This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

Joint BS-user Association, Power Allocation and User Side Interference Cancellation in Cell-free Heterogeneous Networks

An Liu, *Member IEEE*, and Vincent Lau, *Fellow IEEE*

***Abstract*—Future wireless networks are expected to be heterogeneous with dense deployment of pico (small cell) base stations (BSs) overlaid with traditional macro BSs. There are two performance bottlenecks in such heterogeneous networks (HetNet): the interference issue and cell-edge effect. We propose to combine cell-free BS-user association (BUA), power control and dynamic user side interference cancellation (IC) to mitigate these bottlenecks. By dynamically selecting the "best" serving BSs for each user, the cell-free BUA can exploit the *multi-BS diversity gain* to mitigate the interference and cell-edge effect. Furthermore, the user side IC eliminates the strong cross-tier interference. We formulate the joint optimization of cell-free BUA, power allocation and user side IC as a weighted sum-rate (WSR) maximization problem, and propose a WMMSE alternating optimization algorithm to solve it. Specifically, a generalized WMMSE method is proposed to solve the power optimization subproblem with non-differentiable WSR function. Furthermore, by exploiting the specific problem structure, low complexity search methods are designed to find the optimal solutions of the combinatorial cell-free BUA and user-side IC subproblems. The proposed algorithm is shown to converge to a stationary solution of the joint problem. Simulations verify the significant gain of the proposed solution over existing solutions.**

***Index Terms*—Cell-free heterogeneous networks, Multi-BS diversity, User side interference cancellation, WSR maximization**

I. Introduction

In HetNet, the dense deployment of pico BSs with small coverage areas can significantly improve the spectral efficiency per unit area. On the other hand, the overlaid macro BSs with larger coverage areas can be used to eliminate the coverage holes. Despite the significant benefits of HetNet, the interference in HetNet is more complicated than traditional cellular networks. First, to avoid underutilization of pico BSs, cell range expansion (CRE) has been introduced in HetNet to expand the coverage region of pico BSs. With CRE, the serving BS of a pico cell user may no longer be the BS with the strongest received signal strength (RSS) and thus the pico cell user may receive strong cross-tier interference from the nearest macro BS. To address this issue, various enhanced inter-cell interference coordination (eICIC) schemes have been proposed in LTE and LTE-A, such as the almost blank sub-frames (ABS) or reduced power sub-frames. Many radio resource optimization algorithms have also been proposed for HetNet with eICIC. For example, in [1], the authors proposed an algorithm for victim pico user partition and optimal synchronous ABS rate selection. In [2], almost blank resource block (ABRB), which can be viewed as a generalization of ABS, was proposed for more flexible interference control in HetNet. However, resource block (RB) blanking reduces the spectral efficiency in macro cells. Moreover, the cell-edge users in the HetNet is still the bottleneck of system performance.

Recently, the concepts of cell-free HetNet has been proposed in both industry and academia [3]. In cell-free HetNet, the identity, pilot, physical channel (broadcast/control/data) and so on for each user are decoupled from any BS to allow dynamic *cell-free BS-user association* (BUA). As a result, a user can access the network any time anywhere without any interruption across one gigantic area. However, many technical challenges remain unsolved in cell-free HetNet. For example, how to optimize the dynamic BUA to improve the system performance? And how to do resource allocation and interference control under dynamic BUA?

There are already some works on joint user association and interference/power control. The joint optimization of user association and RB blanking has been studied in [4]–[6]. In [7], a joint power control and load-aware user association algorithm is proposed to realize traffic offloading from the macro cell to pico cells and inter-cell interference mitigation in the HetNet. In [8], the authors considered joint optimization of the transmit powers, user scheduling, and user association in cellular networks to maximize the WSR based on an interference pricing approach. In addition, a joint user association and interference management algorithm has recently been proposed in [9] for HetNets with Massive MIMO.

On the other hand, as the processing capability of user equipment increases, interference mitigation at user side becomes possible and is currently under study by 3GPP [10]. Various practical issues associated with user side interference cancellation (IC) has been studied in [11]. In [12], the performance of using successive IC (SIC) is analyzed in HetNets with the BS locations distributed according to a Poisson point process. In [13], some advanced interference management schemes are discussed for 5G cellular networks, with an emphasis on the importance of user side IC.

In this paper, we combine the *cell-free BUA*, power control and *dynamic user side IC* to mitigate the interference and cell-edge effect in HetNet. Unlike most existing works where each

This work was partially supported by RGC N-HKUST605/13, and partially supported by Huawei.

An Liu and Vincent K. N. Lau are with the Department of ECE, The Hong Kong University of Science and Technology (email: eewendaol@ust.hk; eeknlau@ece.ust.hk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

user is only allowed to be associated with one BS, in the more flexible cell-free BUA, each user can be dynamically associated with multiple BSs according to the channel conditions and system loadings. The cell-free BUA provides a new degree of freedom to exploit the *multi-BS diversity* to mitigate the interference and cell-edge effect. Moreover, most existing works focus on static user side IC where each user always cancels a fixed number of interferers. In this paper, we consider a more flexible dynamic user side IC which is adaptive to the instantaneous channel gains and can further improve the performance over the static user side IC. The main contributions are elaborated below.

- **Joint cell-free BUA, power control and dynamic user side IC:** Due to the consideration of cell-free BUA and dynamic user side IC, the joint optimization problem is a mixed non-convex and combinatorial optimization problem, which cannot be solved by the existing optimization algorithms for HetNet.
- **Problem transformation based on virtual network:** To simplify the joint optimization problem, we introduce the notion of *virtual network* and show that the WSR maximization problems in the virtual network and actual HetNet are equivalent in Theorem 1. Theorem 1 reveals that the combinatorial optimization of cell-free BUA can be equivalently solved by the power allocation optimization in the virtual network.
- **Generalized WMMSE algorithm:** With user-side IC, the WSR objective function of the power allocation subproblem is non-differentiable and thus existing algorithms designed for differentiable WSR maximization, such as the WMMSE in [14], cannot be applied. To address this challenge, we propose a new generalized WMMSE algorithm to solve a stationary point of the non-differentiable WSR maximization problem.
- **Structural property of the optimal user-side IC:** In Theorem 4, we obtain the structural property of the optimal user-side IC, which is exploited to design a greedy algorithm that is able to find the optimal solution of the user side IC subproblem with only linear complexity.

The rest of the paper is organized as follows. The system model and the problem formulation are given in Section II and III, respectively. The proposed algorithm and the convergence analysis are given in Section IV. Simulations in Section V show that the proposed solution can achieve significant gain over existing solutions. The conclusion is given in Section VI.

II. System Model

A. HetNet Topology and Channel Model

Consider the downlink of a cell-free HetNet with N_0 macro BSs, $N-N_0$ pico BSs, and K users. For clarity, assume that the BSs and users are equipped with a single antenna. Without loss of generality, the macro BSs are indexed by $1,...,N_0$ and the pico BSs are indexed by $N_0+1,...,N$. The maximum transmit power of each macro BS and pico BS are denoted by P_M and P_P respectively. The macro BS has a larger maximum transmit power and thus a larger coverage area. Within the coverage area of one macro BS, there are usually multiple pico BSs with smaller coverage areas. The channel coefficient between BS n and user k is modeled as $h_{k,n}=L_{k,n}\overline{h}_{k,n}$, where $L_{k,n}$ is the large scale fading factor used to model the combined effect of path loss and shadow fading, and $\overline{h}_{k,n}$ is the small scale fading factor.

B. Cell-free BS-user Association

We consider a flexible cell-free BUA scheme in which the serving BSs of each user is dynamically selected from a set of candidate BSs according to the channel conditions and system loadings. Let $\mathcal{C}_k$ denote the set of candidate BSs for user k and $\mathcal{B}_k\subseteq\mathcal{C}_k$ denote the set of serving BSs for user k. Each BS in $\mathcal{B}_k$ sends an independent data stream to user k. For practical considerations, the BSs in $\mathcal{B}_k$ are not allowed to perform MIMO cooperation. As a result, each BS can only transmit to a single user at a given time slot, i.e., $\mathcal{B}_k\cap\mathcal{B}_l=\emptyset,\forall k\neq l$. For convenience, let $\boldsymbol{\mathcal{B}}=\{\mathcal{B}_k,k=1,...,K\}$ denote the BUA control variable and

$$\Omega_{\mathcal{B}}=\{\boldsymbol{\mathcal{B}}:\ \mathcal{B}_k\subseteq\mathcal{C}_k,\forall k;\mathcal{B}_k\cap\mathcal{B}_l=\emptyset,\forall k\neq l\}$$

denote the set of feasible BUA variables.

Define the biased large scale fading factor between BS n and user k as

$$\tilde{L}_{k,n}=\begin{cases}L_{k,n}, & n\in\{1,...,N_0\}\\ \sqrt{\beta}L_{k,n}, & n\in\{N_0+1,...,N\}\end{cases},$$

where $\beta>1$ is the *candidate BS selection bias* for pico BSs. Then the candidate BS set $\mathcal{C}_k$ for user k is chosen as the N_c BSs which have the largest biased large scale fading factors with user k. The purpose of introducing the candidate BS selection bias is to balance the loading between macro and pico BSs [15]. In practice, N_c is usually chosen to be a small number (e.g., 2 or 3) because typically, a user can only receive strong signals from 2 to 3 BSs. Clearly, the cell-free BUA helps to mitigate the cell-edge effect.

Under the cell-free BUA, the received signal of user k is

$$y_k=\sum_{n\in\mathcal{B}_k}h_{k,n}x_{k,n}+\sum_{l\neq k}\sum_{n\in\mathcal{B}_l}h_{k,n}x_{l,n}+z_k,$$

where $x_{k,n}$ with $\mathbb{E}\left|x_{k,n}\right|^2=p_n$ is the data symbol from BS n to user k, p_n is the transmit power at BS n, and $z_k\in\mathcal{CN}\left(0,1\right)$ is the additive white Gaussian noise (AWGN). For a user receiving multiple data streams from the BSs in $\mathcal{B}_k$, successive decoding [16] is employed to decode its desired data streams. Specifically, the decoding order at user k is represented by a permutation mapping π_k from $\{1,...,|\mathcal{B}_k|\}$ to the set $\mathcal{B}_k$. Given decoding order π_k, the data stream from BS $\pi_k\left(i\right)$ to user k is the i-th one to be decoded, and thus it is not interfered by BS $\pi_k\left(1\right),\pi_k\left(2\right),...,\pi_k\left(i-1\right)$.

