transmission only when the downlink transmission of the MBS does not incur interference owing to the directional mmWave beam propagation characteristics. The detailed simulation parameters are listed in Table II while numerical simulations with respect to the Tx power and number of Tx/Rx beam sectors variations were performed.

B. Simulation results for deterministic case

In this subsection, we evaluate the performance of the proposed transmission coordination method under the deterministic case where the IBAMs are perfectly known in advance.

Figure 2(a) shows the aggregate number of successful downlink transmissions with respect to the number of SBSs in the network. The figure shows that as the density of the SBSs increases, the performance of the naïve and e-ABSF methods gradually decreases owing to the increase in inter-beam interference caused by both intra- and inter-tier concurrent downlink transmissions. On the other hand, as the density of the SBSs increases, the performance of the proposed algorithm gradually increases and levels off. As a result, the performance difference between the proposed method and other methods becomes significant. This is because the proposed algorithm successfully learns the IBAMs and exploits them to avoid packet collisions when the BSs transmit their downlink packets. Thus, the proposed algorithm achieves an optimal transmission policy capable of fully exploiting the high spatial diversity of beamforming in the two-tier heterogeneous mmWave UDN scenario.

Figure 2(b) shows the aggregate number of successful downlink transmissions with respect to the number of Tx and Rx beam sectors when the number of SBSs in the network is 150. The figure shows that, as the number of the beam sectors increases, the performances of all methods increase. This is because the capability to exploit spatial diversity through directional beamforming is enhanced, resulting in alleviation of the inter-beam interference incurred by concurrent intra- and inter-tier transmissions. The figure also shows that the performance difference between the proposed method and other methods increases as the number of beam sectors decreases (i.e., the inter-beam interference becomes severe). This implies that the proposed method coordinates concurrent transmissions of the BSs in the network to minimize the downlink packet reception failure caused by intra- and inter-tier interferences in order to improve the downlink performance. In the e-ABSF method, all the SBSs that interfered with the MBS's downlink transmission are prohibited from performing their downlink transmission regrardless of the level of interference. Although the e-ABSF allows the SBSs to avoid inter-tier interference, it is unable to prevent the intra-tier interferences incurred by concurrent transmissions of the SBSs.

It is evident that if the number of beam sectors becomes sufficiently large, the effectiveness of the proposed method will decline because of negligible inter-beam interference. However, the increase in the number of beam sectors may result in increased costs for the antenna array infrastructure and beam alignment overhead. This implies that there is a trade-off between the infrastructure expenditure for antenna arrays and the inter-beam interference in beamforming capable downlink scenarios in ultra-dense HetNets. Note that when the number of beam sectors is 1 (i.e., omni-directional Tx/Rx scenario), the probability of successful transmissions for all methods is almost 0. In the naïve method, the SBSs' transmissions are significantly affected by the concurrent transmissions of the MBS because the MBS transmits its downlink packets with 100 times higher transmission power than that of the SBSs. On the other hand, the e-ABSF and proposed methods prohibit all SBSs from transmitting their packets because all SBSs in the networks cannot to avoid the inter-tier interference incurred because of the MBS's transmission under the omni-directional

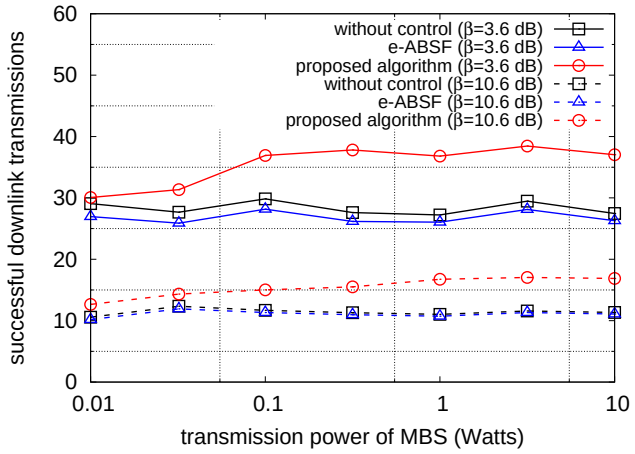


Fig. 3. Aggregate number of successful downlink transmissions with respect to the transmission power ratio between the MBS and SBS when the number of SBSs is 150.

Tx/Rx scenario.

Figure 3 shows the aggregate number of successful downlink transmissions with respect to the transmission power ratio between the MBS and SBSs when the number of SBSs is 150. The figure shows that as $P_t^{n_M}$ increases, the performance discrepancy between both methods increases and levels off. In general heterogeneous networks, an MBS transmits downlink packets with higher transmission power than that of SBSs; hence, the proposed method is an appropriate solution to coordinate concurrent downlink transmissions in heterogeneous mmWave UDNs by taking into account the impact of intra- and inter-tier interferences.

