1-1-2018

Joint Power Allocation in Interference-Limited Networks via Distributed Coordinated Learning

Roohollah Amiri
Boise State University

Hani Mehrpouyan
Boise State University

David Matolak
University of South Carolina

Maged Elkashlan
Queen Mary University of London

Joint Power Allocation in Interference-Limited Networks via Distributed Coordinated Learning

Roohollah Amiri*, Hani Mehrpouyan*, David Matolak†, Maged Elkashlan‡

**School of Electrical and Computer Engineering, Boise State University, {roohollahamiri,hanimehrpouyan}@boisestate.edu*

†Department of Electrical Engineering, University of South Carolina, matolak@cec.sc.edu

‡School of Electronic Engineering and Computer Science, Queen Mary University of London, maged.elkashlan@qmul.ac.uk

***Abstract*—Dense deployment of small base stations (SBSs) is one of the main methods to meet the 5G data rate requirements. However, high density of independent SBSs will increase the interference within the network. To circumvent this interference, there is a need to develop self-organizing methods to manage the resources of the network. In this paper, we present a distributed power allocation algorithm based on multi-agent Q-*learning* in an interference-limited network. The proposed method leverages coordination through simple message passing between SBSs to achieve an optimal joint power allocation. Simulation results show the optimality of the proposed method for a two-user case.**

I. Introduction

Ultra-densification through the use of smaller base stations is a promising technology in the next generation of cellular networks (5G) [1]. The small base stations (SBSs) might be mounted by users in a plug-and-play fashion, and their backhaul may be supported by broadband connections. The user-mounted feature, introduces unplanned deployment of SBSs, which may result in unavoidable co-channel interference.

The problem of power allocation in an interference-limited network has been investigated widely in the literature. In [2] and [3], the optimal power allocation for a two-user interference channel is derived for sum and individual power constraints, respectively. In [4] a more general solution is proposed for multi-transmitter systems with individual power constraints. The solution depends on the signal-to-interference-plus-noise ratio (SINR) value. In high SINR regime, the optimal solution is derived through transforming the problem into a geometric programming (GP) problem, while in the low SINR regime, a heuristic solution based on solving multiple GPs is used. It is important to note that all of these prior approaches are based on interior point methods. Hence, they require a centralized network management approach which may be impossible in dense networks. In [4], a distributed method based on decomposing the optimization problem into local problems is proposed. The solution is based on message-passing and applies to high SINR case with full channel state information (CSI). Nonetheless, in a dense plug-and-play network, with a changing architecture, the assumptions of high SINR and the availability of full CSI at all nodes may not hold.

In an ultra-dense network, in which the architecture of the network changes sporadically, a self-organizing method is a viable solution to manage the network resources. To this end, cooperative multi-agent reinforcement learning (MARL) methods have been used in resource management of communication networks [5]–[9]. Radio measurements such as SINR, are part of the Big data in cellular network [10], and one of the main advantages of MARL solutions is to utilize the measured SINR values. Generally most of the classic optimization solutions are based on channel coefficients. Thus, the prior methods require full CSI to find the solution while the MARL methods only need access to existing radio measurements, i.e., the measured SINR values. However, the existing MARL approaches in communication network management do not address the optimality of their cooperation methods. This is an important research topic to address since finding the optimal joint power allocation is directly impacted by the nature of the cooperation approach.

In this paper, we find an optimal joint power allocation solution via coordination between deployed SBSs. To address the optimality of the MARL approach, we model the whole system as a Markov decision process (MDP) with the SBSs being represented as the agents of the MDP. Subsequently, the value function of the MDP is approximated by a linear combination of local value functions of the SBSs. As we mentioned before, in order to remove the need for access to CSI, and develop an adaptable algorithm that handles a changing network architecture, each SBS uses a model-free reinforcement learning approach, i.e., Q-*learning*. Q-*learning* is used to update the SBS's local value function. Subsequently, we leverage the ability of SBSs to communicate over the backhaul network to build a simple message passing structure to coordinate them, based on variable elimination [11]. Finally, we propose a distributed algorithm which finds an optimal joint power allocation in an interference-limited network.

The paper is organized as follows. In Section II, the system model is presented. Section II-A first introduces the optimization problem, then analyzes the convexity of the problem. Section III presents the general framework of the proposed solution. Section IV outlines the proposed solution while Section V presents simulation results. Finally, Section VI concludes the paper.