C. User Side Interference Cancellation

In the cell-free BUA scheme, the candidate BS selection bias β helps to expand the effective coverage area of pico BSs. However, this will also cause strong cross-tier interference from the macro BS to pico users. In this paper, power control will be used at the macro BSs to mitigate the strong cross-tier

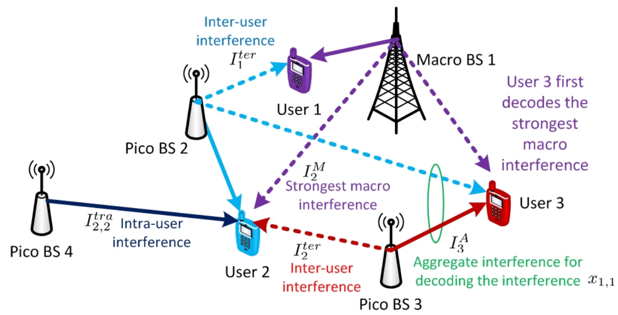


Figure 1: Illustration of interference and achievable rate under the cell-free BUA and user side IC. The serving BS sets of the three users are $\mathcal{B}_1 = \{1\}$, $\mathcal{B}_2 = \{2, 4\}$ and $\mathcal{B}_3 = \{3\}$.

interference. Furthermore, we consider *user side interference cancellation* (IC) scheme where each user is allowed to first decode and cancel the interference from the strongest macro BS (i.e., the macro BS which has the largest channel gain with the user), and then decode its desired data streams. Such user side IC is motivated by the information theoretical study of interference channel, where it has been shown that decoding and completely cancelling the interference is optimal in the strong interference channel [17]. Specifically, let $b_k = \underset{n\in\{1,...,N_0\}}{\text{argmax}} \ g_{k,n}$ denote the strongest macro BS of user k, where $g_{k,n} = |h_{k,n}|^2$ denotes the channel gain between BS n and user k. Let $\theta_k \in \{0, 1\}$ denote the user side IC control variable of user k, where $\theta_k = 0$ means that user k will not decode the interference from the strongest macro BS b_k, and $\theta_k = 1$ means that user k will decode and cancel the interference from the strongest macro BS b_k. Let $\boldsymbol{\theta}_n = [\theta_k, \forall k \in \{l : \ b_l = n\}]^T$ denote the set of user side IC control variables associated with macro BS n, for $n \in \{1, ..., N_0\}$.

When user k cancels the interference from its strongest macro BS b_k, we may have to decrease the data rate of the user served by macro BS b_k (say user l) so that user k can correctly decode the interference from macro BS b_k (i.e., the data stream target for user l), as will be illustrated by the example in Section II-D. Hence, although user side IC at user k helps to increase the data rate of user k, it does not necessarily improve the overall system performance. As a result, we must dynamically control the user side IC according to the instantaneous channel gains in order to improve the overall system performance. In this paper, we only consider cancelling the strongest interferer because such design achieves the best tradeoff between performance and complexity. As shown by both the analytical results and simulation results in [12], allowing more interferers being cancelled only provides marginal gain with much higher complexity and signaling overhead.

D. Interference and Achievable Rate

With user side IC, user k not only needs to decode its desired data streams $x_{k,n}, \forall n \in \mathcal{B}_k$ but also the interference from the strongest macro BS when $\theta_k = 1$. For the decoding of a desired data stream $x_{k,n}, n \in \mathcal{B}_k$ at user k, there are three types of interference as elaborated below.

Intra-user interference $I_{k,n}^{tra}$ associated with the decoding of desired data stream $x_{k,n}$ at user k: This is the interference from the other serving BSs $\mathcal{B}_k \backslash \{n\}$ of user k and it depends on the decoding order π_k. Specifically, the intra-user interference is given by $I_{k,n}^{tra} = \sum_{i=\pi_k^{-1}(n)+1}^{|\mathcal{B}_k|} g_{k,\pi_k(i)} p_{\pi_k(i)}$, where π_k^{-1} denotes the inverse mapping of π_k such that $\pi_k^{-1}(\pi_k(i)) = i$. For example, for user 2 in Fig. 1, it receives two desired data streams $x_{2,2}$ and $x_{2,4}$ from pico BS 2 and pico BS 4 respectively. Suppose the decoding order is given by $\pi_2(1) = 4$ and $\pi_2(2) = 2$, i.e., user 2 first decodes $x_{2,4}$ and then decodes $x_{2,2}$. Then the intra-user interference associated with the decoding of $x_{2,2}$ is $I_{2,2}^{tra} = 0$. On the other hand, if the decoding order is reversed, then the intra-user interference associated with the decoding of $x_{2,2}$ is $I_{2,2}^{tra} = g_{2,4} p_4$.

Strongest macro interference I_k^M associated with the decoding of desired data streams $x_{k,n}, \forall n \in \mathcal{B}_k$ at user k: This is the interference from the strongest macro BS b_k of user k and it depends on the user side IC θ_k. Note that the strongest macro interference is the same for the decoding of all desired data stream $x_{k,n}, \forall n \in \mathcal{B}_k$ and thus it only has a single subscript k. When $\theta_k = 1$, the strongest macro interference is decoded and cancelled at user k and thus $I_k^M = 0$ after user side IC. When $b_k \in \mathcal{B}_k$, the strongest macro interference I_k^M is also zero because BS b_k is a serving BS of user k and thus the interference from BS $b_k \in \mathcal{B}_k$ is counted as part of the intra-user interference which depends on the decoding order π_k. On the other hand, when $\theta_k = 0$ and $b_k \notin \mathcal{B}_k$, the strongest macro interference is given by $I_k^M = g_{k,b_k} p_{b_k}$. In summary, we have $I_k^M = \mathbb{I}(b_k \notin \mathcal{B}_k)(1 - \theta_k) g_{k,b_k} p_{b_k}$, where $\mathbb{I}(\cdot)$ is the indication function. For example, for user 2 in Fig. 1, if $\theta_2 = 0$, then user 2 does not decode and cancel the strongest macro interference and thus the strongest macro interference at user 2 is given by $I_2^M = g_{2,1} p_1$. On the other hand, if $\theta_3 = 1$, then user 3 will decode and cancel the strongest macro interference and thus the strongest macro interference at user 3 is $I_3^M = 0$.

Inter-user interference I_k^{ter} associated with the decoding of desired data streams $x_{k,n}, \forall n \in \mathcal{B}_k$ at user k: This is the interference from the BSs other than the serving BSs in $\mathcal{B}_k$ or the strongest macro BS b_k (i.e., $\{1, ..., N\} \backslash \{\mathcal{B}_k \cup b_k\}$), and it does not depend on the decoding order or user side IC. The inter-user interference is also the same for the decoding of all desired data stream $x_{k,n}, \forall n \in \mathcal{B}_k$. For example, for user 2 in Fig. 1, the inter-user interference is from pico BS 3 and is given by $I_2^{ter} = g_{2,3} p_3$.

For the decoding of strongest macro interference at user k, the *aggregate interference* I_k^A is contributed from all BSs other than macro BS b_k, and it also does not depend on the decoding order or user side IC. Specifically, $I_k^A = \sum_{n' \in \{1,...,N\} \backslash \{b_k\}} g_{l,n'} p_{n'}$. For example, for user 3 in Fig. 1, the aggregate interference for decoding the strongest macro interference $x_{1,1}$ is given by $I_3^A = g_{3,2} p_2 + g_{3,3} p_3 + g_{3,4} p_4$.

With the above interference expressions, we can obtain the achievable rate for any data stream $x_{k,n}, n \in \mathcal{B}_k$ at user k. There are two types of data streams, namely, *pico data stream* sent from a pico BS and *macro data stream* send from a macro BS. Due to user side IC, the calculations of achievable rate for these two types of data streams are different.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

Achievable rate for the pico data stream $x_{k,n}$ from pico BS n to user k: For the pico data stream $x_{k,n}$, the data rate only depends on the SINR of $x_{k,n}$ seen at the desired user k. Hence, the data rate of $x_{k,n}$ decoded at user k is given by

$$\overline{R}_{k,n,k}=\log_{2}\left(1+\frac{g_{k,n}p_{n}}{I_{k,n}^{tra}+I_{k}^{M}+I_{k}^{ter}+1}\right),\qquad(1)$$

where $g_{k,n}p_n$ is the signal power, and $I_{k,n}^{tra}, I_k^M, I_k^{ter}$ are the three types of interference associated with the decoding of desired data stream $x_{k,n}$ at user k as elaborated above. Note that the subscript $_{k,n,l}$ indicates that this is the achievable rate associated with the decoding of data stream $x_{k,n}$ at user l (in general, k,l can be different due to the user side IC).

Achievable rate for the macro data stream $x_{k,n}$ from macro BS n to user k: The major difference from the pico data stream is that, the achievable rate of a macro data stream $x_{k,n}$ not only depends on the SINR of $x_{k,n}$ seen at the desired user k, but also the SINR of $x_{k,n}$ seen at the users in $\overline{\mathcal{U}}_n=\{l:\ b_l=n,\theta_l=1\}$ who need to decode $x_{k,n}$ for interference cancellation. The maximum allowable rate for the desired user k to successfully decode $x_{k,n}$ is still given by (1). On the other hand, the maximum allowable rate for a user $l\in\overline{\mathcal{U}}_n$ to successfully decode the interference $x_{k,n}$ is given by

$$\overline{R}_{k,n,l}=\log\left(1+\frac{g_{l,n}p_{n}}{1+I_{l}^{A}}\right).\qquad(2)$$

where $g_{l,n}p_n$ is the power of the strongest macro interference seen at user l, and $I_l^A=\sum_{n'\in\{1,...,N\}\setminus\{n\}}g_{l,n'}p_{n'}$ is the aggregate interference for the decoding of strongest macro interference at user l as elaborated above. Finally, the achievable rate of a macro data stream $x_{k,n}$ is given by

$$R_{k,n}=\min_{l\in k\cup\overline{\mathcal{U}}_{n}}\overline{R}_{k,n,l},\qquad(3)$$

where $\overline{R}_{k,n,l}$ is given in (1) when $l=k$ and (2) when $l\in\overline{\mathcal{U}}_n$.

Note that for a pico BS $n\in\{N_0+1,...,N\}$, we have $\overline{\mathcal{U}}_n=\emptyset$ and thus $R_{k,n}=\overline{R}_{k,n,k}$. Hence, we can use $R_{k,n}$ as the unified expression for the achievable rate of both pico and macro data streams. Finally, the sum rate of all data streams of user k is $R_k=\sum_{n\in\mathcal{B}_k}R_{k,n}$.