Figure 4 shows the aggregate number of successful downlink transmissions with respect to the downlink traffic load at the MBS and SBSs when the number of SBSs is 150. Fig. 4(a) shows that as the amount of MBS downlink traffic load increases, the performance of all methods decreases gradually. Nevertheless, the proposed method achieves better performance than the other methods. In addition, as depicted in Fig. 4(b), the proposed method achieves better performance than the other methods as SBS traffic load increases. The simulation results verify that the proposed method enhances the network performance of the two-tier heterogeneous mmWave UDN under the various traffic load scenarios.

C. Simulation results for stochastic case

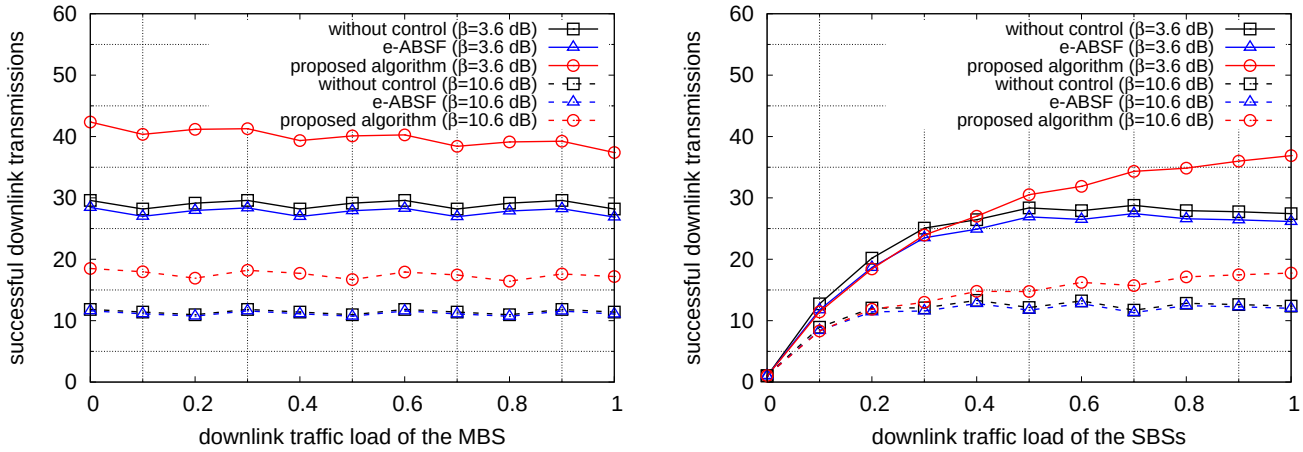
In this subsection, we evaluate the performance of the proposed method under the stochastic case, where inter-beam interferences are gradually learned and exploited to coordinate transmissions. For the simulation, we set the number of SBSs to 5, in which the SBSs are densely deployed in an area of 200×200 m^2. We randomly generated the location of the UEs from homogeneous Poisson point process in the transmission range of their associated BSs. The proposed online learning algorithm iteratively measures and updates the elements in the IBAMs $\mathbf{A}_{i,j}$ for all $i, j \in \mathcal{N}$. For comparison purposes, a naive online learning approach using UCB1 policy in [18] was considered. The UCB1 policy is designed to learn the aggregate number of successful downlink transmissions for all possible beam transmission sets.

Figure 5(a) shows the simulation results of the *regret* divided by the number of iterations. This is the average difference between the maximum rewards obtained by the optimal decision and those obtained by the online learning approaches. The figure shows that for both online learning approaches, the results of the *regret* divided by the number of iterations decrease as the iteration increases. This implies that the both approaches learn the unknown stochastic characteristics gradually and exploit them to find an optimal transmission policy. In comparison with the results of the UCB1 policy, the simulation results of the proposed algorithm converges rapidly to 0. This shows that the proposed algorithm achieves better transmission coordination performance than UCB1 in terms of convergence speed.

In addition, as the number of beam sector (M) increases, the convergence speed for both online learning approaches decreases. In particular, the performance degradation of the UCB1 policy becomes worse than the proposed online learning algorithm in terms of learning convergence. The proposed online learning algorithm learns the inter-beam affectances directly and exploits the learned inter-beam affectances to find the optimal transmission policy. Therefore, as the number of beam sectors increases, the number of unknown objects to be learned by the proposed online learning approach increases linearly. On the other hand, the UCB1 policy is designed to learn the aggregate number of successful transmissions for all beam combinations of the BSs in the network; as a result, as the number of beam sectors increases, the number of unknown objects increases exponentially. Moreover, as the number of SBSs (N) in the network increases, the performance of the UCB1 policy becomes worse. This implies that the UCB1 policy is infeasible for applications to mmWave UDN scenarios where a number of directional beamforming-capable SBSs exist.

Figure 5(b) shows the *regret* divided by the number of iterations in a dynamically changing scenario, where the IBAMs are completely changed owing to the re-deployment of SBSs in the network at the $10,000$th iteration. The simulation results show that the average *regret* of the online learning algorithm rapidly decreases after the $10,000$th iteration, and then levels off after the $12,000$th iteration. For example, if a single downlink session time is 1 ms, the proposed online learning algorithm requires only 2 s to converge. This shows that even though the environment dynamically changes, the proposed online learning algorithm is able to converge consistently.

