Optimizing Downlink Resource Allocation in Multiuser MIMO Networks via Fractional Programming and the Hungarian Algorithm

Ahmad Ali Khan, *Student Member, IEEE,* Raviraj Adve, *Fellow, IEEE,* and Wei Yu, *Fellow, IEEE*

***Abstract*—Optimizing the sum-log-utility for the downlink of multi-frequency band, multiuser, multiantenna networks requires joint solutions to the associated beamforming and user-scheduling problems through the use of cloud radio access network (CRAN) architecture; optimizing such a network is, however, non-convex and NP-hard. In this paper, we present a novel iterative beamforming and scheduling strategy based on fractional programming and the Hungarian algorithm. The beamforming strategy allows us to iteratively maximize the chosen objective function in a fashion similar to block-coordinate ascent. Furthermore, based on the crucial insight that, in the downlink, the interference pattern remains fixed for a given set of beamforming weights, we use the Hungarian algorithm as an efficient approach to optimally schedule users for the given set of beamforming weights. Specifically, this approach allows us to select the best subset of users (amongst the larger set of all available users). Our simulation results show that, in terms of average sum-log-utility, as well as sum-rate, the proposed scheme substantially outperforms both the state-of-the-art multicell weighted minimum mean-squared error (WMMSE) and greedy proportionally fair WMMSE schemes, as well as standard interior-point and sequential quadratic solvers. Importantly, our proposed scheme is also far more computationally efficient than the multicell WMMSE scheme.**

I. Introduction

The improvements in spectral efficiency, throughput and quality-of-service achieved by utilizing multiantenna networks have been extensively documented in the literature [2], [3]. In particular, optimizing the resource allocation in such multiantenna networks, by designing beamforming weights and scheduling specific users from the larger pool of potential users, is central to fully exploiting the finite wireless resources available and maximizing spectral efficiency [4], [5]. However, designing efficient resource allocation schemes remains challenging, since many utility functions of practical interest, such as sum-rate, sum-log-utility and and min-rate, are inherently nonconvex; in fact, the associated optimization problems for each of these objective functions have been found to be NP-hard [6]. Thus, solving these optimization problems to global optimality entails impractical computational complexity even for very small network sizes [7].

One solution to these problems is to utilize single-cell schemes, such as zero-forcing or matched filtering, in which base stations ignore intercell interference when designing beamforming weights and making scheduling decisions [8]. While far from globally optimal, such uncoordinated schemes offer three significant advantages: first, they are analytically tractable in the sense that they can be analyzed using tools from probability theory and stochastic geometry to yield accurate estimates of the data rates achieved by users (and hence objective functions like the network sum-rate) [9]; second, these schemes are computationally efficient, especially compared to globally optimal techniques or iterative block-coordinate ascent based algorithms like weighted minimum mean squared error (WMMSE) [10] processing; third, and likely most important, in these schemes each base-station (BS) requires channel state information (CSI) from only its own users, not for users in other cells[1]. As such, these uncoordinated schemes offer a useful benchmark against which to evaluate the performance of more sophisticated resource allocation strategies.

Coordinated resource allocation schemes, in which base stations jointly design their scheduling and beamforming decisions, improve on uncoordinated schemes. Such joint design leads to improved quality-of-service since it helps to mitigate the effects of both inter-cell and intra-cell interference [11]. In doing so, however, such schemes inevitably incur increased computational complexity, as compared to uncoordinated schemes, since these algorithms optimize across multiple BSs. Since the objective functions for most utility maximization problems are nonconvex, such schemes typically rely on block-coordinate-ascent [10], successive convex approximation [12] or other heuristic methods [13], [14] to reach, at best, a local optimum.

A number of coordinated schemes have been developed in the literature; for example, in [15] the authors develop an interference pricing and greedy proportionally fair (PF) scheduling algorithm to maximize the weighted sum rate (WSR) for the downlink. The proposed scheme demonstrates excellent performance in terms of average sum-log-utility but is not guaranteed to be nondecreasing in the objective function since greedy scheduling is used. In [12], Weeraddana et al. propose an algorithm based on the successive convex approximation approach to optimize the needed beamforming weights and power allocation in order to solve the general WSR maximization problem for the downlink of a multiple input multiple output (MIMO) cellular network. The algorithm

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, Ontario, ON M5S 3G4, Canada. E-mails: (akhan, rsadve, weiyu)@ece.utoronto.ca. The materials in this paper have been presented in part at IEEE Global Communications Conference (Globecom) 7th International Workshop on "Emerging Technologies for 5G and Beyond Wireless and Mobile Networks (ET5GB)", Abu Dhabi, December 2018 [1].