III. Problem Formulation and Transformation

A. WSR Maximization Formulation

In this paper, we consider the following WSR maximization problem under per-BS power constraints

$$\mathcal{P}_{0}:\max_{\boldsymbol{\mathcal{B}}\in\Omega_{\mathcal{B}},\{p_{n}\},\boldsymbol{\theta},\boldsymbol{\pi}}\sum_{k=1}^{K}\mu_{k}R_{k},\ \text{s.t.}\ p_{n}\leq P_{n},\forall n\in\cup_{k}\mathcal{B}_{k},\qquad(4)$$

where $\{p_n\}$ is the transmit powers for the active BSs in $\cup_k\mathcal{B}_k$, $\boldsymbol{\theta}=\{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{N_0}\}$ is the set of all user side IC control variables and $\boldsymbol{\pi}=\{\pi_1,...,\pi_K\}$ is the set of decoding orders for all users. In the per BS power constraint $p_n\leq P_n$, we have $P_n=P_M$ for $n\in\{1,...,N_0\}$ (macro BS) and $P_n=P_P$ for $n\in\{N_0+1,...,N\}$ (pico BS). The WSR has been widely applied to wireless radio resource control. Many other popular network utility maximization problems, such as proportional fairness, can be decomposed into WSR maximization subproblems.

Problem $\mathcal{P}_0$ is a mixed non-convex and combinatorial problem whose solution space is exponentially large w.r.t. the number of BSs N and the number of users K. To address this challenge, we first transform Problem (4) into a more tractable problem which can be interpreted as a WSR maximization problem in a virtual HetNet. Then we combine the alternating optimization and WMMSE [14] approaches to solve the transformed problem in Section IV.

B. Problem Transformation

We first formulate a WSR maximization problem for a virtual HetNet. Then we show that the WSR maximization problems in the virtual and actual HetNets are equivalent. Note that from mathematical point of view, we can directly pose the transformed problem without explaining the physical meaning/intuition behind it. However, introducing the concept of virtual HetNet can help the readers to better understand the transformed problem and the associated solution structure discussed in the later sections.

In a virtual HetNet, each user always receives N_c data streams from all candidate BSs in $\mathcal{C}_k$. Unlike the actual HetNet considered in Section II where each BS can only transmit a single data stream, each BS in the virtual HetNet may transmit multiple data streams to multiple users since the candidate BSs of different users may overlap. Let $p_{k,n}$ denote the transmit power for the data stream from BS n to user k and let $\mathcal{S}_n=\{k:\ n\in\mathcal{C}_k\}$ denote the set of users served by BS n, in the virtual HetNet. *We define* the rate of the data stream $x_{k,n}$ from BS n to user $k\in\mathcal{S}_n$ in the virtual HetNet as

$$R_{k,n}^{'}=\min_{l\in k\cup\overline{\mathcal{U}}_{n}}\overline{R}_{k,n,l}^{'},\qquad(5)$$

where $\overline{R}_{k,n,l}^{'}$ with $l=k$ and $l\in\overline{\mathcal{U}}_n$ are respectively given by

$$\overline{R}_{k,n,k}^{'}=\log\left(1+\frac{g_{k,n}p_{k,n}}{I_{k,n,k}^{M'}+I_{k,n,k}^{tra'}+I_{k,n,k}^{ter'}+1}\right),\qquad(6)$$

$$\overline{R}_{k,n,l}^{'}=\log\left(1+\frac{g_{l,n}p_{k,n}}{1+I_{k,n,l}^{A'}}\right).\qquad(7)$$

where $I_{k,n,k}^{M'}=(1-\mathbb{I}(b_k\neq n)\,\theta_k)\sum_{l\in\mathcal{S}_{b_k}\setminus\{k\}}g_{k,b_k}p_{l,b_k}$, $I_{k,n,k}^{tra'}=\sum_{i=\pi_k^{-1}(n)+1}^{|\mathcal{C}_k|}g_{k,\pi_k(i)}p_{k,\pi_k(i)}$, $I_{k,n,k}^{ter'}=\sum_{l\neq k}\sum_{n'\in\mathcal{C}_l\setminus\{b_k\}}g_{k,n'}p_{l,n'}$, and $I_{k,n,l}^{A'}=\sum_{l'}\sum_{n'\in\mathcal{C}_{l'}}g_{l,n'}p_{l',n'}-g_{l,n}p_{k,n}$. Then, the sum rate of all data streams of user k in the virtual HetNet *is defined as* $R_k^{'}=\sum_{n\in\mathcal{C}_k}R_{k,n}^{'}$.

Note that the virtual HetNet does not correspond to an implementable physical HetNet. Hence, we have freedom to define the virtual data rate $R_k^{'}$ in the virtual HetNet, as long as the WSR maximization problem in the virtual HetNet is mathematically equivalent to that in the actual HetNet. Specifically, the virtual data rate $R_k^{'}$ in the virtual HetNet is defined such that the achievable rate in the actual HetNet with control variables $\boldsymbol{\mathcal{B}} \in \Omega_{\mathcal{B}}, \{p_n\}, \boldsymbol{\theta}, \boldsymbol{\pi}$ is equal to that of the virtual HetNet with control variables $\boldsymbol{\theta}, \boldsymbol{\pi}$ and

$$p_{k,n} = \begin{cases} p_n, & \forall n \in \mathcal{B}_k \\ 0, & \forall n \in \mathcal{C}_k \backslash \mathcal{B}_k \end{cases}. \quad (8)$$

For example, consider the system in Fig. 1, and assume the candidate BS sets are $\mathcal{C}_1 = \{1,2\}, \mathcal{C}_2 = \{2,4\}, \mathcal{C}_3 = \{1,3\}$. The achievable rate in the actual HetNet with control variables $\mathcal{B}_1 = \{1\}$, $\mathcal{B}_2 = \{2,4\}$, $\mathcal{B}_3 = \{3\}$ is equal to that of the virtual HetNet with power allocation $p_{1,1} = p_1$, $p_{1,2} = 0$, $p_{2,2} = p_2$, $p_{2,4} = p_4$, $p_{3,1} = 0$ and $p_{3,3} = p_3$.

Finally, the WSR maximization problem in the virtual HetNet can be formulated as

$$\mathcal{P}: \max_{\mathbf{p},\boldsymbol{\theta},\boldsymbol{\pi}} \sum_{k=1}^{K} \mu_k R_k^{'}, \text{ s.t. } \sum_{k \in \mathcal{S}_n} p_{k,n} \leq P_n, \forall n \in \cup_k \mathcal{C}_k, \quad (9)$$

where $\mathbf{p} = \left[\mathbf{p}_{1,\mathcal{C}_1}^T, \mathbf{p}_{2,\mathcal{C}_k}^T, ..., \mathbf{p}_{K,\mathcal{C}_K}^T\right]^T$ with $\mathbf{p}_{k,\mathcal{C}_k} = [p_{k,n}, \forall n \in \mathcal{C}_k]^T$ is the aggregate power allocation vector for all users. Clearly, any feasible solution $\boldsymbol{\mathcal{B}}, \{p_n\}, \boldsymbol{\theta}, \boldsymbol{\pi}$ of Problem $\mathcal{P}_0$ can be mapped to a feasible solution of Problem $\mathcal{P}$ using (8) such that $\sum_{k=1}^K \mu_k R_k^{'} = \sum_{k=1}^K \mu_k R_k$. Hence, the optimal objective value of Problem $\mathcal{P}$ provides an upper bound for that of Problem $\mathcal{P}_0$.

The following theorem formally establishes the equivalence between $\mathcal{P}_0$ and $\mathcal{P}$. Please refer to Appendix A for the proof.

Theorem 1 (Equivalence between Problem $\mathcal{P}_0$ and $\mathcal{P}$)**.** *Let* $(\mathbf{p}^\star, \boldsymbol{\theta}^\star, \boldsymbol{\pi}^\star)$ *denote the optimal solution of Problem (9) and let* $\mathcal{B}_k^\star = \left\{n : n \in \mathcal{C}_k, p_{k,n}^\star > 0\right\}$. *Then we have* $\sum_{k \in \mathcal{S}_n} \mathbb{I}\left(p_{k,n}^\star > 0\right) \leq 1, \forall n$ *(i.e., at the optimal solution of (9), each BS in the virtual HetNet has at most one outgoing data stream with non-zero transmit power) and thus* $\mathcal{B}_k^\star \cap \mathcal{B}_l^\star = \emptyset, \forall k \neq l$. *Moreover,* $\boldsymbol{\mathcal{B}}^\star = \{\mathcal{B}_k^\star, k = 1, ..., K\}, \boldsymbol{\theta}^\star, \boldsymbol{\pi}^\star, \left\{p_n^\star = \sum_{k \in \mathcal{S}_n} p_{k,n}^\star\right\}$ *is the optimal solution of Problem (4), and the optimal objective of Problem (4) is equal to that of Problem (9).*

Hence, we shall focus on solving Problem $\mathcal{P}$.

C. Stationary Solution of Problem $\mathcal{P}$

Problem $\mathcal{P}$ is still a mixed non-convex and combinatorial optimization problem. Solving the optimal solution of $\mathcal{P}$ involves exhaustive search over exponentially large solution space which is intractable. In this paper, we aims at designing a low complexity algorithm to find a *stationary solution* of Problem $\mathcal{P}$ defined below.

Definition 1 (Stationary solution of $\mathcal{P}$)**.** A solution $(\mathbf{p}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)$ is called a stationary solution of Problem $\mathcal{P}$ if

1) $\mathbf{p}^*$ is a stationary point[1] of Problem $\mathcal{P}$ with fixed $\boldsymbol{\theta}^*, \boldsymbol{\pi}^*$.
2) $\boldsymbol{\theta}_n^*$ is the optimal solution of Problem $\mathcal{P}$ with fixed $\mathbf{p}^*, \boldsymbol{\pi}^*$ and $\boldsymbol{\theta}_{-n}^* \triangleq \left\{\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_{n-1}^*, \boldsymbol{\theta}_{n+1}^*, ..., \boldsymbol{\theta}_{N_0}^*\right\}$.
3) $\boldsymbol{\pi}^*$ is the optimal solution of Problem $\mathcal{P}$ with fixed $\mathbf{p}^*, \boldsymbol{\theta}^*$.

The global optimal solution of $\mathcal{P}$ must be a stationary solution. However, the reverse is not necessarily true. For the special case of non-convex problem, a stationary solution reduces to a conventional stationary point (i.e., a point satisfying the KKT conditions of the non-convex problem). In the simulations, it is observed that the proposed algorithm always converges to a stationary solution with good performance.

The problem of finding a stationary solution of $\mathcal{P}$ is still highly non-trivial. First, for fixed $\boldsymbol{\theta}, \boldsymbol{\pi}$, the optimization of $\mathbf{p}$ belongs to a non-convex problem with non-differentiable functions. Second, for fixed $\mathbf{p}, \boldsymbol{\pi}$ ($\mathbf{p}, \boldsymbol{\theta}$), the optimization of $\boldsymbol{\theta}$ ($\boldsymbol{\pi}$) is combinatorial and there is no generic low complexity algorithm which is guaranteed to find the optimal solution.