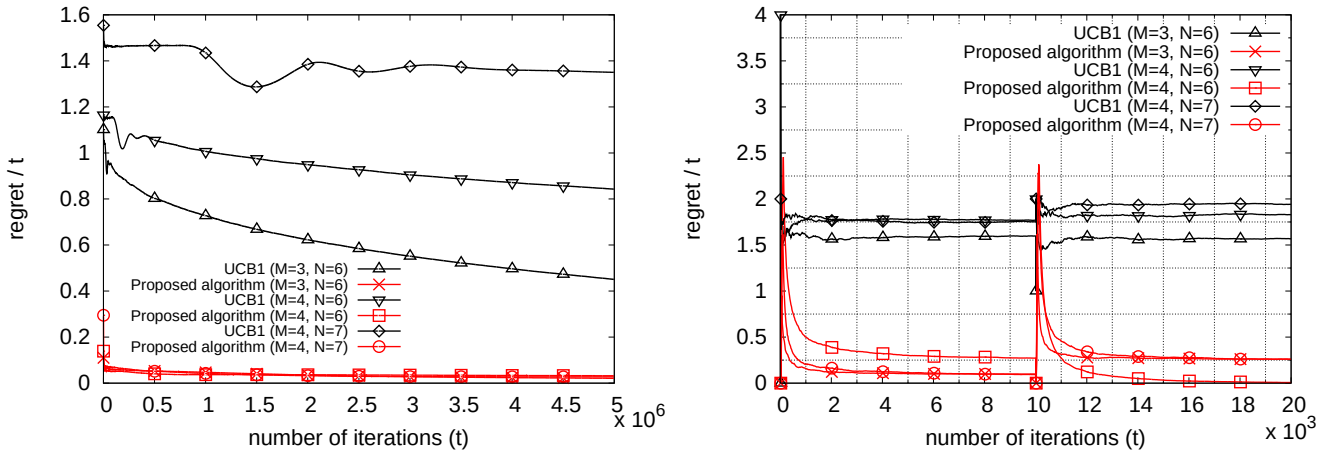
Figure 6 shows the aggregate number of successful downlink transmissions with respect to the iterations. The simulation results show that the proposed online learning algorithm gradually converges as time passes regardless of the amount of downlink traffic load. However, the learning speed of the proposed algorithm to converge to an optimal solution depends on the network traffic load. As shown in the simulation results, as the amount of downlink traffic loads at the BSs increases, the proposed online learning algorithm converges to the optimal solution more rapidly. The reason is that as more downlink transmissions are performed, the proposed algorithm



(a) MBS downlink traffic load change

(b) SBS downlink traffic load change

Fig. 4. Aggregate number of successful downlink transmissions with respect to the amount of downlink traffic generated when the number of SBSs is 150.



(a)

(b)

Fig. 5. Regret divided by the number of iterations when the SINR threshold is 10.6 dB: a) static case and b) dynamic case when the locations of all BSs are abruptly changed at the 10,000th iteration.

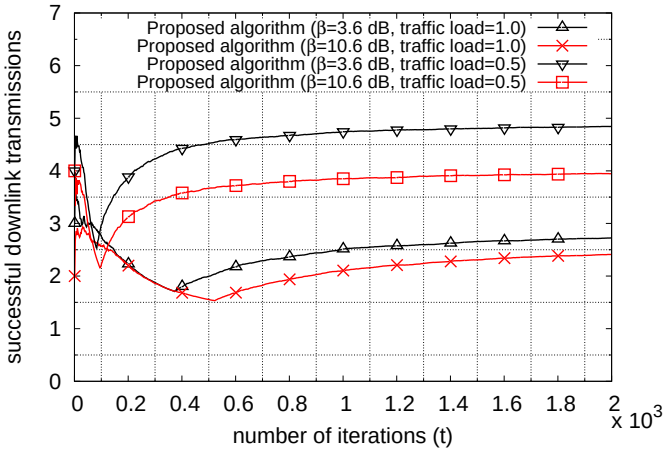


Fig. 6. Aggregate number of successful downlink transmissions with respect to the iteration when the number of SBSs is 5.

gathers more measurement data used for learning IBAMs. Therefore, the learning speed of the proposed online learning algorithm is directly proportional to the traffic in the network. The simulation results in various mmWave HetNet scenarios indicate that the proposed method is applicable to downlink transmission coordination in beamforming-capable mmWave UDNs.

VI. Conclusion and Future Work

In this paper, we considered online learning-aided ICI management in heterogeneous mmWave UDNs. Given the downlink scheduling of BSs in the network, the proposed online learning-based transmission coordination method could find an optimal transmission policy that maximized the number of successful downlink transmissions. The proposed algorithm adopted the MAB framework to learn and exploit the unknown characteristics of ICI to decide which BSs were to be allowed to transmit their downlink packets. The results of the numerical simulations verified that the proposed online learning algorithm significantly improved the downlink performance.

In future work, we will transform the proposed transmission coordination algorithm into a distributed online learning-based transmission decision algorithm that will be applicable for self-organization networks.

