Distributed Approximating Global Optimality with Local Reinforcement Learning in HetNets

Yawen Fan
Department of Electrical Engineering
and Computer Science
The University of Tennessee
Knoxville, Tennessee 37996
Email: yfan12@vols.utk.edu

Husheng Li
Department of Electrical Engineering
and Computer Science
The University of Tennessee
Knoxville, Tennessee 37996
Email: hli31@utk.edu

***Abstract*—This paper proposes a distributed multiagent reinforcement learning framework for resource allocation problems in heterogeneous wireless networks (HetNet). The base station, pico transmitters and user equipments collaborate to improve the system throughput and energy efficiency by handling the inter-cell interference based on the local observable environment. Utilizing the sparsity and tree structure of connection graph of the network, the collaboration is carried out through a distributed calculation of the optimal action via the message passing method. The tradeoff between the system throughput and the energy efficiency is addressed by a weighting parameter in the reward function. Simulation results show that the proposed framework outperforms the traditional inter-cell interference coordination method and improves overall performance.**

I. Introduction

We study the resource allocation problem in heterogeneous wireless networks (HetNets). The Inter-Cell Interference Coordination (ICIC), user association and energy efficiency are jointly considered for downlink transmissions. In a typical HetNet, the high-power macro network nodes are deployed for blanket coverage, while the pico nodes with small RF coverage areas are deployed for possible traffic hotspots or the edge area [1]. Picocells have a coverage of hundreds of meters. However, since the pico transmitters typically have low transmit powers and share the same frequency band with base-stations (BSs) in macrocells, there are two problems in the HetNets:

- Most user equipments (UEs) will connect to BSs in the macrocell by greedily choosing the highest received signal reference power (RSRP). The pico transmitters will connect to few UEs or even remain idle.
- The link between UE and pico transmitter may suffer severe inter-cell interference. [1]

The performance degradation in energy efficiency and overall throughput from the above problems may exacerbate in the future 5G networks, when the densities of BSs and pico transmitters increase.

The enhanced inter-cell interference coordination (eICIC) [2] is proposed to mitigate the above problems. With eICIC, each BS remains silent for certain subframes in each frame, termed Almost Blank Subframes (ABS), to reduce the interference from BS, as shown in Fig 1. Each UE is allowed to connect to the neighboring pico transmitter whose signal power is even less than the BSs. Moreover, to improve the energy efficiency within HetNets, a dynamic base station power allocation problem is considered, such that BSs could adjust the signal power when the neighboring wireless resource is abundant for the UE requests in the near future.

Previous studies are focused on the optimal user association policy, the adaptive number of ABS or effective BS sleep managements eparately [3], [4]. Jointly optimizing those three factors remains challenging since it has been proven to be an NP-hard problem [4], [5]. For example, [6] determined the optimal amount of radio resource that a microcell offers to picocells and association rules for each UE by relaxing the integer constraint.

Reinforcement learning is considered as an alternative solution. [7] proposed decentralized procedures for joint interference management and cell association mechanism. Their system models fall in the single-agent reinforcement learning framework (SARL), where the BS and pico transmitters are modeled as independent agents with no collaboration. However, previous theoretical analysis revealed that in SARL the agents may change their respective actions frequently, or oscillate between actions, such that the convergence to the optimal solution is not assured [8].

In this paper, we propose a distributed resource allocation framework to improve the system's overall performance, based on collaborative multiagent reinforcement learning using local observations. The multiagent method has been previously applied in cognitive radio [9], [10] for spectrum sensing. However, they only consider homogeneous models where the agents within the system are secondary users.

In our work, the HetNet is considered as the heterogeneous learning environment, while each agent is defined by its own state/action space and receives different rewards. BS learns the optimal number of ABS, pico transmitters learn the sleep mode and UEs learn the association policy based on previous experience. To carry out the collaboration, the distributed agents need to train the globally optimal strategy based on

[1] Although in this paper, we consider the single macrocell case, UE may suffer the interference from the neighboring transmitters. Such interference is considered as inter-cell interference.