IV. WMMSE Alternating Optimization for $\mathcal{P}$

In this section, we propose a novel WMMSE alternating optimization (WMMSE-AO) algorithm to solve a stationary solution of $\mathcal{P}$. In WMMSE-AO, the power allocation $\mathbf{p}$, user side IC $\boldsymbol{\theta}$ and decoding order $\boldsymbol{\pi}$ are optimized in an alternating way. Specifically, for fixed $\boldsymbol{\theta}, \boldsymbol{\pi}$, the power optimization subproblem is

$$\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi}): \max_{\mathbf{p}} \sum_{k=1}^{K} \mu_k R_k^{'}, \text{ s.t. } \sum_{k \in \mathcal{S}_n} p_{k,n} \leq P_n, \forall n \in \cup_k \mathcal{C}_k.$$

In Section IV-A, we will propose a *generalized WMMSE algorithm* to find a stationary point of $\mathcal{P}_p$. For fixed $\mathbf{p}, \boldsymbol{\theta}_{-n}, \boldsymbol{\pi}$, the user side IC optimization subproblem is

$$\mathcal{P}_{\theta_n}(\mathbf{p}, \boldsymbol{\theta}_{-n}, \boldsymbol{\pi}): \max_{\boldsymbol{\theta}_n} \sum_{k=1}^{K} \mu_k R_k^{'}.$$

In Section IV-B, we will propose a *greedy IC algorithm* to find the optimal solution of $\mathcal{P}_{\theta_n}$. Finally, for fixed $\mathbf{p}, \boldsymbol{\theta}$, the decoding order optimization subproblem is

$$\mathcal{P}_\pi(\mathbf{p}, \boldsymbol{\theta}): \max_{\boldsymbol{\pi}} \sum_{k=1}^{K} \mu_k R_k^{'}.$$

In Section IV-C, we will propose a *per-user decoding order determination (PU-DoD)* algorithm to find the optimal solution of $\mathcal{P}_\pi$. The overall WMMSE-AO algorithm is summarized in Algorithm 1 and the detailed algorithms to solve the subproblems will be elaborated in the subsequent subsections.

A. Generalized WMMSE for Solving $\mathcal{P}_p$

Note that the original WMMSE algorithm and the associated convergence proof in [14] for linear MMSE receiver cannot be directly applied to Problem $\mathcal{P}_p$ with non-linear receiver (user-side IC). This is because with user-side IC, the objective function of Problem $\mathcal{P}_p$ (WSR) is non-differentiable and the original WMMSE approach for solving a stationary point of differentiable WSR maximization cannot be directly applied. This challenge will be addressed in the following.

Introduce two sets of auxiliary variables $\mathbf{q} = \{q_{k,n}, \forall k, \forall n \in \mathcal{C}_k\}$ and $\mathbf{u} =$

[1] A stationary point for a non-convex problem with possibly non-differentiable objective function is a point which satisfies the generalized KKT conditions defined in [18]. When the objective function is differentiable, the generalized KKT conditions reduces to the conventional KKT conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

Algorithm 1 WMMSE Alternating Optimization for $\mathcal{P}$

1: **Initialization:** Choose a feasible initial solution $\mathbf{p}^0, \boldsymbol{\theta}^0, \boldsymbol{\pi}^0$.
2: **Let** $i = 0$.
3: **Step 1: (Power optimization):** Call the generalized WMMSE algorithm with input $\boldsymbol{\theta}^i, \boldsymbol{\pi}^i, \mathbf{p}^i$ to obtain a stationary point $\mathbf{p}^{i+1}$ of $\mathcal{P}_p\left(\boldsymbol{\theta}^i, \boldsymbol{\pi}^i\right)$.
4: **Step 2 (User side IC optimization):** For $n = 1$ to N_0, call the greedy IC algorithm with input $\mathbf{p}^{i+1}, \boldsymbol{\pi}^i, \boldsymbol{\theta}_n^i, \boldsymbol{\theta}_{-n}^{i+1} \triangleq \left\{\boldsymbol{\theta}_1^{i+1}, ..., \boldsymbol{\theta}_{n-1}^{i+1}, \boldsymbol{\theta}_{n+1}^{i}, ..., \boldsymbol{\theta}_{N_0}^{i}\right\}$ to obtain the optimal solution $\boldsymbol{\theta}_n^{i+1}$ of $\mathcal{P}_{\theta_n}\left(\mathbf{p}^{i+1}, \boldsymbol{\theta}_{-n}^{i+1}, \boldsymbol{\pi}^i\right)$.
5: **Step 3 (Decoding order optimization):** Call the PU-DoD algorithm with input $\mathbf{p}^{i+1}, \boldsymbol{\theta}^{i+1}, \boldsymbol{\pi}^i$ to obtain the optimal solution $\boldsymbol{\pi}^{i+1}$ of $\mathcal{P}_\pi\left(\mathbf{p}^{i+1}, \boldsymbol{\theta}^{i+1}\right)$.
6: **Let** $i = i + 1$ **and return to** Step 1 until convergence.

$\left\{u_{k,n,l}, \forall k, \forall n \in \mathcal{C}_k, \forall l \in k \cup \overline{\mathcal{U}}_n\right\}$, where $q_{k,n}^2 = p_{k,n}$ is the transmit power of the data stream from BS n to user k, and $u_{k,n,l}$ is the receive coefficient for the decoding of data stream $x_{k,n}$ at user l. Then the interference-plus-noise power associated with the rate term $\overline{R}_{k,n,k}^{'}$ is

$$IN_{k,n,k} = \Bigg(1 - \sum_{l \in \mathcal{S}_{b_k} \backslash k} \mathbb{I}\left(b_k \notin n\right) \theta_k g_{k,b_k} q_{l,b_k}^2 + \sum_{i=\pi_k^{-1}(n)+1}^{|\mathcal{C}_k|} g_{k,\pi_k(i)} q_{k,\pi_k(i)}^2 + \sum_{l \neq k} \sum_{n' \in \mathcal{C}_l} g_{k,n'} q_{l,n'}^2\Bigg),$$

and the corresponding MSE associated with $\overline{R}_{k,n,k}^{'}$ is

$$e_{k,n,k} = \left(1 - u_{k,n,k}\sqrt{g_{k,n}} q_{k,n}\right)^2 + u_{k,n,k}^2 IN_{k,n,k}.$$

Similarly, the interference-plus-noise power associated with the rate term $\overline{R}_{k,n,l}^{'}, l \in \overline{\mathcal{U}}_n$ is

$$IN_{k,n,l} = 1 + \sum_{l'} \sum_{n' \in \mathcal{B}_{l'}} g_{l,n'} q_{l',n'}^2 - g_{l,n} q_{k,n}^2,$$

and the MSE associated with $\overline{R}_{k,n,l}^{'}, l \in \overline{\mathcal{U}}_n$ is

$$e_{k,n,l} = \left(1 - u_{k,n,l}\sqrt{g_{l,n}} q_{k,n}\right)^2 + u_{k,n,l}^2 IN_{k,n,l}.$$

Define the weighted MMSE associated with $R_{k,n}^{'}$ in (5) as

$$\gamma_{k,n} = \min_{l \in k \cup \overline{\mathcal{U}}_n} w_{k,n,l} e_{k,n,l} - \log w_{k,n,l}, \tag{10}$$

where the auxiliary variable $w_{k,n,l}$ is the MSE weight associated with the rate term $\overline{R}_{k,n,l}^{'}$. Let $\gamma_k = \sum_{n \in \mathcal{C}_k} \gamma_{k,n}$. Then the equivalent weighted MMSE minimization problem for $\mathcal{P}_p\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$ can be formulated as:

$$\mathcal{P}_p^{'}\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right): \min_{\mathbf{q},\mathbf{u},\mathbf{w}} \sum_{k=1}^K \mu_k \gamma_k, \text{ s.t. } \sum_{k \in \mathcal{S}_n} q_{k,n}^2 \leq P_n, \forall n \in \cup_k \mathcal{C}_k, \tag{11}$$

where $\mathbf{w} = \left\{w_{k,n,l}, \forall k, \forall n \in \mathcal{C}_k, \forall l \in k \cup \overline{\mathcal{U}}_n\right\}$ is the set of all MSE weights.

Theorem 2 (Equivalence between Problem $\mathcal{P}_p$ and $\mathcal{P}_p^{'}$)**.** *Let* $\left(\mathbf{q}^\star, \mathbf{u}^\star, \mathbf{w}^\star\right)$ *denote the optimal solution of Problem* $\mathcal{P}_p^{'}\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$*. Then* $\mathbf{p}^\star \triangleq \mathbf{q}^\star \circ \mathbf{q}^\star$ *is the optimal solution of Problem* $\mathcal{P}_p\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$*, where the notation* $\circ$ *denotes the Hadamard product, and the optimal objective value of Problem* $\mathcal{P}_p\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$ *is equal to that of Problem* $\mathcal{P}_p^{'}\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$*.*

Please refer to Appendix B for the detailed proof.

Hence we only need to solve $\mathcal{P}_p^{'}$, which is convex in each of the optimization variables $\mathbf{q}, \mathbf{u}, \mathbf{w}$. We can use the block coordinate decent method to solve $\mathcal{P}_p^{'}$. First, for fixed $\mathbf{q}, \mathbf{u}$, the optimal $\mathbf{w}$ is given by $w_{k,n,l} = e_{k,n,l}^{-1}, \forall k, n, l$. Second, for fixed $\mathbf{q}, \mathbf{w}$, the optimal $\mathbf{u}$ is given by the MMSE receiver: $u_{k,n,l} = IN_{k,n,l}^{-1}\sqrt{g_{l,n}} q_{k,n}, \forall k, n, l$. Finally, for fixed $\mathbf{u}, \mathbf{w}$, it can be verified that Problem $\mathcal{P}_p^{'}$ is a convex problem:

$$\min_{\mathbf{q}} \sum_{k=1}^K \mu_k \gamma_k, \text{ s.t. } \sum_{k \in \mathcal{S}_n} q_{k,n}^2 \leq P_n, \forall n \in \cup_k \mathcal{C}_k, \tag{12}$$

which can be solved using existing convex optimization software. The overall generalized WMMSE algorithm is summarized in Algorithm 2 and the convergence of the algorithm is established in the following theorem.

Algorithm 2 Generalized WMMSE for $\mathcal{P}_p^{'}\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$

1: **Input:** $\boldsymbol{\theta}, \boldsymbol{\pi}$ and $\mathbf{p}$.
2: **Initialization: Let** $l = 0$ and $q_{k,n} = \sqrt{p_{k,n}}, \forall k, n$.
3: **Step 1 (Update u): Let** $u_{k,n,l} = IN_{k,n,l}^{-1}\sqrt{g_{l,n}} q_{k,n}, \forall k, n, l$.
4: **Step 2 (Update w): Let**

$$w_{k,n,l} = \left(1 - u_{k,n,l}\sqrt{g_{l,n}} q_{k,n}\right)^{-1}, \forall k, n, l.$$

5: **Step 3 (Update q):** Solve the convex problem (**12**) with fixed $\mathbf{u}, \mathbf{w}$ to obtain the optimal solution $\mathbf{q}^{(l+1)}$. Let $\mathbf{q} = \mathbf{q}^{(l+1)}$.
6: **Let** $l = l + 1$ **and return to** Step 1 until convergence.
7: **Output** $\mathbf{p} = \mathbf{q} \circ \mathbf{q}$.