APPENDIX A
APPROXIMATION FACTOR DERIVATION FOR THE PROPOSED GREEDY ALGORITHM

In this appendix, we derive the worst-case performance of the greedy algorithm described in Section IV-B by mathematically analyzing the approximation factor. Let $f_a(\mathcal{A}_{\boldsymbol{\delta}}) = |\mathcal{A}_{\boldsymbol{\delta}}|$ and $D_a(\mathcal{A}_{\boldsymbol{\delta}}) = \max_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}\setminus\{i\}} w_{i,j}\delta_i\delta_j$. Then, the problem in (7) can be re-written by

$$\max_{\boldsymbol{\delta}\in\{0,1\}^{N\times 1}} |\mathcal{A}_{\boldsymbol{\delta}}| \quad \text{subject to} \quad D_a(\mathcal{A}_{\boldsymbol{\delta}}) \le 1. \tag{15}$$

Because the objective function $f_a(\cdot)$ is an increasing and modular set function; for an arbitrary sets S and T, the following condition must hold:

$$f_a(S) = f_a(T) + \sum_{i\in T\setminus S} (f_a(S\cup\{i\}) - f_a(S)). \tag{16}$$

Assume that $\boldsymbol{\delta}^*$ is an globally optimal solution of the problem in (15) and $\boldsymbol{\delta}_g^t$ is a solution obtained by the proposed greedy algorithm after the t-th steps. We sort the index set of $\{k_i\} \in \mathcal{A}_{\boldsymbol{\delta}^*}$ such that

$$D_a(k_1,\cdots,k_i) = \min_{k_i\in\mathcal{A}_{\boldsymbol{\delta}^*}\setminus\{k_1,\cdots,k_{i-1}\}} D_a(\{k_1,\cdots,k_{i-1}\}\cup\{k_i\}), \quad \forall i\in\{1,\cdots i^*\},$$

where i^*+1 is the first step of the greedy algorithm for which the algorithm dose not update the solution, i.e., $D_a(\mathcal{A}_{\delta_g^{i^*+1}}) > 1 > D_a(\mathcal{A}_{\delta_g^{i^*}})$. Let $V = \{k_1, k_2, k_3\}$ denote the set of first three elements of the index set $\mathcal{A}_{\boldsymbol{\delta}^*}$. Then, for any element $k_i \in \mathcal{A}_{\boldsymbol{\delta}^*}$, $i \ge 4$, and any set $W \in \mathcal{A}_{\mathbb{1}_N} \setminus \{V\cup\{k_i\}\}$, the following equality hold:

$$f_a(V\cup W\cup\{k_i\}) - f_a(V\cup W) = \frac{1}{3}f_a(V) = 1. \tag{17}$$

We define a new set function $g_a(S) = f_a(S) - f_a(V)$, which is also an increasing and modular set function. Then, the following inequalities also hold:

$$\begin{aligned} g_a(\mathcal{A}_{\boldsymbol{\delta}^*}) &\le g_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}) + \sum_{k\in\mathcal{A}_{\boldsymbol{\delta}^*}\setminus\mathcal{A}_{\boldsymbol{\delta}_g^i}} g_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}\cup\{k\}) - g_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}) \\ &= g_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}) + \sum_{k\in\mathcal{A}_{\boldsymbol{\delta}^*}\setminus\mathcal{A}_{\boldsymbol{\delta}_g^i}} f_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}\cup\{k\}) - f_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}) \\ &\le g_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}) + (1 - D_a(V))\times\theta_{i+1}, \quad 0\le i\le i^*-1, \end{aligned} \tag{18}$$

where θ_i represents the inverse of the minimum increase of the maximum weight constraint set function $D_a(\cdot)$ at the i-th steps as follows:

$$\theta_i = \max_{k\in\mathcal{A}_{\mathbb{1}_N}\setminus\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}}} \frac{f_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}}\cup\{k\}) - f_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}})}{D_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}}\cup\{k\}) - D_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}})}. \tag{19}$$

Let $c_i = \min_{k\in\mathcal{A}_{\mathbb{1}_N}\setminus\mathcal{A}_{\boldsymbol{\delta}_g^{i-1}}} D_a(\mathcal{A}_{\boldsymbol{\delta}_g^i}\cup\{k\}) - D_a(\mathcal{A}_{\boldsymbol{\delta}_g^i})$, then $\sum_{j=1}^i c_j = D_a(\mathcal{A}_{\boldsymbol{\delta}_g^i})$ for all $i = \{1,\cdots,i^*\}$. Let $D_i = \lceil\gamma\sum_{j=1}^i c_j\rceil$ and $D_0 = 0$, where γ is an arbitrary scale-up factor to ensure that the difference between D_i and D_{i+1} become a positive integer for all $i = \{1,\cdots,i^*\}$. Then, we define positive integers $\hat{D}$ and $\bar{D}$ such that the following inequality hold: $\hat{D} = D_{i^*+1} \ge \lceil\gamma\times(1-D_a(V))\rceil = D_3 = \bar{D}$. For $j = \{1,\cdots,\hat{D}\}$, we also define $\rho_j = \theta_i/\gamma$ if $j = \{1,\cdots,i^*\}$.