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

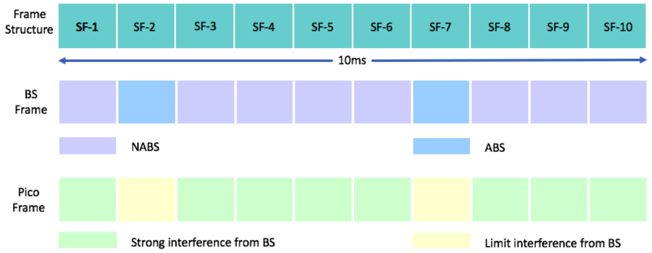

Figure 1: Illustration of ABS subframes in eICIC

their local observations. Inspired by the similarities between the wireless network topology and the structure of the low-density parity-check (LDPC) code (i.e., the corresponding graphs are both sparse), we employ distributed message passing based algorithm to solve the max-sum problem. The proposed framework is efficient and requires little communication overhead.

The remainder of this paper is organized as follows. In Section II, the system model is presented. In Section III, we formulate the resource allocation problem in the reinforcement learning framework, and in Section IV we explain the implementation of the proposed framework. In Section V, system level simulations are given, and Section VI concludes the paper.

II. System Model

A. Spatial Model

We focus on analyzing one macrocell network with multiple pico transmitters considering the downlink. Both BS $\mathcal{M}$ and pico transmitters $\mathcal{P}$ operate in the same frequency band with different transmission powers. The BS is located at the center of the cell, while pico transmitters are deployed to either the edge region or traffic hotspot. We consider a fixed number of UEs within the cell that are distributed uniformly in $\mathcal{R}^2$, according to homogeneous Poisson point processes (PPS) with density λ_U. A hotspot area has more UEs that are uniformly distributed.

B. Transmission Model

We assume that BS transmits with its maximum power P_m. The pico transmitters can switch between active and idle modes with transmission power P_{p_i} and 0, respectively. In practice, the received signal power could be accessed based on Reference Signal Received Quality(RSRQ). The received signal power in the downlink (DL) for UE u from BS m or pico transmitter p_i is given by

$$P^k_{re}(u) = P_k h_{u,k} L_k(k-u)^{-\alpha}, k \in \{\mathcal{M}, \mathcal{P}\}, \qquad (1)$$

where P_k is the transmission power of transmitter $k \in \{\mathcal{M}, \mathcal{P}\}$, h is the small scale fading power gain where we consider Rayleigh fading here, L is the pathloss and α is the pathloss exponent.

C. Interference Model

Consider a single UE u. Denote the power it receives from the connected transmitter by $P_{re}(u)$. Let $P^{macro}(u)$ and $P^{pico}_i(u)$ be the signal power from BS m and pico transmitter p_i, respectively. For those UEs connected to BS, the interferences come from nearby pico transmitters. Thus we have

$$SINR_m(u) = \frac{P_{re}(u)}{N + \sum_{p_i \in \mathcal{P}} P^{pico}_i(u)}. \qquad (2)$$

For those UEs u connected to the pico transmitter p_j, in the non-ABS subframes, the SINR is given by

$$SINR_{p,NA}(u) = \frac{P_{re}(u)}{N + \sum_{p_i \neq p_j} P^{pico}_i(u) + P^{macro}(u)}, \qquad (3)$$

while for ABS subframes the SINR is given by

$$SINR_{p_i,A}(u) = \frac{P_{re}(u)}{N + \sum_{p_i \neq p_j} P^{pico}_i(u)}. \qquad (4)$$

Denoting by F the total number of subframes in each frame and by F_A the number of ABS, we derive the rate for each UE, which is given by

$$R_m(u) = \frac{W}{N_m}\frac{F_A}{F}\log(1 + SINR_m(u)), \qquad (5)$$

and

$$R_{p_i}(u) = \frac{W}{N_{p_i}}\left[\frac{F_A}{F}R_{p_i,A}(u) + (1 - \frac{F_A}{F})R_{p_i,NA}(u)\right], \qquad (6)$$

where N_m is the number of UEs connected to BS m and N_{p_i} is the number of UEs connected to pico transmitter p_i. $R_{p_i,A}(u)$ and $R_{p_i,NA}(u)$ can be calculated based on (3) and (4) with the Shannon formula, respectively. Assume that, at each time slot, UE u is allowed to connect to one transmitter.