Theorem 3 (Convergence of Generalized WMMSE)**.** *Algorithm 2 converges to a stationary point* $\left(\mathbf{q}^*, \mathbf{u}^*, \mathbf{w}^*\right)$ *of Problem* $\mathcal{P}_p^{'}\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$*, and the corresponding* $\mathbf{p}^* \triangleq \mathbf{q}^* \circ \mathbf{q}^*$ *is a stationary point of Problem* $\mathcal{P}_p\left(\boldsymbol{\theta}, \boldsymbol{\pi}\right)$*. Moreover,* $\mathbf{p}^*$ *satisfies* $\sum_{k \in \mathcal{S}_n} \mathbb{I}\left(p_{k,n}^* > 0\right) \leq 1, \forall n$*.*

Please refer to Appendix C for the proof.

B. Greedy IC for Solving $\mathcal{P}_{\theta_n}$

Let $\mathcal{K}_n = \{l : b_l = n\}, \forall n \in \{1, ..., N_0\}$ denote the set of users whose strongest macro BS is BS n. The optimization of $\boldsymbol{\theta}_n$ is a combinatorial problem and finding the optimal solution using exhaustive search has exponential complexity $O\left(2^{|\mathcal{K}_n|}\right)$, which is infeasible in practice. Since in Step 2 of Algorithm 1, $\mathbf{p}^{i+1}$ is a stationary point of $\mathcal{P}_p$ found by the generalized WMMSE algorithm, we only need to focus on solving $\mathcal{P}_{\theta_n}\left(\mathbf{p}, \boldsymbol{\theta}_{-n}, \boldsymbol{\pi}\right)$ with $\mathbf{p}$ satisfying $\sum_{k \in \mathcal{S}_n} \mathbb{I}\left(p_{k,n} > 0\right) \leq 1, \forall n$ according to Theorem 3. In the following, we exploit the specific structure of problem $\mathcal{P}_{\theta_n}$ to design a greedy IC algorithm which can find the optimal solution of $\mathcal{P}_{\theta_n}$with only linear complexity $O\left(|\mathcal{K}_n|\right)$, when $\sum_{k \in \mathcal{S}_n} \mathbb{I}\left(p_{k,n} > 0\right) \leq 1, \forall n$.

Let $k_n^* = \arg\max_{l \in \mathcal{S}_n} p_{l,n}$ denote the user which is allocated with the largest transmit power at BS n. Let $\overline{K}_n = |\mathcal{K}_n \backslash k_n^*|$ (it is possible that $k_n^* \notin \mathcal{K}_n$ and $\mathcal{K}_n \backslash \{k_n^*\} = \mathcal{K}_n$) and let σ_n be a permutation mapping from $\left\{1, ..., \overline{K}_n\right\}$ to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

Algorithm 3 Greedy IC Algorithm for $\mathcal{P}_{\theta_n}\left(\mathbf{p},\boldsymbol{\theta}_{-n},\boldsymbol{\pi}\right)$

1: **Input:** $\mathbf{p},\boldsymbol{\pi}$ and $\boldsymbol{\theta}_n,\boldsymbol{\theta}_{-n}$, where $\mathbf{p}$ satisfying $\sum_{k\in\mathcal{S}_n}\mathbb{I}\left(p_{k,n}>0\right)\leq 1,\forall n$.

2: **Initialization:** Let $\boldsymbol{\theta}_n^*=\boldsymbol{\theta}_n$ and

$$WSR_n^*=\sum_{k\in\mathcal{K}_n\backslash k_n^*}\mu_k R_k^{'}\left(\theta_k\right)+\mu_{k_n^*}\min_{l\in k_n^*\cup\overline{\mathcal{U}}_n}\overline{R}_{k,n,l}^{'},$$

where $R_k^{'}\left(\theta_k\right)$ is the data rate of user k with user side IC control variables $\boldsymbol{\theta}_{-n},\theta_k$ (note that $R_k^{'},\forall k\in\mathcal{K}_n\backslash k_n^*$ only depends on $\boldsymbol{\theta}_{-n},\theta_k$).

3: **If** $p_{k_n^*,n}=0$, **goto** Line 9.

4: **For** $i=0$ **to** $\overline{K}_n$

5: **Let** $\theta_{\sigma_n(j)}=1,j=1,...,i$ **and** $\theta_{\sigma_n(j)}=0,j=i+1,...,\overline{K}_n$.

6: **Let**

$$WSR_n=\sum_{k\in\mathcal{K}_n\backslash k_n^*}\mu_k R_k^{'}\left(\theta_k\right)+\mu_{k_n^*}\min_{l\in k_n^*\cup\{\sigma_n(j),j=1,...,i\}}\overline{R}_{k,n,l}^{'}.$$

7: **If** $WSR_n>WSR_n^*$, **let** $WSR_n^*=WSR_n^*$ and $\boldsymbol{\theta}_n^*=\boldsymbol{\theta}_n$.

8: **End For**

9: **Output** $\boldsymbol{\theta}_n^*$.

the set $\mathcal{K}_n\backslash k_n^*$ such that $\overline{R}_{k_n^*,n,\sigma_n(1)}^{'}\geq\overline{R}_{k_n^*,n,\sigma_n(2)}^{'}\geq...\geq\overline{R}_{k_n^*,n,\sigma_n\left(\overline{K}_n\right)}^{'}$, where $\overline{R}_{k,n,l}^{'},\forall k,n,l$ is given in (7). Then the optimal solution of $\mathcal{P}_{\theta_n}$ must have the following structure.

Theorem 4 (Structural Properties of the Optimal Solution of $\mathcal{P}_{\theta_n}$)**.** *Consider Problem $\mathcal{P}_{\theta_n}\left(\mathbf{p},\boldsymbol{\theta}_{-n},\boldsymbol{\pi}\right)$ with $\mathbf{p}$ satisfying $\sum_{k\in\mathcal{S}_n}\mathbb{I}\left(p_{k,n}>0\right)\leq 1,\forall n$. If $p_{k_n^*,n}=0$, any feasible solution $\boldsymbol{\theta}_n\in\{0,1\}^{|\mathcal{K}_n|}$ is optimal for $\mathcal{P}_{\theta_n}\left(\mathbf{p},\boldsymbol{\theta}_{-n},\boldsymbol{\pi}\right)$. Otherwise, the optimal solution $\boldsymbol{\theta}_n^*$ of $\mathcal{P}_{\theta_n}\left(\mathbf{p},\boldsymbol{\theta}_{-n},\boldsymbol{\pi}\right)$ must satisfy: $\theta_{\sigma_n(j)}^*=1,j=1,...,i^*$ and $\theta_{\sigma_n(j)}^*=0,j=i^*+1,...,\overline{K}_n$ for some $i^*\in\left\{0,...,\overline{K}_n\right\}$.*

Please refer to Appendix D for the detailed proof.

Based on optimal solution structure in Theorem 4, we propose a greedy IC algorithm in Algorithm 3 to find the optimal solution of $\mathcal{P}_{\theta_n}$ with only linear complexity.

C. PU-DoD Algorithm for Solving $\mathcal{P}_\pi$

Clearly, for fixed $\mathbf{p},\boldsymbol{\theta}$, the data rate of user k only depends on the decoding order π_k at user k. Hence, Problem $\mathcal{P}_\pi\left(\mathbf{p},\boldsymbol{\theta}\right)$ can be decomposed into K independent subproblems as

$$\mathcal{P}_\pi^k\left(\mathbf{p},\boldsymbol{\theta}\right):\ \max_{\pi_k}\sum_{n\in\mathcal{B}_k}R_{k,n}^{'},$$

where $\mathcal{B}_k=\left\{n:\ n\in\mathcal{C}_k,p_{k,n}>0\right\}$. If $\mathcal{B}_k\cap\{1,...,N_0\}=\emptyset$, i.e., the serving BSs of user k do not contain macro BSs, $\mathcal{P}_\pi^k\left(\mathbf{p},\boldsymbol{\theta}\right)$ reduces to a sum-rate maximization problem in multi-access channel (MAC). In this case, it is well known that any decoding order is optimal for $\mathcal{P}_\pi^k\left(\mathbf{p},\boldsymbol{\theta}\right)$. Otherwise, $R_{k,n}^{'}$ for $k\in\mathcal{B}_k$ and $n\in\{1,...,N_0\}$ is in general the minimum of several rate terms as given in (5). In this case, different decoding orders π_k may result in different sum-rate $\sum_{n\in\mathcal{B}_k}R_{k,n}^{'}$, and we can simply use exhaustive search to find the optimal decoding order. The complexity of exhaustive search over all decoding orders is $O\left(|\mathcal{B}_k|!\right)\leq O\left(N_c!\right)$. Since the number of candidate BSs N_c for each user is usually set as a small number (e.g., 2 or 3), such complexity is acceptable in practice. The overall PU-DoD algorithm for solving $\mathcal{P}_\pi$ is summarized in Algorithm 4.

Algorithm 4 PU-DoD Algorithm for $\mathcal{P}_\pi\left(\mathbf{p},\boldsymbol{\theta}\right)$

1: **Input:** $\mathbf{p},\boldsymbol{\theta}$ and $\boldsymbol{\pi}$.

2: **Initialization:** Let $\boldsymbol{\pi}^*=\boldsymbol{\pi}$.

3: **For** $k=1$ to K

4: **If** $\mathcal{B}_k\cap\{1,...,N_0\}\neq\emptyset$

5: **Calculate** $\pi_k^*=\arg\max_{\pi_k}\sum_{n\in\mathcal{B}_k}R_{k,n}^{'}$ using exhaustive search.

6: **End If**

7: **End For**

8: **Output** $\boldsymbol{\pi}^*$.

D. Convergence of the Overall Algorithm

In Step 1 of the WMMSE-AO algorithm, the generalized WMMSE algorithm with initial point $\mathbf{p}^i$ ensures that the WSR will be strictly increased unless $\mathbf{p}^i$ is already a stationary point of $\mathcal{P}_p\left(\boldsymbol{\theta}^i,\boldsymbol{\pi}^i\right)$. In Step 2, since the optimal solution is found for $\mathcal{P}_{\theta_n}\left(\mathbf{p}^{i+1},\boldsymbol{\theta}_{-n}^{i+1},\boldsymbol{\pi}^i\right)$, the WSR will be strictly increased unless $\boldsymbol{\theta}_n^i$ is already the optimal solution. Similarly, in step 3, the WSR will also be strictly increased unless $\boldsymbol{\pi}^i$ is already the optimal solution of $\mathcal{P}_\pi\left(\mathbf{p}^{i+1},\boldsymbol{\theta}^{i+1}\right)$. As a result, the following convergence result can be proved.