II. Network Model

This paper considers downlink transmission in a dense deployment of N small base stations (SBSs). We assumed each SBS supports one user equipment (UE), and all SBSs share the same frequency resource block. This system can represent a single cluster of a large network, which uses different frequency in each cluster to avoid interference between clusters. It is also assumed the SBSs are interconnected via a backhaul network supported by, for example, a broadband connection. Here, we use the same model of interference as [2]. Thus, the received signal at the ith UE, r_i is given by

$$r_i = \sqrt{g_i P_i} d_i + \sum_{j \in D_i} \sqrt{g_i P_j \beta_{ji}} d_j + n_i, \tag{1}$$

where g_i represents the channel gain between the ith SBS and the UE it is serving, d_i is the transmitted signal from the ith SBS, P_i is the transmitted power at the ith SBS, D_i represents the set of interfering SBSs to the ith UE, β_{ji} $(0 \leq \beta_{ji} \leq 1)$ for $1 \leq i \leq N$ and $j \in D_i$ is the ratio of the unintended power of the jth SBS when measured at the ith UE, and n_i is the zero mean additive white Gaussian noise (AWGN) at the ith UE with variance σ^2.

According to the signal representation in (1), the SINR at the ith UE, SINR_i, can be determined as

$$\text{SINR}_i = \frac{g_i P_i}{\sum_{j \in D_i} g_i P_j \beta_{ji} + \sigma^2}, \tag{2}$$

and the throughput at the ith UE normalized by the transmission bandwidth, R_i, is calculated as

$$R_i = \log_2 \left(1 + \text{SINR}_i\right). \tag{3}$$

A. Problem Analysis

Let us define $\underline{\mathbf{P}} = \{P_1, P_2, ..., P_N\}$ as the set containing the transmitted power of the SBSs. The goal of the optimization is to find the optimal joint power allocation between SBSs, $\underline{\mathbf{P}}^* = \{P_1^*, P_2^*, ..., P_N^*\}$, that maximizes the total throughput of the network. The optimization problem ($\mathbf{OP}_1$) can be formulated as

$$\underset{\underline{\mathbf{P}}}{\text{maximize}} \quad \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \log_2 \left(1 + \text{SINR}_i\right), \tag{4a}$$

$$\text{subject to} \quad P_i \leq P_{i,max}, \; i = 1, \ldots, N. \tag{4b}$$

Here, the objective function in (4a) maximizes the sum throughput of the network. The constraint (4b) refers to the individual power limitation of every SBS.

The objective function in (4a) contains the interference term in the denominator of SINR term. In a dense network the interference term cannot be ignored [12]. Due to the presence of the interference term, the objective function (4a) is a nonconcave function [13], which leads to non-convexity of the optimization problem.

III. Distributed Coordinated Q-*learning*

In this section, the proposed optimal solution based on the Markov decision process (MDP) is presented. Then, the dimensionality issues of the optimal solution will be investigated. The dimensionality is important since it affects the tractability of the problem. Next, we use the coordination method introduced by [11] to solve the problem in a distributed fashion. We show that the resulting method, provides a joint solution for the MDP via message passing between the agents of the network.

A. Optimal Solution via Q-learning

Consider a system with N agents, where each agent j selects its actions from its action set, A_j. Further, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ is the set of state variables which define the state of the system. Let us denote $\underline{x} \subset X$ to represent a single state of the system. In a fully cooperative game, we look for an optimal joint solution that is a Pareto optimal Nash equilibrium. One obvious solution to this problem is to model the whole system as a large MDP with its action set representing the joint action set of all the agents in the system. We consider $\mathbf{A}$ as the joint action set of all the agents, and $\underline{a} \subset \mathbf{A}$ as a single joint action of this set.

The MDP framework will be modeled as $(\mathbf{X}, \mathbf{A}, Pr, \mathbf{R})$, where $\mathbf{X}$ denotes the finite set of states of the system, $\mathbf{A}$ is a finite set of joint actions, Pr is the transition model which represents the probability of taking action $\underline{a}$ at state $\underline{x}$ and ending up in state $\underline{x}'$, $Pr\left(\underline{x}, \underline{a}, \underline{x}'\right)$, and $\mathbf{R}$ is the immediate reward received by taking action $\underline{a}$ at state $\underline{x}$, $\mathbf{R}\left(\underline{x}, \underline{a}\right)$.