III. Optimal Resource Allocation

The problem is formulated as follows. Given a set of UEs $\mathcal{U}$, pico transmitters $\mathcal{P}$ and BS $\mathcal{M}$. $x^k(u)$ describes the connectivity between UE u and transmitter k: $x^k(u) = 1$ if u and k are connected and 0 otherwise. Define the connection success rate q_r as the portion of reliable links, which is given by

$$q_r = \frac{1}{|\mathcal{U}|}\sum_{u_j \in \mathcal{U}} \mathbf{1}_{R_k(u_j) > \beta}, \qquad (7)$$

and q_a as the portion of inactive pico transmitters, which is given by

$$q_a = \frac{1}{|\mathcal{P}|}\sum_{i=1}^{\mathcal{P}} \mathbf{1}_{P_{p_i} = 0}. \qquad (8)$$

We want to compute the number of ABSs, F_A, a binary decision on each UE u's association policy and the status of pico transmitters P_{p_i} so that the following optimization problem is solved:

$$\begin{aligned} \underset{F_A, x^k(u), P_{p_i}}{\text{maximize}} \quad & q_r + \alpha q_a \\ \text{subject to} \quad & \sum_{k \in \mathcal{M} \cup \mathcal{P}} x^k(u) = 1, \forall u \in \mathcal{U} \\ & F_A \in \boldsymbol{Z}^+ \quad \text{and} \quad F_A \leq F \\ & P_{p_i} \in \{0, P_0\} \end{aligned}, \tag{9}$$

where $\boldsymbol{Z}^+$ denotes the nonnegative integer and P_0 is the maximum transmission power for pico transmitters. α controls the weight between energy efficiency and throughput. When $\alpha \to 0$, we focus on the system's overall throughput while larger α indicates more consideration on the energy cost.

Solving (9) is NP-hard, as both F_A and $x^k(u)$ are integers. Here we consider a dynamic Q-learning procedure that only uses previous experiences to estimate the optimal value automatically from the environment.

A. *Q-learning*

Q-learning [11] is used to find the optimal state-action policy for any finite state Markov decision process. It has been applied to many fields for its guaranteed convergence to the optimal policy. Q-learning problem is characterized by the agent with its state $\mathcal{S}$, the set of action $\mathcal{A}$ per state and the reward $\mathcal{R}$, after the action is executed by the agent. A policy is the agent's choice of actions for each state. The goal of Q-learning is to find the optimal policy that maximizes the expected value of the total reward over all successive steps.

$$Q(s, a) = E\left\{ \sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) | s_0, a_0 \right\} \tag{10}$$

where $Q(s, a)$ is the quantity of the state-action pair (s, a). In HetNets, the players, state, action and reward functions are defined as

- **Player**: BS, pico transmitter and UE
- **State**: Each type of player has its own state space as $s_{\mathcal{M}}$, s_{p_i} and s_{u_j}

$$s_{\mathcal{M}} = \{F_A, N_m, N^r_{\mathcal{M}}\}$$

$$s_{p_i} = \{P_{p_i}, N_{p_i}, N^r_{p_i}\}$$

$$s_{u_j} = \{P^m_{re}(u_j), P^p_{re}(u_j), F_A, N_m, N^r_{\mathcal{M}}, P_{p_i}, N_{p_i}, N^r_{p_i}\}$$

- **Action**:

$$(a_{\mathcal{M}}, a_{p_i}, a_{u_j}) = (\Delta F_A, P_{p_i}, x^k(u_j))$$

- **Reward**: The total reward for the learning framework in time slot t is defined as

$$r^t = q^t_r + \alpha q^t_a + \eta q_p \tag{11}$$

  $q_p = \mathbf{1}_{F_A \notin [0, F]}$ is the penalty term when F_A is beyond the reason range.

The detailed descriptions for the above variables are summarized in Table I.

Table I: Descriptions about State/Action Variable

Variable	Description
F_A	Number of ABS
N_m	Number of UEs connected to BS
$N^r_{\mathcal{M}}$	Number of reliable links for BS
P_{p_i}	Transmission power for p_i
N_{p_i}	Number of UEs connected to pico transmitter p_i
$N^r_{p_i}$	Number of reliable links for pico transmitter p_i
$P^m_{re}(u_j)$	The received power from BS for u_j
$P^{p_i}_{re}(u_j)$	The received power from pico transmitter p_i for u_j
ΔF_A	The change of ABS

The standard algorithm for updating the Q-function at iteration t is given by

$$\begin{aligned} & Q(s_t, a_t) \\ = \quad & (1-\gamma) Q(s_t, a_t) + \gamma \left[r^t + c \max_a Q(s_{t+1}, a) \right] . \end{aligned} \tag{12}$$

At iteration t, the agent first checks the Q-table for the current state s_t. It finds the optimal action a^*_t with the largest Q value $Q(s_t, a_t*)$ and executes it. The system moves to the next state, as the response of a^*_t, and returns the reward. Then it updates the current Q function. γ controls the weight between the previous experience and the future reward. c is the learning rate.