[1]We note that, like other works that focus on algorithm development [6], [10], the acquisition of CSI, its overhead and quality is beyond the scope of this paper. However, we do acknowledge that this is a vitally important problem in wireless networks.

requires minimal exchange of information between cooperating BSs; however, the algorithm is also shown to underperform the WMMSE scheme of [10]. Additionally, one alternative is to employ worst-case weighted sum-rate maximization [16], although such an approach is generally better suited to settings with uncertainty in channel vectors.

The work in [10] develops the WMMSE algorithm by demonstrating the equivalence of minimizing the weighted MSE and maximizing the WSR and adopting a block-coordinate ascent strategy to reach a (guaranteed) local optimum of the original WSR objective function. The algorithm iterates between obtaining beamforming weights and a set of auxiliary variables, optimizing one while the other is kept fixed. This WMMSE approach demonstrates excellent performance and is, thus, widely utilized as a benchmark against which the performance of other coordinated resource allocation schemes is compared.

The work in [10] does not address the important problem of user scheduling, i.e., choosing a set of users to serve from the larger set of available users. One solution is to use the multicell WMMSE scheme, with *all users across all cells* are scheduled – the BSs then jointly design beamforming weights for each and every user, as in [10]. Eventually, after a number of iterations, the power assigned (the norm of the beamforming vectors) to most users will be essentially zero and these users are, then, *implicitly* not scheduled. However, since this set of "unscheduled" users is unknown a priori, beamforming weights have to be calculated for *all users* in the network for *each iteration of the algorithm*. This is extremely computationally expensive since a large matrix needs to be inverted in each step. Furthermore, it is worth emphasizing that the multicell WMMSE scheme, as described, is *not globally optimal*. Additionally, as the authors in [17] have observed, when scheduling all users, the WMMSE algorithm tends to get stuck in a low-quality locally optimal solution. Despite these drawbacks, the WMMSE algorithm remains the benchmark against which other resource allocation algorithms are evaluated [18]–[20].

A lower-complexity approach is to alternately optimize the scheduling and beamforming variables in a fashion similar to that proposed by [21] and [22], by alternately utilizing the WMMSE algorithm for beamforming and updating the scheduling decisions using the greedy PF scheme. However, because of the greedy step, this approach is also not guaranteed to be nondecreasing in the original WSR objective function.

Globally optimal schemes to solve sum-rate and weighted-sum-rate optimization problems have also been formulated in the literature, using the framework of monotonic optimization [4], [7], [23], [24], as well as geometeric and airthmetic-mean methods [25]. For example, the authors in [7] develop an algorithm to find the globally optimal beamformers to maximize the WSR for the downlink of a multiuser multiantenna network. However, as the authors in [23] note, these globally optimal schemes require impractical computational complexity for even small systems, and are thus used almost exclusively as benchmarks for very small network sizes.

It is worth emphasizing that extensive CSI exchange between BSs and computational resources are required in order to enable both locally and globally optimal resource management schemes. These requirements are best served by the utilization of cloud radio access networks (CRANs), which allow for flexible deployment of resource allocation algorithms and on-demand processing while utilizing relatively inexpensive hardware at the BS [26]–[30]. Deploying dedicated hardware at the BS level to implement individual algorithms is both technically challenging and cost ineffective [26], [30]; on the other hand, through the use of CRAN, low-cost remote radio heads can be utilized at each BS, while virtualized baseband processing units for the entire network can be implemented in the cloud and easily altered to enable different resource management schemes and capabilities [28], [29]. Thus, the CRAN architecture is necessary in order to enable coordinated resource allocation and is stated explicitly or assumed implictly in the various coordinated schemes detailed in the literature [12].

In summary, effective multicell resource allocation schemes with relatively low complexity are, as yet, not available. It is this gap in the literature that we address here. Specifically, we develop an iterative scheduling and beamforming strategy to find an effective solution to the problem of maximizing the average sum-log-utility function for the downlink of a multiuser MIMO network. Using the framework of fractional programming, originally developed in [18] and [17] for uplink problems and extended to the matrix setting in [31], we derive a scheme similar to a block-coordinate ascent scheme. In [18], fractional programming has shown large performance benefits for utility maximization in the *uplink* setting. Similarly, in [32], the authors utilize fractional programming to jointly optimize power control and scheduling decisions for energy efficiency maximization; the proposed algorithm can be implemented in both distributed and centralized fashion and provides excellent convergence and performance properties. In both these scenarios, the interference pattern changes with the scheduling decisions; thus, optimization across multiple cells can provide considerable benefit. This paper demonstrates the efficacy of fractional programming for the *downlink*, where we exploit the *fixed* interference pattern to improve performance and reduce computational complexity in the user scheduling step.