To derive the approximate factor of the greedy algorithm, we use the Wolsey inequality in [19]: If P and Q are arbitrary positive integers, ρ_is are arbitrary nonnegative reals for $i = \{1,\cdots,P\}$, and $\rho_1 \ge 0$, then

$$\frac{\sum_{i=1}^P \rho_i}{\min_{t=1,\cdots,P}(\sum_{i=1}^{t-1}\rho_i + Q\rho_t)} \ge 1 - \left(1-\frac{1}{Q}\right)^P > 1 - e^{P/Q}. \tag{20}$$

Based on (18) and (20), the following inequalities hold:

$$\begin{aligned} \frac{g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*-1}}\cup\{k_{i^*}\})}{g_a(\mathcal{A}_{\boldsymbol{\delta}^*})} &\ge \frac{\sum_{j=1}^{\hat{D}}\rho_j}{\min_{l=\{1,\cdots,\hat{D}\}}\left\{\sum_{j=1}^{l-1}\rho_j + \bar{D}\rho_l\right\}} \\ &\ge 1 - e^{-\hat{D}/\bar{D}} \\ &> 1 - e^{-1}. \end{aligned} \tag{21}$$

By combining (17) and (21),

$$\begin{aligned} f_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}) &= f_a(V) + g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}) \\ &= f_a(V) + g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}\cup\{k_{i^*+1}\}) \\ &\quad - (g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}\cup\{k_{i^*+1}\}) - g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}})) \\ &= f_a(V) + g_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}\cup\{k_{i^*+1}\}) \\ &\quad - (f_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}}\cup\{k_{i^*+1}\}) - f_a(\mathcal{A}_{\boldsymbol{\delta}_g^{i^*}})) \\ &\ge f_a(V) + (1-e^{-1})g_a(\mathcal{A}_{\boldsymbol{\delta}^*}) - \frac{1}{3}f_a(V) \\ &\ge (1-e^{-1})f_a(\mathcal{A}_{\boldsymbol{\delta}^*}). \end{aligned}$$

Consequently, we derive that the proposed greedy algorithm achieves a constant approximation factor $1-e^{-1}$.

APPENDIX B
UPPER BOUND OF THE COUNTER $C_{i,j,k,l}(T)$

We derive the upper bound of the counter $C_i(T)$. Let $\eta_{i,j,c_{i,j}(t)}$ denote $\sqrt{\frac{(N(N-1)+1)\ln n_{i,j}}{c_{i,j}(t)}}$ and I_t denote the indicator of the counter selected at the t-th iteration. Then, the upper bound of the counter $C_i(T)$ can be derived as follows:

$$\begin{aligned} &\mathbb{E}[C_i(T)] = \sum_{t=1}^T P\{I_t = i\} \le \kappa + \sum_{t=1}^T P\left\{I_t = i, C_i(t-1)\ge\kappa\right\} \\ &\le \kappa + \\ &\sum_{t=1}^T P\Bigg\{\sum_{i\in\mathcal{N}}\delta_i^*(t)\cdot\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t) - \eta_{i,j,c_{i,j}(t)}\right)\delta_j^*(t)\le 1\right] \\ &\qquad \le \sum_{i\in\mathcal{N}}\delta_i(t)\cdot\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t) - \eta_{i,j,c_{i,j}(t)}\right)\delta_j(t)\le 1\right], \\ &\qquad C_i(t-1)\ge\kappa\Bigg\} \\ &\le \kappa + \sum_{t=1}^T P\Bigg\{\min_{1<c_{1,2},\cdots,c_{N,N-1}<t}\sum_{i\in\mathcal{N}}\delta_i^*(t) \end{aligned}$$

$$\cdot \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}}\right)\delta_j^*(t)\le 1\right]$$
$$\le \max_{1<c_{1,2},\cdots,c_{N,N-1}<t}\sum_{i\in\mathcal{N}}\delta_i(t)$$
$$\cdot \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}}\right)\delta_j(t)\le 1\right],$$
$$C_i(t-1)\ge\kappa\Bigg\}$$
$$\le \kappa+\sum_{t=1}^{T}\sum_{c_{1,2}=1}^{t-1}\cdots\sum_{c_{N,N-1}=1}^{t-1}\sum_{\bar{c}_{1,2}=1}^{t-1}\cdots\sum_{\bar{c}_{N,N-1}=1}^{t-1}$$
$$P\Bigg\{\min_{1<c_{1,2},\cdots,c_{N,N-1}<t}\sum_{i\in\mathcal{N}}\delta_i^*(t)$$
$$\cdot \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}}\right)\delta_j^*(t)\le 1\right]$$
$$\le \max_{1<\bar{c}_{1,2},\cdots,\bar{c}_{N,N-1}<t}\sum_{i\in\mathcal{N}}\delta_i(t)$$
$$\cdot \mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,\bar{c}_{i,j}}\right)\delta_j(t)\le 1\right],$$
$$C_i(t-1)\ge\kappa\Bigg\}, \tag{22}$$

where κ is an arbitrary positive integer. To hold the inequality in (22), at least one of the three inequalities (23)–(25) must hold.