A policy, $\pi : \underline{x} \rightarrow \underline{a}$, for an MDP is defined as a strategy which shows at state $\underline{x}$, action $\pi\left(\underline{x}\right)$ will be taken. In order to evaluate a policy, a value function $V\left(\underline{x}\right)$, is defined which defines the value of policy at each state. In order to compute the value function for a given policy, we need to calculate the action-value function, also known as Q-function, defined as follows

$$\mathbf{Q}\left(\underline{x}, \underline{a}\right) = \mathbf{R}\left(\underline{x}, \underline{a}\right) + \gamma \sum_{\underline{x}'} Pr\left(\underline{x}' | \underline{x}, \underline{a}\right) V\left(\underline{x}'\right), \tag{5}$$

in which $\gamma \in [0, 1]$ is a discount factor. The optimal value at state $\underline{x}$ is the maximum value that can be reached by taking any action at this state. The optimal value function V^*, which gives the optimal policy π^*, satisfies the Bellman operation as follows [14]

$$V^*\left(\underline{x}\right) = \max_{\underline{a}} \mathbf{Q}^*\left(\underline{x}, \underline{a}\right). \tag{6}$$

Q-*learning* is a model-free reinforcement learning, which solves the Bellman equation through direct observations without knowledge of the transition model. In Q-*learning*, the agent observers the state, $\underline{x}$, takes an action, $\underline{a}$, receives a reward, $\mathbf{R}$, and ends in a next state, $\underline{x}'$. Then, it will update its Q-function as follows

$$\mathbf{Q}\left(\underline{x}, \underline{a}\right) = \mathbf{Q}\left(\underline{x}, \underline{a}\right) + \alpha[\mathbf{R}\left(\underline{x}, \underline{a}\right) + \gamma \max_{\underline{a}'} \mathbf{Q}\left(\underline{x}', \underline{a}'\right) - \mathbf{Q}\left(\underline{x}, \underline{a}\right)], \tag{7}$$

where, α is the learning rate of the algorithm. If any action-state pair is repeatedly visited, the Q-function will converge to the optimal value [15].

One issue with this method is that the size of the joint action set is exponential with respect to the number of agents. If there are N agents in the network, and each one has $|A|$ number of actions as the size of their action set, the size of the joint action set, $|\mathbf{A}|$, will be $|A|^N$. The exponential size of the joint action set makes the computation of the Q-function expensive and in most cases intractable.

B. Factored MDP

In most cases, for both representational and computational advantages, the state and action sets of an MDP can be factored into subsets based on the structure of the problem [16]. In large MDPs, the global Q-function can be approximated by the linear combination of local Q-functions, i.e. $\mathbf{Q} = \sum_j Q_j(\underline{a_j})$ [11]. The jth local Q-function, Q_j, has the joint action set which is a subset of the global joint action set, $\mathbf{A}$. Here, we will define the joint action set of Q_j by $Scope\,[Q_j] \subset \mathbf{A}$ for which $\underline{a_j}$ is a single joint action of this set.

In a communication network, each SBS plays the role of an agent in the multi-agent network. The action of SBS j, is the transmit power, P_j, that is used to transmit its signal to the intended user. From this point, an agent in a communication network, refers to the SBS. Generally, in wireless communication systems, each access point receives interference from specific local access points. Therefore, the approximation of global Q-function by linear combination of local Q-functions, applies to interference-limited communication networks.

C. Decomposition of Global Q-function

The decomposition of the global Q-function, relies on the dependencies between the agents of the network. These dependencies can be represented by *coordination graphs* (CGs) [11]. Generally, there are two decomposition methods: agent-based and edge-based. The agent-based decomposition provides a suitable architecture for a distributed system with exact solution, while the edge-based decomposition is recommended for CGs with densely connected nodes [17] and provides suboptimal solution. In this paper we will choose the agent-based decomposition since we are focused on achieving the optimal solution.