In contrast to our conference-length work in [1], this paper considers the most general setting of the problem: we derive the algorithm and present results for proportionally-fair WSR and sum-rate maximization through scheduling and beamforming across multiple frequency bands with both joint and decoupled power constraints across the bands. Deriving the algorithm for this setting is considerably more challenging than the single-band case considered in [1]; this is especially true for the joint power allocation across multiple bands in which, despite the orthogonality of the bands, *all* beamforming weights and scheduling decisions become coupled due to the power constraint. Nonetheless, we demonstrate that fractional programming allows us to decouple these optimization variables and solve for an effective solution with guaranteed nondecreasing convergence. Specifically, the contributions of this paper are:

- We formulate the downlink sum-log-utility maximization problem as a WSR problem in the general case of multiple interfering cells, multiple frequency bands, and a large number of potential users per cell.
- We develop a joint beamforming and user scheduling algorithm based on fractional programming and the Hungarian algorithm. The Hungarian algorithm selects the optimal set of users from the much larger pool of potential users, for a given set of beamforming weights, in *polynomial* time. The development of these two aspects of our overall algorithm is our key contribution, as the scheduling step allows us to reach an effective solution while simultaneously reducing computational complexity.
- We compare the performance of joint power allocation across all frequency bands with the simpler case in which power constraints are de-coupled across bands. We show that the simpler approach, in fact, suffers little performance loss.
- We show that each iteration of the proposed algorithm leads to nondecreasing objective function values; the overall algorithm outperforms several competing approaches, including the state-of-the-art multicell WMMSE, as well as standard interior-point and sequential quadratic programming solvers widely utilized in the literature, with significantly lower computational complexity.
- Our proposed algorithm outperforms the aforementioned competing state-of-the-art approaches over a wide range of BS maximum transmit power values.

This paper is organized as follows: In Section II, we present our system model and formulate the desired optimization problem. In Section III, we describe the proposed solution approach in detail, while also presenting a proof for its convergence. In Section IV, we present the results and compare the performance and computational complexity of the proposed scheme against the benchmarks described previously. We draw some conclusions in Section V.

Prior to proceeding further, we define some notation used in this paper. $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ represent the set of real numbers, non-negative real numbers and positive numbers respectively. We denote scalars using lowercase (eg. x), vectors using lowercase boldface (eg. $\mathbf{x}$), matrices using uppercase boldface (eg. $\mathbf{X}$) and sets using script typeface (eg. $\mathcal{X}$). The operator $|.|$ denotes the absolute value when applied to a scalar and cardinality when applied to a set; we use $\|.\|_2$ to denote the ℓ_2-norm of a vector. The conjugate of a complex scalar z is denoted by z^*; the Hermitian of a complex vector $\mathbf{z}$ is denoted by $\mathbf{z}^H$. Likewise, $\mathbb{C}$ represents the set of complex numbers. A complex multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu},\mathbf{K})$. Finally, $\mathbf{I}$ represents the identity matrix.

II. System Model and Problem Formulation

We consider a time-duplexed network, with base-stations located in a regular hexagonal pattern; we denote the set of BSs in the network by $\mathcal{B}$. Each user associates with the geographically closest BS, with K_b users associating with the b^{th} base-station; under the hexagonal grid layout of the BSs, this leads to identically sized hexagonal cells. We choose this hexagonal pattern purely for convenience; the derivations and algorithms that follow are applicable to any distribution of BSs in a multi-cell network. We assume that there are a total of F orthogonal frequency bands, each of bandwidth W, available for transmission to each base-station. Each BS is equipped with M transmit antennas which are capable of simultaneously transmitting on all available frequency bands; each user is equipped with a single receive antenna capable of simultaneously receiving signals on all available frequency bands. Furthermore, we also assume that the number of users associated with each BS significantly exceeds the number of transmit antennas available at the base-station (i.e., $K_b \gg M$ for all b). Figure 1 illustrates the system at hand with hexagonal cells.