$$\sum_{i\in\mathcal{N}}\delta_i^*(t)\cdot\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\delta_j^*(t)\le 1\right]\le f(\boldsymbol{\delta}^*(t)), \tag{23}$$

$$\sum_{i\in\mathcal{N}}\delta_i(t)\cdot\left(2\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\bar{w}_{i,j}(t)\delta_j(t)\le 1\right]-\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\le 1\right]\right)\ge f(\boldsymbol{\delta}(t)), \tag{24}$$

$$f(\boldsymbol{\delta}(t))+2\sum_{i\in\mathcal{N}}[\delta_i(t)\cdot\left(\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\delta_j(t)\le 1\right]-\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\bar{w}_{i,j}(t)\delta_j(t)\le 1\right]\right)]>f(\boldsymbol{\delta}^*(t)). \tag{25}$$

The upper bound for (23) is given by

$$P\{(23)\}$$
$$\le\sum_{i\in\mathcal{N}}\delta_i^*(t)P\left\{\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\delta_j^*(t)\le 1\right]\le\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}w_{i,j}(t)\delta_j^*(t)\le 1\right]\right\}$$
$$\le\sum_{i\in\mathcal{N}}\delta_i^*(t)P\left\{\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\delta_j^*(t)\ge\sum_{j\in\mathcal{N}\setminus\{i\}}w_{i,j}(t)\delta_j^*(t)\right]\right\}$$
$$\le\sum_{i\in\mathcal{N}}\delta_i^*(t)\sum_{j\in\mathcal{N}\setminus\{i\}}\delta_j^*(t)P\left\{\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\ge w_{i,j}(t)\right\}$$
$$\le N(N-1)\times P\left\{\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\ge w_{i,j}(t)\right\}. \tag{26}$$

Here, we apply the Chernoff-Hoeffding bound in (26) to derive its upper bound. Let $X_1,\cdots,X_n$ be random variables in the range $[0,1]$ such that $\mathbb{E}[X_t|X_1,\cdots,X_{t-1}]=\mu$, and $S_n=X_1+\cdots+X_n$. Then, for all $a\ge 0$

$$P\{S_n\ge n\mu+a\}\le e^{-2a^2/n}\text{ and }P\{S_n\le n\mu-a\}\le e^{-2a^2/n}. \tag{27}$$

Then, the upper bound of the probability $P\left\{\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\ge w_{i,j}(t)\right\}$ is given by

$$P\left\{\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\ge w_{i,j}(t)\right\}\le e^{-2(N(N-1)+1)\ln n_{i,j}}=(n_{i,j})^{-2(N(N-1)+1)}\le t^{-2(N(N-1)+1)}. \tag{28}$$

Similarly, the upper bound of the probability for (28) is also given by

$$P\left\{\bar{w}_{i,j}(t)+\eta_{i,j,c_{i,j}(t)}\le w_{i,j}(t)\right\}\le e^{-2(N(N-1)+1)\ln n_{i,j}}\le t^{-2(N(N-1)+1)}. \tag{29}$$

Last, we consider the inequality in (25). Let $f(\boldsymbol{\delta},\boldsymbol{w})=\sum_{i\in\mathcal{N}}h_i(\boldsymbol{\delta},\boldsymbol{w}_i)$ denote the aggregate number of successful downlink transmission function where $h_i(\boldsymbol{\delta},\boldsymbol{w}_i)=\delta_i\sum_{i\in\mathcal{N}}\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}w_{i,j}(t)\delta_j^*(t)\le 1\right]$ and $\boldsymbol{w}_i=[w_{i,1},\cdots,w_{i,N}]$. Note that the aggregate number of successful downlink transmission function $h_i(\boldsymbol{\delta},\boldsymbol{w}_i)$ is increasing function with respect to $\boldsymbol{w}_i$.

$$f(\boldsymbol{\delta}^*(t))-f(\boldsymbol{\delta}(t))-2\sum_{i\in\mathcal{N}}\left[\delta_i(t)\cdot\left(\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\eta_{i,j,c_{i,j}(t)}\right)\delta_j(t)\le 1\right]-\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\bar{w}_{i,j}(t)\delta_j(t)\le 1\right]\right)\right]$$
$$=f(\boldsymbol{\delta}^*(t))-f(\boldsymbol{\delta}(t))-2\sum_{i\in\mathcal{N}}\left[\delta_i(t)\cdot\left(\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\left(\bar{w}_{i,j}(t)-\sqrt{\frac{N\ln n_{i,j}}{c_{i,j}(t)}}\right)\delta_j(t)\le 1\right]-\mathbb{I}\left[\sum_{j\in\mathcal{N}\setminus\{i\}}\bar{w}_{i,j}(t)\delta_j(t)\le 1\right]\right)\right]$$
$$\ge f(\boldsymbol{\delta}^*(t))-f(\boldsymbol{\delta}(t))-2\sum_{i\in\mathcal{N}}\{h_i(\boldsymbol{\delta},\bar{\boldsymbol{w}}_i-\zeta_{\min}^2(\boldsymbol{\delta},\bar{\boldsymbol{w}}))-h_i(\boldsymbol{\delta},\bar{\boldsymbol{w}}_i)\}$$
$$\ge\theta_{\boldsymbol{\delta}(t)}-2\sum_{i\in\mathcal{N}_{\boldsymbol{\delta}(t)}}\frac{\theta_{\min}}{2N}\ge\theta_{\boldsymbol{\delta}(t)}-\theta_{\min}$$
$$\ge 0. \tag{30}$$