Theorem 5 (Convergence of WMMSE-AO)**.** *The WMMSE-AO algorithm monotonically increases the WSR $\sum_{k=1}^K\mu_k R_k^{'}$ and converges to stationary solution $\left(\mathbf{p}^*,\boldsymbol{\theta}^*,\boldsymbol{\pi}^*\right)$ of Problem $\mathcal{P}$. Moreover, $\mathbf{p}^*$ satisfies $\sum_{k\in\mathcal{S}_n}\mathbb{I}\left(p_{k,n}^*>0\right)\leq 1,\forall n$.*

After obtaining a stationary solution $\left(\mathbf{p}^*,\boldsymbol{\theta}^*,\boldsymbol{\pi}^*\right)$ of $\mathcal{P}$ using the WMMSE-AO algorithm, the solution for the original problem in (4) is given by $\boldsymbol{\mathcal{B}}^*=\left\{\mathcal{B}_k^*,k=1,...,K\right\},\boldsymbol{\theta}^*,\boldsymbol{\pi}^*,\left\{p_n^*=\sum_{k\in\mathcal{S}_n}p_{k,n}^*\right\}$, where $\mathcal{B}_k^*=\left\{n:\ n\in\mathcal{C}_k,p_{k,n}^*>0\right\}$. The corresponding objective value (WSR) of Problem (4) is the same as that of Problem $\mathcal{P}$ at the stationary solution $\left(\mathbf{p}^*,\boldsymbol{\theta}^*,\boldsymbol{\pi}^*\right)$.

E. Low Complexity WMMSE-AO Algorithm

By exploiting the specific problem structure, the combinatorial optimizations of the user-side IC and decoding order in Step 2 and Step 3 of WMMSE-AO only have linear complexity w.r.t. the number of users K. The main complexity lies in the generalized WMMSE in Step 1, which needs to solve a sequence of convex problems in (12) until convergence. In this section, we propose a modified WMMSE-AO algorithm to reduce the complexity without affecting the performance. Specifically, in Step 1 of WMMSE-AO, instead of running the generalized WMMSE until convergence, only one iteration of the generalized WMMSE is performed. Compared to the WMMSE-AO, each iteration of the low complexity WMMSE-AO only needs to solve one convex problem in (12) and thus the complexity is significantly reduced. Moreover, the low complexity WMMSE-AO is still guaranteed to converge to a stationary solution of $\mathcal{P}$ as explained below. In Step 2 of the low complexity WMMSE-AO, $\mathbf{p}^{i+1}$ may no longer be a stationary point of $\mathcal{P}_p$. In this case, the greedy IC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

Parameters	Values
Network layout	7 macro cells, 6 pico BSs & 7 users per macro cell
Inter-site distance	500m
Number of antennas	BS 2 antennas, user 1 antenna
BS transmit power	Macro: 46 dBm, Pico 30 dBm
Channel model	IMT-Advanced Channel Model [20, Annex B]
Thermal noise	- 174 dBm/Hz
Bandwidth	10 MHz
BS selection bias	β =9 dB

Table I: Key simulation parameters.

algorithm in Step 2 may not find the optimal solution of $\mathcal{P}_{\theta_n}$. However, each step of the low complexity WMMSE-AO can still monotonically increase the WSR $\sum_{k=1}^{K}\mu_k R_k^{'}$ and converge to a solution $(\mathbf{p}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)$. At the converged solution, it can be shown that $\mathbf{p}^*$ must be a stationary point of $\mathcal{P}_p$. Otherwise, Step 1 will strictly increase the WSR. As a result, at the converged solution, the greedy IC algorithm in Step 2 must find the optimal solution of $\mathcal{P}_{\theta_n}$ and thus the converged solution must be a stationary solution of $\mathcal{P}$.

The proposed algorithm based on AO has both theoretical and practical importance. First, by decomposing the problem into subproblems using AO and further exploiting specific structures of subproblems, the proposed algorithm only has polynomial complexity. Second, as pointed out in [19], AO-based algorithms facilitate parallel and distributed implementation, which may potentially reduce the signaling overhead and is more scalable to large scale systems. Finally, the proposed algorithm reveals that the optimal power allocation and user side IC must satisfy the structural properties in Theorem 1 and 4, respectively. These structural properties can be used as a guidance for the design of heuristic/distributed algorithms that are more suitable for practical implementation.

V. Simulations

Consider a 7-cell wrapped-around HetNet with the simulation parameters listed in Table I. We compare the performance of the proposed scheme with the following 4 baselines.

- **Baseline 1 (WMMSE)** [14]: Fixed BUA is used where each user is associated with a single BS according to the *biased cell selection* mechanism in [15] with bias β =9 dB. The WMMSE algorithm in [14] is used to calculate the power allocation. There is no user side IC.
- **Baseline 2 (Static user side IC)** [11]: Fixed BUA is used. The generalized WMMSE algorithm is used to calculate the power allocation and the static user side IC in [11] is employed at each user to cancel the strongest interference.
- **Baseline 3 (Dynamic user side IC)**: This scheme is a special case of the proposed scheme when the number of candidate BSs for each user is $N_c = 1$, i.e., fixed BUA is used together with the dynamic user side IC.
- **Baseline 4 (IPA JPCSA)** [8]: This is the interference pricing based joint power control, scheduling, and association algorithm in [8]. There is no user side IC.

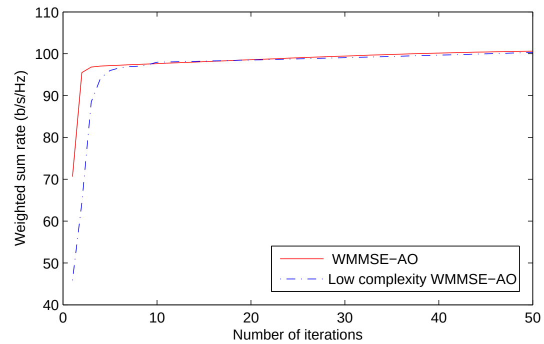


Figure 2: Convergence of (low complexity) WMMSE-AO algorithms. The number of candidate BSs for each user is $N_c = 2$.

Schemes	Sum rate	Gain (3)	Gain (2)
Proposed, $N_c = 3$	1056 Mbps	0%	-4.42%
Proposed, $N_c = 2$	1009 Mbps	4.63%	0%
Baseline 1	750.8 Mbps	40.65%	34.43%
Baseline 2	823.2 Mbps	28.28%	22.6%
Baseline 3	926.1 Mbps	14.02%	8.98%
Baseline 4	900.0 Mbps	17.33%	12.14%

Table II: Sum rates $\sum_{k=1}^{K} R_k^{'}$ of different schemes when the weights are chosen to be $\mu_k = 1, \forall k$. The third column is the gain of the proposed scheme with $N_c = 3$ over the other schemes. The fourth column is the gain of the proposed scheme with $N_c = 2$ over the other schemes.

A. Convergence of the WMMSE-AO

In Fig. 2, we plot the objective value $\sum_{k=1}^{K}\mu_k R_k^{'}$ achieved by the (low complexity) WMMSE-AO algorithms versus the number of iterations, where the weights are chosen to be $\mu_k = 1, \forall k$. It can be seen that the WMMSE-AO converges quickly. Moreover, the low complexity WMMSE-AO algorithm has similar convergence speed as the WMMSE-AO but with lower complexity.

B. Performance Evaluations

In Table II, we set the weights as $\mu_k = 1, \forall k$ and compare the sum rate performance of different schemes. It can be seen that the proposed scheme has *large* gain over all baselines. In particular, the performance gap between the proposed scheme and baseline 3 (dynamic user side IC) represents the multi-BS diversity gain due to cell-free BUA. The multi-BS diversity gain increases with the number of candidate BSs N_c at the cost of higher complexity. On the other hand, the performance gap between baseline 3 and baseline 1/baseline 2 represents the gain due to the dynamic user side IC. This gain demonstrates

Schemes	PFS Utility	Gain (3)	Gain (2)
Proposed, $N_c = 3$	21.68	0%	-10.66%
Proposed, $N_c = 2$	19.37	11.94%	0%
Baseline 1	5.60	287%	245.74%
Baseline 2	10.28	110.87%	88.39%
Baseline 3	16.11	34.5%	20.16%
Baseline 4	9.72	123.11%	99.32%

Table III: PFS utility $\sum_{k=1}^{K}\log\overline{r}_k$ of different schemes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

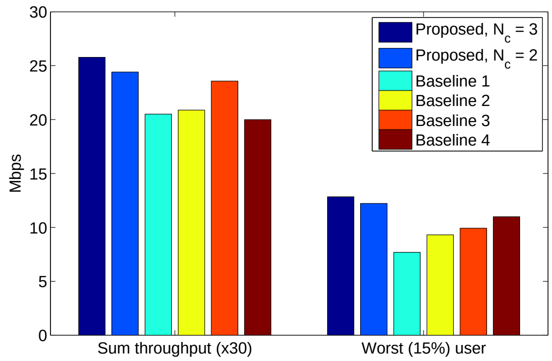


Figure 3: Throughput comparisons of different schemes under PFS.

the importance of adapting the user side IC according to instantaneous channel gains.

In Table III, we set the weights as $\mu_k = 1/\overline{r}_k$ according to the proportional fairness (PFS) criterion and compare the PFS utility $\sum_{k=1}^{K} \log \overline{r}_k$ of different schemes, where $\overline{r}_k$ is the long term average data rate for user k. In this case, the proposed scheme achieves a *huge* gain over all baselines. This shows that the proposed scheme can achieve a much better tradeoff between sum throughput and fairness due to the joint optimization of cell-free BUA and dynamic user side IC.

Fig. 3 compares the throughput of different schemes under PFS criterion. The proposed scheme achieves both higher sum throughput and worst user throughput than the baselines, and the throughput gain increases with N_c. Note that the worst user throughput gain is much larger than the sum throughput gain because the edge users can benefit more from the dynamic user side IC and multi-BS diversity gain. In practice, it is very important to improve the performance of the edge (worst) users since they are the performance bottleneck of the system.

C. Impact of Imperfect User Side IC

In practice, the user side IC cannot be perfect because of the decoding error for the interference. In Fig. 4, we plot the PFS utility versus the block error rate (BLER) of decoding the interference. The performance of all static/dynamic user side IC schemes degrades as the BLER of decoding the interference increases. However, the proposed scheme degrades with the BLER at a much lower rate compared to the baseline IC schemes. This is because in the proposed scheme, the strongest interference can be potentially avoided by the more flexible cell-free BUA, and thus the probability of having to cancel the strongest interference using dynamic user side IC become smaller. As a result, the proposed scheme is less sensitive to the BLER of decoding the interference. As the BLER increases, baseline 2 and 3 may become even worse than baseline 1 and 4 which has no user side IC. However, the proposed scheme still has large gain over all baselines even when the BLER is as high as 0.5.