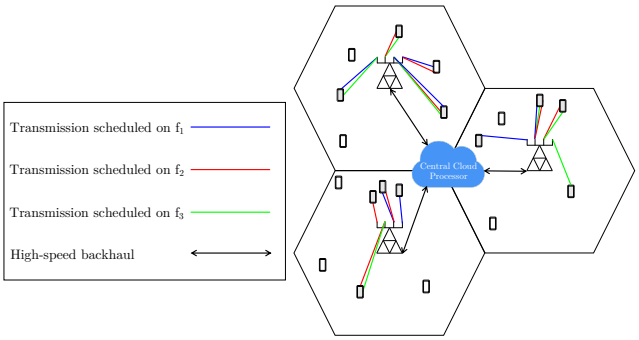

Fig. 1. Network model for the proposed system model. As shown, each BS in the network serves K users within its cell with M transmit antennas, and is thus capable of serving multiple users on each of $F = 3$ orthogonal frequency bands. BSs are connected to a central cloud processor via high-speed backhaul and forward their downlink CSI. This processor computes the scheduling and beamforming strategy and forwards it back to the BSs.

Prior to stating the channel model, it is important to emphasize that this paper focuses on analyzing the best system-level performance. In this regard, we make two important assumptions that are common to the papers in this area [15], [31]. First, we assume that all the BSs have access to perfect CSI of all their associated users on all frequency bands. Second, we assume that all the BSs are connected via high-speed backhaul links to a central cloud server that is capable of performing system-level optimization based on the CSI received from each of the BSs, and relaying back the beamforming weight vectors and scheduling information; thus our system model falls within the general realm of CRAN. This is necessary for a coordinated transmission strategy to achieve the best system-level performance.

The downlink channel from the b'^{th} base-station to the k^{th} user associated with the b^{th} BS on the f^{th} frequency band is a complex $M \times 1$ vector denoted by $\mathbf{h}_{kb,b',f}$. As stated earlier, each BS serves only the set of users associated with it. Accordingly, the beamforming weight vector for the k^{th} user associated with the b^{th} BS on the f^{th} frequency band is a complex $M \times 1$ vector denoted as $\mathbf{v}_{kb,f}$.

In each time slot, each base-station schedules a *subset* of its associated users. Specifically, we impose the constraint that each base-station can schedule no more than M users per time slot on each available frequency band. The binary variable $u_{kb,f}$ is used to indicate whether the k^{th} user associated with the b^{th} BS is scheduled ($u_{kb,f}=1$) or not ($u_{kb,f}=0$) on the f^{th} frequency band. The symbol intended for the k^{th} user associated with the b^{th} BS on the f^{th} frequency band is a complex scalar denoted by $s_{kb,f}$.

It follows that the received downlink signal $r_{kb,f}$ at the k^{th} user associated with the b^{th} base-station on the f^{th} frequency band is given by

$$r_{kb,f}=\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}u_{kb,f}s_{kb,f} + \sum_{\substack{b'=1\\(b',k')\neq}}^{|\mathcal{B}|}\sum_{\substack{k'=1\\(b,k)}}^{K_{b'}}\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}u_{k'b',f}s_{k'b',f}+z_{kb,f} \tag{1}$$

where $z_{kb,f}$ denotes the additive zero-mean Gaussian noise with variance $\sigma_{kb,f}^2$. Thus, the signal-to-interference-plus-noise ratio for the user under consideration on the f^{th} frequency band is given by

$$\gamma_{kb,f}=\frac{u_{kb,f}|\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}|^2}{\sum\limits_{\substack{b'=1\\(b',k')\neq}}^{|\mathcal{B}|}\sum\limits_{\substack{k'=1\\(b,k)}}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2}$$

Consequently, the data rate to the user on the f^{th} frequency band is given by $R_{kb,f}=W\log\left(1+\gamma_{kb,f}\right)$. Recalling that we have a total of F frequency bands available to serve the user, it is clear that the combined data rate to the user in a time slot, denoted by $R_{kb,total}$, is given by $R_{kb,total}=\sum\limits_{f=1}^{F}R_{kb,f}$.

We note that the model presented is general for resource allocation across multiple time-frequency resource blocks. Now, we formulate the resource allocation problem for the downlink of multiantenna networks. To do so, our choice of the network sum utility function is the network WSR.

While equal weights focuses on the sum-rate, it has been shown in [33] that maximizing the weighted sum rate (using an appropriate choice of weights in each time slot) leads to maximization of the sum of the logarithm of the long-term average data rates achieved by each of the users. This in turn leads to a proportionally fair allocation of resources amongst all users in the network.