In a wireless network, the $Scope\,[Q_j]$ for agent j, is determined based on the interference model of the system, which is related to set D in (1). For example, in Fig. 1, four agents interfere with each other. Assume that agent A_1, receives interference from A_2 and A_3, and A_4 receives its interference from A_2 and A_3. Based on this model, the CG of the system is shown in Fig. 1. Each edge between agents, shows a dependency between the two agents.

Here, we assume that all agents have the same state $\underline{x}$, hence, $Q\,(x, a)$ is written as $Q\,(a)$. According to the CG in Fig. 1, the global Q-function, $\mathbf{Q}\,(\underline{\mathbf{a}})$, can be written as

$$\mathbf{Q}(\underline{\mathbf{a}}) = Q_1(a_1, a_2) + Q_2(a_2, a_4) + Q_3(a_1, a_3) + Q_4(a_3, a_4). \tag{8}$$

D. Coordinated Action Selection

In multi-agent Q-*learning*, according to (7), the agents will choose a joint action that maximizes the global Q-function. By using the agent-based decomposition, the joint action selection at state $\underline{x}$, $\max_{\underline{a}} \mathbf{Q}\,(\underline{a})$, is written as

$$\max_{a_1,a_2,a_3,a_4} Q_1(a_1, a_2) + Q_2(a_2, a_4) + Q_3(a_1, a_3) + Q_4(a_3, a_4). \tag{9}$$

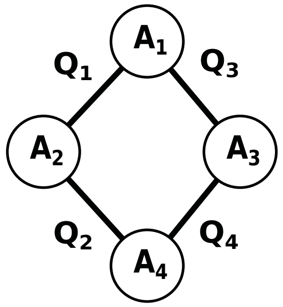

Fig. 1: Coordination graph.

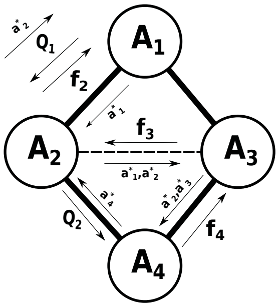

Fig. 2: Message passing.

This maximization problem, can be solved via variable elimination (VE) algorithm, which is basically similar to variable elimination in a Bayesian network [18]. Here, we will review this method for the network in Fig. 1. The key idea is to maximize over one variable at a time, find conditional solutions, passing conditional functions to other agents, and sending back the results of local optimization to the related agents to recover their joint action choices.

We start from agent A_4. a_4 influences Q_2 and Q_4, so the maximization problem can be written as

$$\max_{a_1,a_2,a_3} Q_1(a_1, a_2) + Q_3(a_1, a_3) + [\max_{a_4} Q_2(a_2, a_4) + Q_4(a_3, a_4)]. \tag{10}$$

Agent A_2 communicates Q_2 to A_4, and A_4 solves its local maximization, which results in two functions: $f_4\,(a_2, a_3)$, and $b_4\,(a_2, a_3)$. These functions are defined as follows

$$f_4\,(a_2, a_3) = \max_{a_4} Q_2\,(a_2, a_4) + Q_4\,(a_3, a_4)\,, \tag{11}$$

$$b_4\,(a_2, a_3) = \arg\max_{a_4} Q_2\,(a_2, a_4) + Q_4\,(a_3, a_4)\,. \tag{12}$$

At his stage, the A_4 has a conditional solution for a_4 based on a_2, and a_3, represented as the function b_4. Therefore, A_4 keeps b_4 and sends f_4 to its connecting agent, A_3. Then, A_4 is removed from the CG, and the maximization problem is translated to

$$\max_{a_1,a_2,a_3} Q_1\,(a_1, a_2) + Q_3\,(a_1, a_3) + f_4\,(a_2, a_3)\,, \tag{13}$$

f_4 brings a new edge in the coordination graph, an induced edge, which is shown with dashed line between A_2 and A_3 in Fig. 2. The next agent to be removed is A_3. The maximization problem is rewritten as

$$\max_{a_1,a_2} Q_1\,(a_1, a_2) + \left[\max_{a_3} Q_3\,(a_1, a_3) + f_4\,(a_2, a_3)\right]. \tag{14}$$