Standard Q-learning considers the system as single agent. For the macrocell considered in this paper, if we model the whole system as the single agent, the overall state space is given by

$$\mathcal{S} = s_{\mathcal{M}} \times \boldsymbol{s_{\mathcal{P}}} \times \boldsymbol{s_{\mathcal{U}}} \tag{13}$$

where $\boldsymbol{s_{\mathcal{P}}} = s_{p_1} \times s_{p_2} ... \times s_{p_{|\mathcal{P}|}}$ and $\boldsymbol{s_{\mathcal{U}}} = s_{u_1} \times s_{u_2} ... \times s_{u_{|\mathcal{U}|}}$

Similarly the overall action space is given by

$$\mathcal{A} = a_{\mathcal{M}} \times \boldsymbol{a_{\mathcal{P}}} \times \boldsymbol{a_{\mathcal{U}}} \tag{14}$$

The state/action space size increases exponentially with the number of UEs and pico transmitters. For example, for a cell with 100 UEs, the overall size for the state space could be more than 10^{30}. Finding the globally optimal action becomes computational intractable. In the next section, we will design the distributed collaborative reinforcement learning framework to decompose and solve the problem efficiently.

IV. Distributed Collaborative Q-Learning

We model the wireless network as a coordination graph and decompose the Q-function of the whole network into multiple local Q-functions defined by the connection within the network. The max-sum problem in the process of updating Q-function is handled by the method of message passing.

A. *Coordination Graph*

Coordination graph [12] exploits the fact that in many problems only a few agents depend on each other and the large problem could be decomposed to simpler sub-problems. In a coordination graph, each node represents an agent and

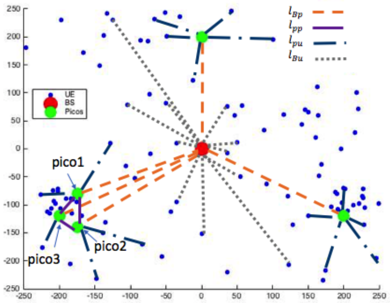

Figure 2: Coordination Graph for Macrocell Network

the edge defines the coordination dependency between the connected nodes. The dependency is further characterized by a reward function based on the agents' actions. The global reward function is the sum of all local reward functions. Instead of finding the globally optimal actions in one step, with the coordination graph, we solve the local sub problems with reduced state/action set cardinality in the first step and then compute the globally optimal action.

In the coordination graph, the global Q-function $Q(\boldsymbol{s}, \boldsymbol{a})$ is defined as the sum of all local edge Q-functions, namely

$$Q(\boldsymbol{s}, \boldsymbol{a}) = \sum_{(i,j)\in \boldsymbol{E}} Q_{i,j}(s_{i,j}, a_i, a_j) \tag{15}$$

where $s_{i,j} \subseteq s_i \cup s_j$ is the subset of the state variables that jointly describe the connected agent j and agent i in the graph. $\boldsymbol{E}$ is the set of edges in the graph. The local edge Q-function $Q_{i,j}$ depends only on the actions of agent i and agent j. Such a decomposition results in the shrink of the state/action space. We can use either the Q-value table or function approximation to represent the local Q-function.

B. Decomposing the Resource Allocation Problem

The macrocell network in this paper can be considered as a coordination graph, where UEs, pico transmitters and BS are the nodes. There are four types of edges in the system. The BS is connected to all the pico transmitters with the backhaul communications l_{Bp} and some of the child UEs by l_{Bu}. Pico transmitters have small transmission power, and their operational status is only dependent on the neighboring pico transmitters which are denoted by l_{pp}. They also connect to UEs by l_{pu}. Fig (2) shows the coordination graph for the macrocell considered in this paper. We omit parts of the connections between UEs and transmitters for clarification.

Observe in 15, the Q-function in the coordination graph is defined on the state space of two neighboring agent, we have to derive the formulation of local reward in order to update the strategy. Notice that the global reward function defined in (11) is separable, as it is the sum of indicator functions. We could distribute global reward to the edge according to the dependency.

- Each UE is connected to one transmitter. $\mathbf{1}_{R_k(u_j)\geq\beta}$ is the reward for each reliable link and thus could be directly assigned to the corresponding local Q-function Q_{k,u_j}.
- Each pico transmitter p_i receives reward when it is idle and does not connect to any UE. Therefore, the reward could be equally assigned to the edges between the idle pico transmitter and the neighboring transmitters.
- BS receives penalty q_p when F_A is beyond the reasonable range. We assume that every edge connected to the BS node contributes equally to this penalty and therefore the penalty is assigned equally to these connected edges.