Then, under the given $\boldsymbol{\delta}$ and $\boldsymbol{w}_i$, there exist $\boldsymbol{\zeta}_i=[\zeta_{i,1},\cdots,\zeta_{i,N}]$ for all $i\in\mathcal{N}$ such that the following holds:

$$h_i(\boldsymbol{\delta},\boldsymbol{w}_i-\boldsymbol{\zeta}_i)-h_i(\boldsymbol{\delta},\boldsymbol{w}_i)=\frac{\theta_{\min}}{2N}, \tag{31}$$

where $\theta_{\boldsymbol{\delta}} = f(\boldsymbol{\delta}^*, \boldsymbol{w}) - f(\boldsymbol{\delta}, \boldsymbol{w})$ and $\theta_{\min} = \min_{\boldsymbol{\delta} \in \mathcal{A}_{1_N} \setminus \mathcal{A}_{\boldsymbol{\delta}^*}} \theta_{\boldsymbol{\delta}}$. Let $\zeta_{\min}(\boldsymbol{\delta}, \boldsymbol{w}) = \min_{i \in \mathcal{N}, j \in \mathcal{N} \setminus \{i\}} \zeta_{i,j}$. Then, for all $i \in \mathcal{N}$, the following holds:

$$h_i(\boldsymbol{\delta}, \boldsymbol{w}_i - \zeta_{\min}) - h_i(\boldsymbol{\delta}, \boldsymbol{w}_i) \leq \frac{\theta_{\min}}{2N}. \tag{32}$$

If we choose the integer $\kappa \geq \lceil \frac{(N(N-1)+1)\ln n}{\zeta^2_{\min}(\boldsymbol{\delta}, \boldsymbol{w})} \rceil$, the inequality in (25) does not hold due to the followings: Based on the above condition, the inequality in (25) dose not hold when $\kappa \geq \lceil \frac{(N(N-1)+1)\ln n}{\zeta^2_{\min}} \rceil$ where $\zeta^2_{\min} = \zeta^2_{\min}(\boldsymbol{\delta}, \bar{\boldsymbol{w}})$. As a result, based on (26), (29), and (30), the upper bound of the counter $C_i(T)$ is given by

$$\begin{aligned}
&\mathbb{E}\left[C_i(T)\right] \\
&\leq \left\lceil \frac{(N(N-1)+1)\ln n}{\zeta^2_{\min}} \right\rceil \\
&\quad \sum_{t=1}^{\infty} \left(\sum_{c_{1,2}=1}^{t-1} \cdots \sum_{c_{N,N-1}=1}^{t-1} \sum_{\bar{c}_{1,2}=1}^{t-1} \cdots \sum_{\bar{c}_{N,N-1}=1}^{t-1} 2Nt^{-2(N(N-1)+1)} \right) \\
&\leq \frac{(N(N-1)+1)\ln n}{\zeta^2_{\min}} + 1 + N \sum_{t=1}^{\infty} 2t^{-2} \\
&\leq \frac{(N(N-1)+1)\ln n}{\zeta^2_{\min}} + 1 + \frac{\pi^2}{3} N.
\end{aligned}$$

REFERENCES

[1] 3GPP, "Evolved universal terrestial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2," 3GPP tech. spec. TS 36.300, Ver. 10.8.0, July 2012.

[2] M. A. Wijaya, K. Fukawa, and H. Suzuki, "Intercell-interference cancellation and neural network transmit power optimization for MIMO channels," in *proc. IEEE Vehicular Technology Conference (VTC Fall)*, Sept. 6–9, 2015.

[3] A. Adeel, H. Larijani, A. Javed, and A. Ahmadinia, "Random neural network based power controller for inter-cell interference coordination in LTE-UL," in *proc. IEEE Int. Conf. Communication Workshop (ICCW)*, June 8–12, 2015.

[4] A. Adeel, H. Larijani, and A. Ahmadinia, "Resource management and inter-cell-interference coordination in LTE uplink system using random neural network and optimization," in *IEEE Access*, vol. 3, pp. 1963–1979, Oct. 2015.

[5] H. Li, Z. Liang, and G. Ascheid, "Artificial neural network aided dynamic scheduling for eICIC in LTE HetNets," in *proc. IEEE 17th Inf. Work. Signal Processing Advances in Wireless Communications (SPAWC)*, July 3–6, 2015.