With the same procedure, A_3 introduces $f_3\,(a_1, a_2)$, and $b_3\,(a_1, a_2)$. Accordingly, the problem reduces to

$$\max_{a_1,a_2} Q_1\,(a_1, a_2) + f_3\,(a_1, a_2)\,. \tag{15}$$

Next agent to choose its action is A_2, for which the problem results in

$$f_1 = \max_{a_1} f_2\,(a_1)\,, \tag{16}$$

where, $f_2\,(a_1) = \max_{a_2} Q_1\,(a_1, a_2) + f_3\,(a_1, a_2)$. Finally, A_1 chooses its action based on maximizing the function $f_2\,(a_1)$. The results at this stage are f_1, and a_1^*. f_1 represents the maximum value of the global Q-function over a_1, a_2, a_3, and a_4, and a_1^* is the optimal joint action for A_1. To recover the joint action choices, A_1 sends a_1^* to A_2. Then A_2 chooses its action, $a_2 = b_2(a_1^*)$, and sends a_1^*, a_2^* to A_3. A_3 and A_4 will choose their actions with the same procedure, $a_3^* = b_3(a_1^*, a_2^*)$, and $a_4^* = b_4(a_2^*, a_3^*)$.

In general, the elimination algorithm maintains a set of functions in each step. It starts with all local functions, $\{Q_1, Q_2, ..., Q_N\}$, and eliminates agents one by one.

E. Local Update Rule

After finding the joint action, each agent will update its local Q-function. The update rule in (7) can be written as

$$\sum_j Q_j\left(\underline{x},\underline{a_j}\right) = \sum_j Q_j\left(\underline{x},\underline{a_j}\right) + \alpha\left[\sum_j R_j\left(\underline{x},\underline{a_j}\right) + \gamma \max_{\underline{a}} \sum_j Q_j\left(\underline{x}',\underline{a}'\right) - \sum_j Q_j\left(\underline{x},\underline{a_j}\right)\right], \tag{17}$$

where, the joint maximization is solved through VE according to the last section. By assuming $\underline{a}^*$ as the solution to the VE, and $\underline{a_j}^* \subset \underline{a}^*$ as the optimal joint action set for Q_j, the update rule for each local Q-function can be derived as

$$Q_j(\underline{x},\underline{a_j}) = Q_j(\underline{x},\underline{a_j}) + \alpha[R_j(\underline{x},\underline{a_j}) + \gamma Q_j(\underline{x}',\underline{a_j}^*) - Q_j(\underline{x},\underline{a_j})]. \tag{18}$$

The Fig. 2 illustrates all messages passed between the agents to solve VE and update local Q-functions.

IV. Power Allocation Using Coordinated Q-*Learning* (Q-CoPA)

To integrate the idea of coordinated multi-agent learning into a communication network, we will model the SBS as an agent, and the whole network as a multi-agent MDP. The goal of the agents is to maximize total throughput of the network, as a cooperative game.

A. Q-CoPA Algorithm

The proposed solution of this paper, *Q-CoPA*, can be summarized as follows
The interference model of the network will be used to derive the coordination graph of the agents. The entire network is modeled as an MDP, and the global Q-function of the MDP is approximated by linear combination of local Q-functions of the agents. Each agent, based on the coordination graph, knows its Scope. Local Q-functions are learned by the agents using cooperative Q-learning. The cooperation method between the agents is to maximize the summation of local Q-functions by choosing an appropriate joint action. This action selection is implemented using variable elimination and message passing between the agents. The backhaul of the network is used as the infrastructure of message passing.

The proposed method is represented in Algorithm 1.

In the Algorithm 1, the loops at lines 5 and 10 are independent, and will be executed in parallel by the agents.

Algorithm 1 The proposed Q-CoPA algorithm

```
 1: Initialize $\underline{x}$
 2: Initialize All $Q_j(\underline{x},\underline{a_j})$ arbitrarily
 3: for all episodes do
 4:     Choose $\underline{a}^*$ according to VE
 5:     for all agents do
 6:         Take action $\underline{a_j}$, observe $R_j$
 7:     end for
 8:     Observe $\underline{x}'$
 9:     Calculate $\max_{\underline{a}'} \mathbf{Q}$ according to VE
10:     for all agents do
11:         Update local Q-function according to Eq. 18
12:     end for
13:     $x_j \leftarrow x_j^{'}$
14: end for
```

B. Q-learning Parameters

In the following the actions, and the reward of the Q-*learning* algorithm implemented by each agent is defined.