Then for each type of edge we have the following updating procedure:

- Updating l_{Bp}:

$$\begin{aligned} & Q_{\mathcal{M},p_i}(s_{\mathcal{M},p_i}, a_{\mathcal{M}}, a_{p_i}) \\ = \ & (1-\gamma)Q_{\mathcal{M},p_i}(s_{\mathcal{M},p_i}, a_{\mathcal{M}}, a_{p_i}) \\ + \ & \gamma\left[\frac{\eta q_p}{N_{\mathcal{M}}} + \frac{\alpha r_{p_i}}{|\mathcal{P}|} + cQ_{\mathcal{M},p_i}(s'_{\mathcal{M},p_i}, a^*_{\mathcal{M}}, a^*_{p_i})\right] \end{aligned} \tag{16}$$

- Updating l_{Bu}:

$$\begin{aligned} & Q_{\mathcal{M},u_j}(s_{\mathcal{M},u_j}, a_{\mathcal{M}}, a_{u_j}) \\ = \ & (1-\gamma)Q_{\mathcal{M},u_j}(s_{\mathcal{M},u_j}, a_{\mathcal{M}}, a_{u_j}) \\ + \ & \gamma\left[\frac{\eta q_p}{N_{\mathcal{M}}} + \frac{\alpha r_{u_j}}{|\mathcal{U}|} + cQ_{\mathcal{M},u_j}(s'_{\mathcal{M},u_j}, a^*_{\mathcal{M}}, a^*_{u_j})\right] \end{aligned} \tag{17}$$

- Updating l_{pp}

$$\begin{aligned} & Q_{p_i,p_j}(s_{p_i,p_j}, a_{p_i}, a_{p_j}) \\ = \ & (1-\gamma)Q_{p_i,p_j}(s_{p_i,p_j}, a_{p_i}, a_{p_j}) \\ + \ & \gamma\left[\frac{\alpha r_{p_i}}{|\mathcal{P}|} + \frac{\alpha r_{p_j}}{|\mathcal{P}|} + cQ_{s_{p_i,p_j}}(s'_{p_i,p_j}, a^*_{p_i}, a^*_{p_j})\right] \end{aligned} \tag{18}$$

- Updating l_{pu}

$$\begin{aligned} & Q_{p_i,u_j}(s_{p_i,u_j}, a_{p_i}, a_{u_j}) \\ = \ & (1-\gamma)Q_{p_i,u_j}(s_{p_i,u_j}, a_{p_i}, a_{u_j}) \\ + \ & \gamma\left[\frac{\alpha r_{u_j}}{|\mathcal{U}|} + cQ_{s_{p_i,u_j}}(s'_{p_i,u_j}, a^*_{p_i}, a^*_{u_j})\right] \end{aligned} \tag{19}$$

In (16)-(19), we follow the standard update for Q-function but the reward received by each agent is equally devided and allocated to its connected edges. Now we prove that updating the local Q-function is equivalent to updating the global Q-function in (12).

Theorem 1. *Suppose each agent in the coordination graph only stores its local Q-function and receives the local edge reward. All the distributed agents have the access to the globally optimal action $\boldsymbol{a}^*$ at state $\boldsymbol{s}'$. Then the local updating procedure defined in (16), (17), (18), (19) is equivalent to updating the global Q-function in (12).*

Remark 1. *Here we assume that each agent has the access to the global predicted action $\boldsymbol{a}^*$ at state $\boldsymbol{s}'$. We will explain the feasibility in the next subsection that, with little communications between the transmitters, we can achieve this by using message passing algorithm even if each agent stores only its local Q-function.*

Proof. . From the definition of global Q-function in (15), we summarize the local updating procedure in (16), (17), (18) and (19) for all $l \in \boldsymbol{E}$ and obtain

$$\begin{aligned} & Q(\boldsymbol{s}, \boldsymbol{a}) \\ = \ & (1-\gamma) Q(\boldsymbol{s}, \boldsymbol{a}) + \gamma \left[q_r + \alpha q_a + \eta q_p + c Q(\boldsymbol{s}^*, \boldsymbol{a}^*) \right] \end{aligned} \tag{20}$$

The equation holds since the overall reward of the system is equal to the sum of all the rewards from the edges, and the sum of local Q-functions is equal to the global Q-function. Therefore, (20) is exactly the standard updating procedure for calculating the global Q-function in (12). □

C. Predicting Globally Optimal Action Using Local Q-functions

Note that in Equations (16), (17), (18), (19), the action $\boldsymbol{a}^*$ represents the optimal action that maximizes the global Q-function in state $\boldsymbol{s}'$. Since each agent only has access to its local Q-function and local reward, $\boldsymbol{a}^*$ cannot be obtained directly. In this subsection we will introduce the message passing algorithm to estimate $\boldsymbol{a}^*$ based on local Q-function in a distributed manner.