[6] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for interference control in OFDMA-based femtocell networks," in *proc. IEEE Vehicular Technology Conference (VTC Spring)*, May 16–19, 2010.

[7] M. Simsek, M. Bennis, and I. Güvenç, "Learning based frequency- and time-domain inter-cell interference coordination in HetNets," in *IEEE Trans. Veh. Technol*, vol. 64, no. 10, pp. 4589–4602, Oct. 2015.

[8] W. Lu, Q. Fan, Z. Li, and H. Lu, "Power control based time-domain inter-cell interference coordination scheme in DSCNs," in *proc. IEEE Int. Conf. Communications (ICC)*, May 22-27, 2016.

[9] A. Feki and V. Capdevielle, "Autonomous resource allocation for dense LTE networks: A multi armed bandit formulation," in *proc. IEEE 22nd Int. Symp. Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept. 11–14, 2011.

[10] P. Coucheney, K. Khawam, and J. Cohen, "Multi-armed bandit for distributed inter-cell interference coordination," in *proc. IEEE Int. Conf. Communications (ICC)*, June 8–12, 2015.

[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," in *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Feb. 2002.

[12] G. R. Maccartney, T. S. Rappaport, M. K. Samimi, and S. Sun, "Millimeter-wave omnidirectional path loss data for small cell 5G channel modeling," in *IEEE ACCESS*, vol. 3, pp. 1573–1580, Aug. 2015.

[13] T. Bai and R. W. Heath, Jr., "Coverage and rate analysis for millimeter-wave cellular networks," in *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.

[14] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," in *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 410–425, Oct. 2012.

[15] M. Čierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On number of almost blank subframes in heterogeneous cellular networks," in *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5061–5073, Oct. 2013.

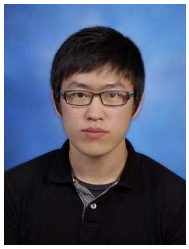
[16] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks," in *proc. IEEE Int. Conf. Communications (ICC)*, June 10–15, 2012.

[17] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," in *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–12, 1985.

[18] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," in *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, May 2002.

[19] L. A. Wolsey, "Maximizing real-valued submodular functions: primal and dual heuristics for location problems," in *Math. Oper. Res.*, vol. 7, no. 3, pp. 410–425, Aug. 1982.

Ryangsoo Kim received the B.S. degree in information and communications from Chungnam National University, Daejeon, Korea, in 2010, the M.S., and the Ph.D. degrees with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 2012 and 2017, respectively. He is currently a researcher with the Honam Research Center, Electronics and Telecommunications Research Institute (ETRI), Gwangju, Korea. His research interests include online learning algorithm for resource management, network protocol design and performance analysis for wired/wireless networks.



Yonggang Kim received his B.S. degree in the School of Computer and Telecommunications Engineering from Yonsei University, Wonju, Republic of Korea, in 2012, and his M.S. degree in the School of Information and Communications from Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, in 2014. He is currently pursuing a Ph.D. degree in the School of Electrical Engineering and Computer Science at GIST. His research interests include network resource optimization for next-generation wireless networks.

Nam Yul Yu (M'07) received the B.S. degree in electronics engineering from Seoul National University, Seoul, South Korea, in 1995, the M.S. degree in electronic and electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2000, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2007. From 2000 to 2003, he was with the Telecommunication Research and Development Center, Samsung Electronics, South Korea, where he worked on channel coding schemes for wireless communication systems. In 2007, he was a Senior Research Engineer with LG Electronics, South Korea, working on the standardization of the 3GPP-LTE. From 2008 to 2014, he was an Assistant/Associate Professor with the Department of Electrical Engineering, Lakehead University, Thunder Bay, ON, Canada. In 2014, he joined the Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently an Associate Professor with the School of Electrical Engineering and Computer Science. His research interests include compressed sensing, sequence design, communications, and cryptography. From 2009 to 2011, he served as an Associate Editor for Sequences in the IEEE Transactions on Information Theory.

Seung-Jun Kim (S'96-M'98-SM'12) received the B.S. and M.S. degrees from Seoul National University, Seoul, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree from the University of California at Santa Barbara in 2005, all in electrical engineering. From 2005 to 2008, he worked as a Research Staff Member at NEC Laboratories America, Princeton, NJ, USA. He was in the Digital Technology Center and the Department of Electrical and Computer Engineering, University of Minnesota, from 2008 to 2014, where his final title was the Research Associate Professor. Since 2014, he has been an Assistant Professor with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County. His research interests include statistical signal processing, optimization, and machine learning, with applications to wireless communication and networking, future power systems, and big data analytics. He is serving as an Associate Editor of IEEE Signal Processing Letters.

Hyuk Lim (M'03) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the School of Electrical Engineering and Computer Science, Seoul National University (SNU), Seoul, Korea, in 1996, 1998, and 2003, respectively. From 2003 to 2006, he was a postdoctoral research associate with the Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently a full professor with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Korea. His research interests include network protocol design, optimization, and performance evaluation of computer and communication networking systems.