- *Actions* : Each SBS has a set of actions, which is defined as the transmit power levels. We define this set as $\{p_1, p_2, ..., p_{N_{power}}\}$. The number of power levels is defined as N_{power}.
- *Reward* : In each episode, SBS chooses a power level, and transmits its data to its intended user. The user measures the SINR of the signal, and will feedback it to the SBS. Then the reward of the SBS j is calculated as $r_j = \log_2(1 + \text{SINR}_j)$.

V. Simulation Results

We consider two SBSs, each supporting one UE, with interfering channels. Each transmitter has omni-directional antenna and separate power source. The channel model is assumed to be time-invariant, i.e. slow fading. The channel gains are assumed to be $g_1 = 2.5$, and $g_2 = 1.5$. The $P_{1,max} = 10$ dBm, $P_{2,max} = 13$ dBm, and $\sigma^2 = 0$ dBm. Without loss of generality we assume that $\beta_{1,2} = \beta_{2,1} = \beta$ in Eq. 1. The objective of the optimization is to find the power allocation to maximize the sum throughput of the network under individual power constraints.

In executing the Q-CoPA algorithm, each Q-function is defined as a table, Q-table. The learning rate is $\alpha = 0.5$, the discount factor as $\gamma = 0.9$, $N_{\text{power}} = 100$, and the maximum number of episodes is set to 50 times the size of a Q-table. The MDP of this problem is assumed to be stateless. The actions of agents are the transmit powers, $a_1 = P_1$, and $a_2 = P_2$, Q-functions are defined as: $Q_1(P_1, P_2)$ and $Q_2(P_1, P_2)$, and the global Q-function is defined as: $\mathbf{Q}(P_1, P_2) = Q_1(P_1, P_2) + Q_2(P_1, P_2)$.

According to [3], the optimal power allocation to maximize the sum-rate of the above network is derived as

$$(P_1^*, P_2^*) = \begin{cases} (P_{1,max}, 0), & \text{if } g_1 P_{1,max} \geq max\left(g_2 P_{2,max}, 1/\beta^2\right), \\ (0, P_{2,max}), & \text{if } g_2 P_{2,max} \geq max\left(g_1 P_{1,max}, 1/\beta^2\right), \\ (P_{1,max}, P_{2,max}), & \text{otherwise.} \end{cases} \tag{19}$$

First we will execute our proposed algorithm for $\beta = 0.3$. According to the optimal solution, $(0, P_{2,max})$ is the optimal solution. According to Q-CoPA, the SBSs will choose the powers that maximizes the global Q-function. The learned global Q-function, $\mathbf{Q}(P_1, P_2)$, is plotted in Fig. 3 with maximum value at $P_1 = 0$ and $P_2 = P_{2,max}$, which is optimal.

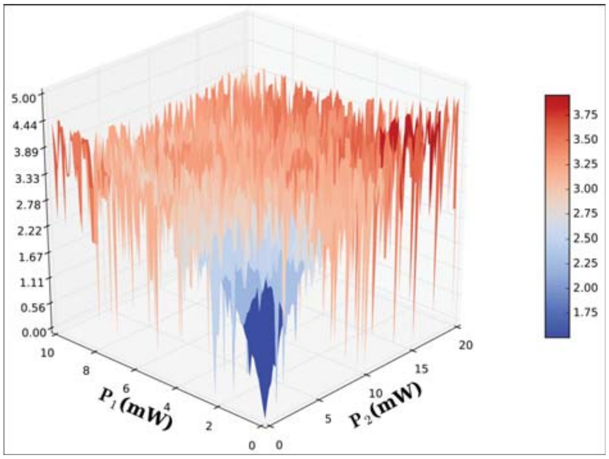

Fig. 3: Global action-value function.

In Fig. 4, the solution of the power allocation for different values of the portion of interference between two channels, $\beta \in [0, 1]$, is plotted. The greedy approach is defined to allocate full power to the transmitter with higher peak power, and zero to the other one. The simultaneous allocation is defined to use maximum power at both transmitters. According to Fig. 4, the Q-CoPA finds the optimal solution for all values of β.