VI. Conclusion

The joint optimization of BUA, power allocation and user side IC in HetNet is formulated as a WSR maximization

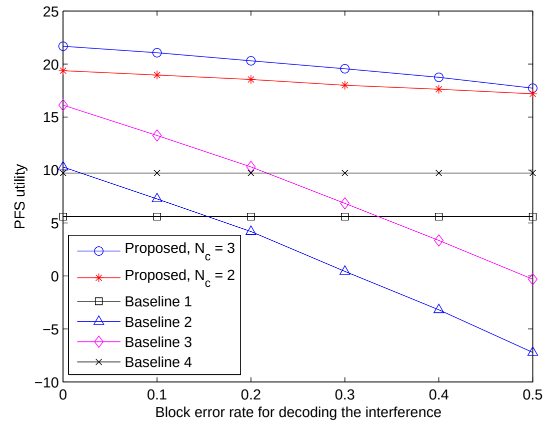


Figure 4: PFS utility versus block error rate for decoding the interference.

problem. We first transform the problem into a more tractable form. Then we propose an efficient WMMSE-AO algorithm to solve a stationary solution of the transformed problem. Specifically, in each iteration, the non-convex non-smooth power allocation subproblem is first solved using a generalized WMMSE algorithm which is a generalization of the WMMSE algorithm in [14] to the case with non-differentiable WSR objective function. Then the combinatorial user side IC subproblem is solved using a greedy IC algorithm, which is able to find the optimal solution with only linear complexity. Finally, simulations show that the proposed scheme has significant gain over various baselines.

Appendix

A. Proof of Theorem 1

Introducing auxiliary variables $\{r_{k,n} : \forall k, n \in \mathcal{C}_k\}$, $\mathcal{P}$ with fixed $\boldsymbol{\theta}, \boldsymbol{\pi}$ can be reformulated in an epigraph form as [18]

$$\begin{aligned} \widetilde{\mathcal{P}}_p(\boldsymbol{\theta}, \boldsymbol{\pi}) : \quad & \max_{\mathbf{p}, \{r_{k,n}:\forall k, n\in\mathcal{C}_k\}} \sum_{k=1}^{K} \mu_k \sum_{n\in\mathcal{C}_k} r_{k,n}, \qquad (13) \\ \text{s.t.} \quad & \sum_{k\in\mathcal{S}_n} p_{k,n} \le P_n, \ \forall n \in \cup_k \mathcal{C}_k, \\ & r_{k,n} \le \overline{R}'_{k,n,l}, \ \forall n, k \in \mathcal{S}_n, l \in k \cup \overline{\mathcal{U}}_n. \end{aligned}$$

Clearly, the optimal power allocation $\mathbf{p}^\star$ is also the optimal solution of $\widetilde{\mathcal{P}}_p(\boldsymbol{\theta}^\star, \boldsymbol{\pi}^\star)$. For fixed $\boldsymbol{\theta} = \boldsymbol{\theta}^\star, \boldsymbol{\pi} = \boldsymbol{\pi}^\star$ and $p_{k,n} = p^\star_{k,n'}, \forall k \in \mathcal{S}_{n'}, n' \neq n$, the rate of data stream $x_{k,n}$ is

$$\begin{aligned} r_{k,n} &= \log\left(1 + \frac{\overline{g}_{k,n} p_{k,n}}{1 + \overline{g}_{k,n} \sum_{l\in\mathcal{S}_n\setminus\{k\}} p_{l,n}}\right), \\ \overline{g}_{k,n} &= \min_{l\in k\cup\overline{\mathcal{U}}_n} \frac{g_{l,n}}{IN^\star_{k,n,l} - \sum_{l\in\mathcal{S}_n\setminus\{k\}} p^\star_{l,n} g_{l,n}}, \forall k, \end{aligned}$$

where $IN^\star_{k,n,l}$ is the interference-plus-noise power for the decoding of $x_{k,n}$ at user l under the optimal power allocation $p_{k,n} = p^\star_{k,n'}, \forall k \in \mathcal{S}_{n'}, \forall n'$, $IN^\star_{k,n,l} - \sum_{l\in\mathcal{S}_n\setminus\{k\}} p^\star_{l,n} g_{l,n}$ is the interference-plus-noise power excluding the interference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

contributed by BS n. Hence, for fixed $\boldsymbol{\theta}^\star, \boldsymbol{\pi}^\star$ and $p_{k,n} = p^\star_{k,n'}, \forall k \in \mathcal{S}_{n'}, n' \neq n$, Problem $\widetilde{\mathcal{P}}_p(\boldsymbol{\theta}^\star, \boldsymbol{\pi}^\star)$ reduces to

$$\max_{p_{k,n}\geq 0, \forall k\in\mathcal{S}_n} \sum_{k\in\mathcal{S}_n} \mu_k \log\left(1+\frac{\overline{g}_{k,n}p_{k,n}}{1+\overline{g}_{k,n}\sum_{l\in\mathcal{S}_n\setminus\{k\}}p_{l,n}}\right) \tag{14}$$
$$\text{s.t.} \sum_{k\in\mathcal{S}_n} p_{k,n} \leq \overline{P},$$

where $\overline{P} = \sum_{k\in\mathcal{S}_n} p^\star_{k,n}$. In the following, we show that the optimal solution of Problem (14) satisfies $\sum_{k\in\mathcal{S}_n} \mathbb{I}\left(p^\star_{k,n} > 0\right) \leq 1$. This can be proved by successively applying the following lemma.

Lemma 1. *Consider the following WSR maximization in a two-user SISO broadcast channel:*

$$\max_{p_1,p_2\geq 0} \sum_{k=1}^{2} \mu_k \log\left(1+\frac{\overline{g}_k p_k}{1+\sum_{l\neq k}\overline{g}_k p_l}\right), \textit{ s.t. } p_1+p_2 \leq \overline{P}. \tag{15}$$

The optimal solution $p^\star_1, p^\star_2$ of (15) must satisfy $p^\star_1 p^\star_2 = 0$.

Proof: It can be shown that the optimal solution must satisfy $p^\star_1 + p^\star_2 = \overline{P}$. Hence, the objective function can be rewritten as a function of p_1 as

$$f(p_1) = \mu_1 \log\left(\frac{1+\overline{g}_1\overline{P}}{1+\overline{g}_1\left(\overline{P}-p_1\right)}\right) + \mu_2 \log\left(\frac{1+\overline{g}_2\overline{P}}{1+\overline{g}_2 p_1}\right).$$

Note that $\frac{df(p_1)}{dp_1} = 0$ has a unique solution given by

$$p^\circ_1 = \frac{1}{\mu_1+\mu_2}\left(\frac{\mu_2\left(1+p\overline{g}_1\right)}{\overline{g}_1} - \frac{\mu_1}{\overline{g}_2}\right).$$

If $p^\circ_1 \geq \overline{P}$ or $p^\circ_1 \leq 0$, then the optimal solution must be one of the boundary points $p^\star_1 = 0$ or $p^\star_1 = \overline{P}$, which satisfies $p^\star_1 p^\star_2 = 0$. Therefore, we only need to consider the case when $0 < p^\circ_1 < \overline{P}$. It follows from $p^\circ_1 > 0$ that $\left.\frac{df(p_1)}{dp_1}\right|_{p_1=0} < 0$. Since $\frac{df(p_1)}{dp_1} = 0$ has a unique solution p°_1, it follows from $\left.\frac{df(p_1)}{dp_1}\right|_{p_1=0} < 0$ that the objective $f(p_1)$ must achieve the minimum at $p_1 = p^\circ_1$ and thus the optimal solution must also be one of the boundary points $p^\star_1 = 0$ or $p^\star_1 = \overline{P}$. ■

For any two users $k \neq l \in \mathcal{S}_n$, if we fix the power for other users as the optimal solution: $p_{k',n} = p^\star_{k',n}, \forall k' \in \mathcal{S}_n\setminus\{k,l\}$, the optimal power allocation $p^\star_{k,n}, p^\star_{l,n}$ of user k and user l must be the optimal solution of a WSR maximization in an equivalent two-user SISO broadcast channel. According to Lemma 1, we have $p^\star_{k,n}p^\star_{l,n} = 0, \forall k \neq l \in \mathcal{S}_n$, from which it can be shown that $\sum_{k\in\mathcal{S}_n} \mathbb{I}\left(p^\star_{k,n} > 0\right) \leq 1$. Finally, Theorem 1 follows from $\sum_{k\in\mathcal{S}_n} \mathbb{I}\left(p^\star_{k,n} > 0\right) \leq 1, \forall n$ immediately.

B. Proof of Theorem 2

First, given the optimal solution $\mathbf{p}^\star$ of Problem $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$, if we let $u^\star_{k,n,l} = IN^{\star-1}_{k,n,l}\sqrt{g_{l,n}p^\star_{k,n}}, \forall k,n,l$ and $w^\star_{k,n,l} = \left(1-u^\star_{k,n,l}\sqrt{g_{l,n}p^\star_{k,n}}\right)^{-1}, \forall k,n,l$ (where $IN^\star_{k,n,l}$ is the interference-plus-noise power under the optimal solution $\mathbf{p}^\star$), then it follows from Theorem 1 in [14] that $w^\star_{k,n,l}e^\star_{k,n,l} - \log w^\star_{k,n,l} = \overline{R}'^\star_{k,n,l}$ (where $e^\star_{k,n,l}$ and $\overline{R}'^\star_{k,n,l}$ are the MSE and achievable rate under the optimal solution $\mathbf{p}^\star$), and thus the objective value of $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ achieved by the feasible solution $\mathbf{q}^\star = \sqrt{\mathbf{p}^\star}, \mathbf{u}^\star, \mathbf{w}^\star$ is the same as the optimal value of Problem $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$. where $\sqrt{\mathbf{p}^\star}$ means $q^\star_{k,n} = \sqrt{p^\star_{k,n}}, \forall k,n$. This shows that the optimal objective value of Problem $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ is no less than that of $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$.