In state $\boldsymbol{s}'$, finding the globally optimal action based on local Q-function is equivalent to solving the following max-sum problem according to (15):

$$\max_{\boldsymbol{a} \in \mathcal{A}} Q(\boldsymbol{s}, \boldsymbol{a}) = \max_{\boldsymbol{a} \in \mathcal{A}} \sum_{l_{i,j} \in \boldsymbol{E}} Q_{i,j}(s'_{i,j}, a_i, a_j) \tag{21}$$

To solve the problem, the straightforward idea is to transfer all the local Q-functions to the BS and solve it in a centralized style. This requires transferring the whole local Q-function table through backhaul communications which is inefficient. In Fig 3, we illustrated the similarities between the coordination graph and the Markov Random Field (MRF) that solves the following Sum-Product problem and is widely used in statistical inference problems (such as LDPC decoding):

$$p(x_a) = \sum_{x \neq x_a} \prod_{\boldsymbol{E}} p(x_i, x_j) \prod_{\boldsymbol{V}} p(x_k) \tag{22}$$

The detailed comparison is summarized in Table II. Note that in MRF, the edge exists only when two variables are dependent, while in the coordination graph an edge exists when there is a link between the two agents. In practical situations, both of them are sparse graphs.

As the message passing method has achieved substantial success in many statistical inference problems, such as LDPC decoding and image denoising, we apply the message passing method to solve the Max-Sum problem defined in this paper.

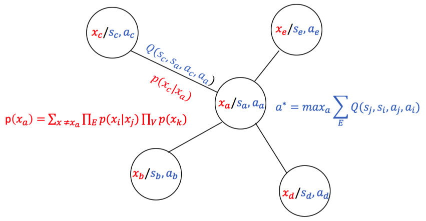
$x_c/s_c, a_c$
$x_e/s_e, a_e$
$Q(s_c, s_a, a_c, a_a)$
$p(x_c|x_a)$
$x_a/s_a, a_a$
$a^* = max_a \sum_E Q(s_j, s_i, a_j, a_i)$
$\mathrm{p}(x_a) = \sum_{x \neq x_a} \prod_E p(x_i|x_j) \prod_V p(x_k)$
$x_b/s_b, a_b$
$x_d/s_d, a_d$

Figure 3: Coordination graph and Markov Random Field

Table II: Comparison between Markov Random Field and Coordination Graph

	Markov Random Field	**Coordination Graph**
Problem	Sum-Product	Max-Sum
Edge	Conditional probability	Edge Q function
Node	Random Variable	Agent(BS,UE,etc)
Connection	Between dependent variables	Between linked agents
Sparsity	Yes	Yes

Briefly speaking, the message passing method is carried out by iteratively sending locally optimized messages to neighboring nodes. It is widely used in belief propagation (BP) or sum-product algorithm. Although in theory it only guarantees to converge when the graph is free of cycles, its empirical results on graphs with cycles in practical problems are surprisingly excellent.

Algorithm 1 Distributed Collaborative Q-learning for Resource Allocation in HetNets

```
Initialize the state of BS and pico transmitters.
UEs are connected to the neighboring transmitters with the maximum RSRP.
for Each UE do
    Transmit the local state to the connected transmitters
end for
Transmitters exchange their state information through backhaul communication
Transmitters calculate the local Q-function based on the received state information
Obtain the global optimal action through message passing and execute it in an $\epsilon-$greedy manner.
Update the local Q-function with the received reward
if Q-function change less than the threshold then
    Obtain the optimal policy under the current situation
end if
```