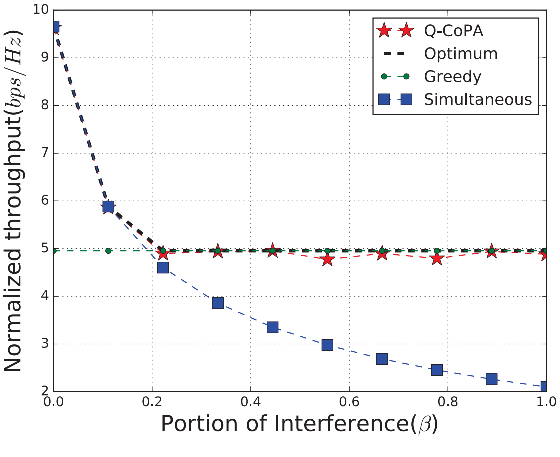

Fig. 4: Normalized throughput versus portion of interference (β).

VI. Conclusion

In this paper we used message-passing and variable elimination to coordinate the power allocation in order to maximize a common goal in an interference-limited network. The proposed solution is based on Q-*learning*, and does not need to know the model of the system, hence, it adapts itself if the architecture or number of SBSs in the network changes. Another advantage of this method is that the Q-functions are learned by just measuring the SINR value at each node (radio measurement), while the optimal solution depends on the channel estimation, for example values of g_1 and g_2 in the simulation in the section V.

The variable elimination algorithm is exact, so as long as the local Q-functions' action set covers all interfering SBSs, the proposed solution is optimal. Although, when each node of the CG gets densely connected, i.e., the size of action set of local Q-function grows, for the sake of computational complexity we need to approximate local Q-functions' action set with smaller sets, which results in suboptimal solution. Therefore, the proposed solution is suitable for indoor applications, or networks in which the number of interferes is low. As the future work the authors will explore the edge-based decomposition to support outdoor networks and highly dense CGs.

References

[1] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Commun. Mag.*, vol. 23, no. 1, pp. 72–79, February 2016.

[2] T. Park, J. Jang, O.-S. Shin, and K. B. Lee, "Transmit power allocation for a downlink two-user interference channel," *IEEE Commun. Lett.*, vol. 9, no. 1, pp. 13–15, Jan 2005.

[3] D. Park, "Optimal power allocation in two-user interference channel under individual power constraint," in *ICTC*, Oct 2016, pp. 530–532.

[4] M. Chiang, C. W. Tan, D. P. Palomar, D. O'neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[5] M. Chen, Y. Hua, X. Gu, S. Nie, and Z. Fan, "A self-organizing resource allocation strategy based on Q-learning approach in ultra-dense networks," in *IC-NIDC*, Sept 2016, pp. 155–160.

[6] S. Lin, J. Yu, W. Ni, and R. Liu, "Radio resource management for ultra-dense smallcell networks: A hybrid spectrum reuse approach," in *Proc. IEEE Veh. Technol. Conf.*, June 2017, pp. 1–7.

[7] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in HetNets considering QoS," *in Proc. IEEE ICC*, pp. 1–7, May 2018.

[8] R. Amiri and H. Mehrpouyan, "Self-organizing mm wave networks: A power allocation scheme based on machine learning," *in Proc. IEEE GSMM*, pp. 1–4, May 2018.

[9] A. Galindo-Serrano and L. Giupponi, "Self-organized femtocells: A fuzzy Q-learning approach," *Wirel. Netw.*, vol. 20, no. 3, pp. 441–455, Apr. 2014.

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

[11] C. Guestrin, M. G. Lagoudakis, and R. Parr, "Coordinated reinforcement learning," in *Proc.*, ser. ICML, 2002, pp. 227–234.

[12] S. Niknam and B. Natarajan, "On the regimes in millimeter wave networks: Noise-limited or interference-limited?" *in Proc. IEEE ICCW*, pp. 1–6, May 2018.

[13] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE J. Select. Areas Commun.*, vol. 24, no. 8, pp. 1426–1438, Aug 2006.

[14] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

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

[16] C. Boutilier, T. L. Dean, and S. Hanks, "Decision-theoretic planning: Structural assumptions and computational leverage," *CoRR*, vol. abs/1105.5460, 2011.

[17] J. R. Kok and N. Vlassis, "Sparse cooperative Q-learning," in *Proc.*, ser. ICML. ACM, 2004, pp. 61–.

[18] C. Guestrin, D. Koller, and R. Parr, "Multiagent planning with factored MDPs," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002, pp. 1523–1530.