Second, given the optimal solution $\mathbf{q}^\star, \mathbf{u}^\star, \mathbf{w}^\star$ of Problem $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$, if we let $u^\circ_{k,n,l} = IN^{\star-1}_{k,n,l}\sqrt{g_{l,n}}q^\star_{k,n}, \forall k,n,l$ and $w^\circ_{k,n,l} = \left(1-u^\circ_{k,n,l}\sqrt{g_{l,n}}q^\star_{k,n}\right)^{-1}, \forall k,n,l$ (where $IN^\star_{k,n,l}$ is the interference-plus-noise power under the optimal solution $\mathbf{q}^\star$), then $\mathbf{q}^\star, \mathbf{u}^\circ, \mathbf{w}^\circ$ must also be the optimal solution of $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$. Moreover, it follows from Theorem 1 in [14] that $w^\circ_{k,n,l}e^\circ_{k,n,l} - \log w^\circ_{k,n,l} = \overline{R}'^\star_{k,n,l}$ (where $e^\circ_{k,n,l}$ and $\overline{R}'^\star_{k,n,l}$ are the MSE and achievable rate under the optimal solution $\mathbf{q}^\star, \mathbf{u}^\circ, \mathbf{w}^\circ$). Hence the objective value of $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ achieved by the feasible solution $\mathbf{p}^\star = \mathbf{q}^\star \circ \mathbf{q}^\star$ is the same as the optimal value of Problem $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$. Finally, Theorem 2 follows immediately from the above analysis.

C. Proof of Theorem 3

Introducing auxiliary variables $\{r_{k,n}: \forall k, n \in \mathcal{C}_k\}$, $\mathcal{P}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ can be reformulated in an epigraph form as [18]

$$\widetilde{\mathcal{P}}'_p(\boldsymbol{\theta}, \boldsymbol{\pi}): \max_{\mathbf{q},\mathbf{u},\mathbf{w},\{r_{k,n}:\forall k,n\in\mathcal{C}_k\}} \sum_{k=1}^{K} \mu_k \sum_{n\in\mathcal{C}_k} r_{k,n},$$
$$\text{s.t.} \sum_{k\in\mathcal{S}_n} q^2_{k,n} \leq P_n, \ \forall n \in \cup_k \mathcal{C}_k,$$
$$r_{k,n} \leq w_{k,n,l}e_{k,n,l} - \log w_{k,n,l}, \ \forall n, k \in \mathcal{S}_n, l \in k \cup \overline{\mathcal{U}}_n.$$

The reformulated problem $\widetilde{\mathcal{P}}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ has a differentiable objective function. The generalized WMMSE algorithm is the block coordinate descent method applied to $\widetilde{\mathcal{P}}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$, where in Step 3, both $\mathbf{q}$ and $r_{k,n}$ are optimized for fixed $\mathbf{u}, \mathbf{w}$ (After solving the convex problem (12) with fixed $\mathbf{u}, \mathbf{w}$ in Step 3, the optimal $r_{k,n}$ is implicitly given by $r_{k,n} = \min_{l\in k\cup\overline{\mathcal{U}}_n} w_{k,n,l}e_{k,n,l} - \log w_{k,n,l}$). It follows from the general optimization theory [21] that the generalized WMMSE algorithm converges to a stationary point $\left(\mathbf{q}^*, \mathbf{u}^*, \mathbf{w}^*, \left\{r^*_{k,n}\right\}\right)$ of $\widetilde{\mathcal{P}}'_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ (where $r^*_{k,n} = \min_{l\in k\cup\overline{\mathcal{U}}_n} w^*_{k,n,l}e^*_{k,n,l} - \log w^*_{k,n,l}$). Following similar arguments as that in Appendix C of [14], it can be shown that $\left(\mathbf{p}^* \triangleq \mathbf{q}^* \circ \mathbf{q}^*, \left\{r^*_{k,n}\right\}\right)$ is a stationary point of Problem $\widetilde{\mathcal{P}}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ defined in (13). Since $\widetilde{\mathcal{P}}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ is an epigraph form of $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$, $\mathbf{p}^*$ must also be a stationary point of $\mathcal{P}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ [18]. Finally, in Appendix B, we have already proved that any stationary point $\mathbf{p}^*$ of Problem $\widetilde{\mathcal{P}}_p(\boldsymbol{\theta}, \boldsymbol{\pi})$ must satisfy $\sum_{k\in\mathcal{S}_n} \mathbb{I}\left(p^*_{k,n} > 0\right) \leq 1, \forall n$. This completes the proof.

D. Proof of Theorem 4

Since $\sum_{k\in\mathcal{S}_n} \mathbb{I}(p_{k,n} > 0) \leq 1$, we must have $p_{k,n} = 0, \forall k \in \mathcal{S}_n \backslash k^*_n$. There are two cases. **Case 1** ($p_{k^*_n,n} = 0$): In this case, the transmit power of macro BS n is zero (i.e., there is no interference from macro BS n to the users) and thus

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2620962, IEEE Transactions on Signal Processing

any feasible $\boldsymbol{\theta}_n$ is optimal for $\mathcal{P}_{\theta_n}$. **Case 2** ($p_{k_n^*,n} > 0$): In this case, if $k_n^* \in \mathcal{K}_n$, macro BS n does not cause interference to user k_n^* because it is the serving BS of user k_n^*, and the choice of $\theta_{k_n^*}$ does not affect the WSR. Hence, we can focus on optimizing $\{\theta_k, k \in \mathcal{K}_n \backslash \{k_n^*\}\}$ only. For fixed $\mathbf{p}, \boldsymbol{\theta}_{-n}, \boldsymbol{\pi}$, the rate $R_k^{'}$ of user $k \in \mathcal{K}_n \backslash \{k_n^*\}$ only depends on θ_k and we can explicitly express it as function of θ_k: $R_k^{'}(\theta_k)$. When a user $k \in \mathcal{K}_n \backslash \{k_n^*\}$ decodes and cancels the interference from macro BS n (i.e., setting $\theta_k = 1$), there will be two effects. First, the data rate of user k will be improved from $R_k^{'}(\theta_k = 0)$ to $R_k^{'}(\theta_k = 1)$. Second, the rate $R_{k_n^*,n}^{'}$ of the data stream from macro BS n to user k_n^* will be limited by the rate term $\overline{R}_{k_n^*,n,k}^{'}$ as defined in (7), which is only a function of $\mathbf{p}$ and does not depend on $\boldsymbol{\theta}, \boldsymbol{\pi}$. Clearly, if $\theta_{\sigma_n(i)} = 1$, setting $\theta_{\sigma_n(j)} = 1, j = 1, ..., i-1$ will not decrease the rate $R_{k_n^*,n}^{'}$ but will increase the data rate of the users in $\{\sigma_n(j) : j = 1, ..., i-1\}$. As a result, the optimal solution of $\mathcal{P}_{\theta_n}$ must have the structure in Theorem 4.

REFERENCES

[1] J. Pang, J. Wang, D. Wang, G. Shen, Q. Jiang, and J. Liu, "Optimized time-domain resource partitioning for enhanced inter-cell interference coordination in heterogeneous networks," in *IEEE WCNC 2012*, pp. 1613–1617.

[2] A. Liu, V. Lau, L. Ruan, J. Chen, and D. Xiao, "Hierarchical radio resource optimization for heterogeneous networks with enhanced inter-cell interference coordination (eicic)," *IEEE Trans. Signal Processing*, vol. 62, no. 7, pp. 1684–1693, April 2014.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *in Proc. IEEE SPAWC 2015*, Jun. 2015, pp. 201–205.

[4] S. Vasudevan, R. N. Pupala, and K. Sivanesan, "Dynamic eICIC - a proactive strategy for improving spectral efficiencies of heterogeneous LTE cellular networks by leveraging user mobility and traffic dynamics," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4956–4969, Oct. 2013.

[5] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.

[6] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[7] P. H. Chiang, P. H. Huang, S. S. Sun, W. Liao, and W. T. Chen, "Joint power control and user association for traffic offloading in heterogeneous networks," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 4424–4429.

[8] J. Francis, S. Kalyanasundaram, B. Natarajan, R. Agrawal, and N. B. Mehta, "Downlink interference penalty algorithm for power control, scheduling, and user association," in *2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2015, pp. 506–512.

[9] Q. Ye, O. Y. Bursalioglu, H. C. Papadopoulos, C. Caramanis, and J. G. Andrews, "User association and interference management in massive MIMO hetnets," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2049–2065, May 2016.

[10] *3GPP Rel. 12 TR 36.866, Study on Network-Assisted Interference Cancellation and Suppression (NAIC) for LTE*, Mar. 2014.

[11] E. Hardouin, M. S. Hassan, and A. Saadani, "Downlink interference cancellation in LTE: Potential and challenges," in *in Proc. IEEE WCNC 2013*, Apr. 2013, pp. 3597–3602.

[12] M. Wildemeersch, T. Q. S. Quek, M. Kountouris, A. Rabbachin, and C. H. Slump, "Successive interference cancellation in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 62, no. 12, pp. 4440–4453, Dec. 2014.

[13] W. Nam, D. Bai, J. Lee, and I. Kang, "Advanced interference management for 5g cellular networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 52–60, May 2014.

[14] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted mmse approach to distributed sum-utility maximization for a mimo interfering broadcast channel," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4331 –4340, sept. 2011.

[15] *LTE Advanced: Heterogeneous Networks*, Qualcomm Incorporated, 2010.

[16] M. K. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," *IEEE Trans. Commun.*, Jan. 1999.

[17] H. Sato, "The capacity of the gaussian interference channel under strong interference," *IEEE Trans. Inform. Th.*, vol. 27, no. 6, pp. 786–788, Nov. 1981.

[18] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradients," 2003. [Online]. Available: https://web.stanford.edu/class/ee392o/subgrad.pdf

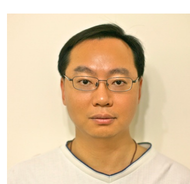
[19] D. P. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. MIT Press, 2003.

[20] *E-UTRA; Further Advancements for E-UTRA Physical Layer Aspects*, 3GPP TR 36.814. [Online]. Available: http://www.3gpp.org

[21] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear gauss-seidel method under convex constraints," *Operat. Res. Lett.*, vol. 26, pp. 127–136, 2000.

An Liu (S'07–M'09) received the Ph.D. and the B.S. degree in Electrical Engineering from Peking University, China, in 2011 and 2004 respectively.

From 2008 to 2010, he was a visiting scholar at the Department of ECEE, University of Colorado at Boulder. From 2011 to 2013, he was a Postdoctoral Research Fellow with the Department of ECE, HKUST, and he is currently a Research Assistant Professor. His research interests include wireless communication, stochastic optimization and compressive sensing.



Vincent K. N. Lau (SM'04–F'12) obtained B.Eng (Distinction 1st Hons) from the University of Hong Kong (1989-1992) and Ph.D. from the Cambridge University (1995-1997). He joined Bell Labs from 1997-2004 and the Department of ECE, Hong Kong University of Science and Technology (HKUST) in 2004. He is currently a Chair Professor and the Founding Director of Huawei-HKUST Joint Innovation Lab at HKUST. His current research focus includes robust and delay-optimal cross layer optimization for MIMO/OFDM wireless systems, interference mitigation techniques for wireless networks, massive MIMO, M2M and network control systems.