Since we use Q-table to represent the Q-function, the message from agent i to agent j is a table of the action, which is defined as

$$\mu_{i,j}(a_j) = \max_{a_i} \left\{ Q_{i,j}(a_i, a_j) + \sum_{k \in \Gamma(i)/j} \mu_{k,i}(a_i) \right\} - c_{i,j} \tag{23}$$

where $\Gamma(i)/j$ is the neighbor of i except j, and $c_{i,j}$ is the normalizing factor [13] to guarantee the convergence. After the convergence, the global optimal action for agent i is obtained by simply solving

$$a_i^* = \arg\max_{a_i} \sum_{k \in \Gamma(i)} \mu_{k,i}(a_i) \tag{24}$$

Table III: Simulation Parameters

Variable	Value
Macro cell radius	289m
Carrier Frequency	2.0 GHz
Bandwidth	10MHz
Thermal noise density	-174 dBm/Hz
Number of BS	1
Number of pico transmitters	5
BS path loss model	128.1 + 37.6log_{10}(D) dB (D[km])
Picos path loss model	140.1 + 36.7log_{10}(D) dB (D[km])
Channel	Reyleigh fading
Number of UE	200
BS power	46 dBm
Pico transmitter power	30 dBm

Table IV: Max-Sum result

	Synthetic Graph	Macro Graph
Number of edge	6	206
Number of node	5	208
$\max Q$ by Exact searching	1	1
$\max Q$ by Message passing	1	0.87

Here we omit s' in the Q-function since it is fixed during the message passing.

For each UE, since it is allowed to connect to only one transmitter, it could be considered as the leaf in a tree. Its parent transmitter can simply collect its current state and calculate the local Q-function. Thus there is no communication overhead between the UE and the transmitter. For pico transmitters, they first collect the UEs states and calculate the local Q-function. Then the global optimal action is obtained by the message passing. Note that the messages are passed only between the transmitters through backhaul communication link. This could be considered as negligible communication overhead. The detailed procedure of the proposed distributed collaborative Q-learning can be found in Algorithm 1. To implement the proposed framework in practice, the BS and pico transmitters have to store the connected UE in eahc time slot and each UE is able to evaluate the received signal power from the neighboring transmitters.

V. Numerical Results

The simulation parameters are given in Table III.

A. Max-Sum Result

We first check the performance for the message passing algorithm introduced in Section IV for the coordination graph. The results on the simple synthetic graph and, the graph in Fig 2 are summarized in Table IV. The synthetic graph is simple, which is composed of one parent node and 5 child nodes. Each child node is connected to the parent node and has no other edge. The demonstration of this simple graph is omitted due to the limit of the space.

We normalize the result by setting the optimal Q-value as 1. For the synthetic graph, the message passing method can find the optimal Q-value as the exact search method. For the graph in Fig 2, the message passing method can only find the suboptimal Q-value. The degraded performance can be explained by the topology of the Fig 2. Notice that, in the bottom left of the graph, there is a loop that contains the three pico transmitters according to their mutual dependency. It is well known that the message passing method can find the exact solution in the tree structure while may not find the optimal solution in the loopy case [14]. However, the exact search method requires the computational cost in the order of $O(A^N)$ while for the message passing method the computational cost is $O(AN)$. Here A is the number of actions for each node and N is the number of nodes. We improve the scalability at the cost of accuracy, wheras we will show in the next subsection that it can find the near-optimal result in the multiagent reinforcement learning.

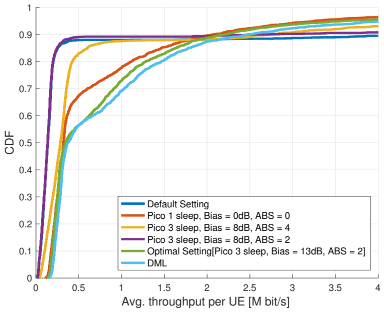

Figure 4: Rate distributions for different settings

B. Simulation Results

Simulation results for the system level performance are presented in Fig 4 in terms of the UE rate distribution. To demonstrate the impact of each agent's action, some reference models are considered. In the default setting, there is no bias for UE and each UE connects to the transmitter with the highest received power. All pico transmitters are active and there is no ABS in the BS's frame to handle the ICIC. By thoroughly searching the parameter space $\{F_A, Bias, P_{p_i}\}$, we obtain the optimal setting with the highest reliable link rate.

It can be observed that deactivating one of the pico transmitter could improve the system's performance both in the reliable link rate and average throughput. This is as the result of the reduce of interference in the hotspot area. To improve the energy efficiency, among all three pico transmitters in the hotspot area, pico 3 could be deactivated according to our simulation. Given the bias term in reasonable range (2dB, 15dB), the optimal number of muted subframes is 2. The optimal bias term for remaining pico transmitters is 13dB.

Our proposed Dynamic Multiagent Learning (DML) framework outperforms the optimal setting regarding the reliable link rate. The F_A and sleep pico are the same as the optimal setting. The better performance could be explained by the fact that in this framework, there is no fixed bias term for each

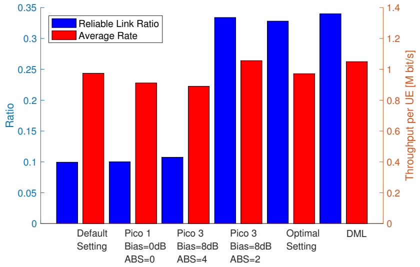

Figure 5: Performance comparison for different settings

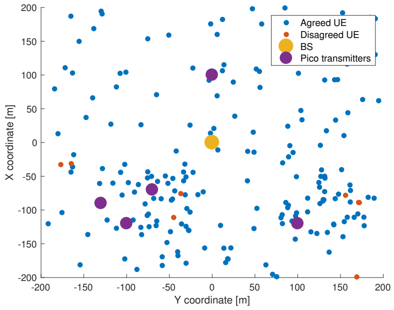

Figure 6: UE's different choice for connections

UE. Instead, in DML, there is no biased term for UE. UE determines the preference according to the received signal power and the transmitters' operational statue dynamically. UE may refuse to connect to the transmitter who has larger (biased) signal power in the classic condition. It can be further observed that the averaged rate in DML is slightly less than the setting when we have the fixed bias term according to Fig 5. The rationale is that the reward function in DML is related to whether the link is reliable, which is more focused on the fairness instead of the overall throughput. Therefore, it is sub-optimal in terms of the overall throughput. We claim that our framework can be easily adapted to maximize the overall throughput by changing the definition of the reward in (11).

Fig 6 demonstrates the connection status in one of the simulations. We compare the UE's choice between the optimal setting and DML setting. The UEs' choices agree under most conditions. The disagreement occurs when the UE is located at the marginal area, where in the optimal setting the biased signal power from pico transmitter is slightly less than the signal power from BS. In DML, a UE still chooses to connect to the pico transmitter; as the result the system level performance is improved.

VI. Conclusion

In this paper, we have investigated the application of multiagent reinforcement learning on the HetNets resource allocation problem. The proposed distributed learning framework can find the optimal strategy for UE, pico transmitter and BS, compared with the static ICIC solutions. Further research may include the investigation on generalizing the proposed framework, where the distribution of the UE is not fixed, and the transmitters can apply different levels of power to ABS instead of muting them. We also believe that a more careful design of the state space for the agent may potentially improve the performance. Besides, we will consider the blockage model and the beamforming, which can be integrated into the future 5G framework.

References

[1] D. Lopez-Perez, I. Guvenc, G. De la Roche, M. Kountouris, T. Q. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, 2011.

[2] K. I. Pedersen, Y. Wang, S. Strzyz, and F. Frederiksen, "Enhanced intercell interference coordination in co-channel multi-layer lte-advanced networks," *IEEE Wireless Communications*, vol. 20, no. 3, pp. 120–127, 2013.

[3] C.-H. Liu and L.-C. Wang, "Optimal cell load and throughput in green small cell networks with generalized cell association," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1058–1072, 2016.

[4] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, "Optimal joint base station assignment and beamforming for heterogeneous networks." *IEEE Trans. Signal Processing*, vol. 62, no. 8, pp. 1950–1961, 2014.

[5] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "Phy-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 037–10 047, 2016.

[6] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM transactions on networking*, vol. 22, no. 1, pp. 137–150, 2014.

[7] M. Simsek, M. Bennis, and A. Czylwik, "Dynamic inter-cell interference coordination in hetnets: A reinforcement learning approach," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 5446–5450.

[8] K.-L. A. Yau, P. Komisarczuk, and P. D. Teal, "Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues," *Journal of Network and Computer Applications*, vol. 35, no. 1, pp. 253–267, 2012.

[9] J. Lundén, S. R. Kulkarni, V. Koivunen, and H. V. Poor, "Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 858–868, 2013.

[10] B. F. Lo and I. F. Akyildiz, "Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on*. IEEE, 2010, pp. 2244–2249.

[11] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[12] C. Guestrin, D. Koller, and R. Parr, "Multiagent planning with factored mdps," in *NIPS*, vol. 1, 2001, pp. 1523–1530.

[13] M. Wainwright, T. Jaakkola, and A. Willsky, "Tree consistency and bounds on the performance of the max-product algorithm and its generalizations," *Statistics and computing*, vol. 14, no. 2, pp. 143–166, 2004.

[14] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.