We are interested in answering the following question: given a network as described above, to maximize the WSR in each time slot which subset of users should each BS serve, in which frequency band, at what power level and with what beamformer design? The optimization problem that encapsulates this question for a single time slot can be expressed as

$$\underset{\mathbf{U},\mathbf{V},\boldsymbol{\Gamma}}{\text{maximize}} \quad \sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}w_{kb}\log\left(1+\gamma_{kb,f}\right) \tag{2a}$$

$$\text{subject to} \quad u_{kb,f}\in\{0,1\}, \quad \begin{matrix}b=1,\ldots,|\mathcal{B}|\,;\\ k=1,\ldots,K_b;\\ f=1,\ldots,F\end{matrix} \tag{2b}$$

$$\sum_{k=1}^{K_b}u_{kb,f}\leq M, \quad \begin{matrix}b=1,\ldots,|\mathcal{B}|\,;\\ f=1,\ldots,F\end{matrix} \tag{2c}$$

$$\sum_{f=1}^{F}\sum_{k=1}^{K_b}\|\mathbf{v}_{kb,f}\|_2^2\leq FP_T, \qquad b=1,\ldots,\ |\mathcal{B}| \tag{2d}$$

$$\gamma_{kb,f}=\frac{u_{kb,f}|\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}|^2}{\sum\limits_{\substack{b'=1\\(b',k')\neq}}^{|\mathcal{B}|}\sum\limits_{\substack{k'=1\\(b,k)}}^{K_{b'}}u_{k'b'}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2} \tag{2e}$$

$$\begin{matrix}b=1,\ldots,|\mathcal{B}|\,;\\ k=1,\ldots,K_b;\\ f=1,\ldots,F\end{matrix}$$

Here, w_{kb} represents the weight for the k^{th} user associated with the b^{th} base-station[2], while $\mathbf{U}$ and $\mathbf{V}$ denote the optimization variables gathered into a matrix: the scheduling variables ($\mathbf{U}$) and the beamformers ($\mathbf{V}$). For simplicity of notation, we drop the index denoting the time slot in the formulation of the optimization problem in (2). The SINR of the k^{th} user associated with the b^{th} BS on the f^{th} frequency band is denoted by $\gamma_{kb,f}$.

Our objective function is the network WSR for a single time slot as expressed in (2a). The constraint in (2b) enforces the scheduling decisions by the BS to be binary; a user is scheduled if its scheduling variable equals one, and vice versa. The second set of constraints in (2c) ensures that a BS is restricted to serving a subset of at most M users from the set of all associated users on each available frequency band. The constraints in (2d) impose a total transmit power constraint FP_T at each BS across the different frequency bands (and thus an average power constraint of P_T across each individual frequency band). Finally, the equality constraints in (2e) enforce the SINR values at each user, BS and frequency band.

III. Proposed Approach

We note that the optimization problem in (2) has a mixed-binary integer form and is nonconvex in the beamforming variables. In fact, as stated earlier, the general WSR maximization problem has been shown to be NP-hard by Luo and Zhang in [6]. We also note that the beamforming and

[2]The proportionally fair weight for the n^{th} time slot is usually determined by finding the inverse of the long-term average data rate achieved by the user in question over an exponentially decaying window [33], i.e., $w_{kb}^{(n)}=1/\bar{R}_{kb}^{(n)}$, where $\bar{R}_{kb}^{(n)}$ represents the exponentially weighted average data rate achieved by the user in the time slots preceding the n^{th} time slot across all frequency bands. This is calculated using the update equation $\bar{R}_{kb}^{(n)}=(1-\alpha)\bar{R}_{kb}^{(n-1)}+\alpha R_{kb}^{(n)}$, where $R_{kb}^{(n)}$ represents the total data rate achieved in the n^{th} time slot across all bands, and α represents the forgetting factor.

scheduling variables are coupled across the different frequency bands due to the sum-power constraint in (2d)), even though the bands themselves are orthogonal and thus noninterfering. To solve this problem and obtain an effective solution, we adopt an iterative optimization strategy, based on fractional programming as developed by Shen and Yu in [17] and [18].

We begin by introducing Lagrange multipliers for each of the equality constraints in (2e), in a similar fashion to [18], as follows

$$\mathcal{L}(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{\Lambda})=\sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}\Bigg[w_{kb}\log\left(1+\gamma_{kb,f}\right)-\lambda_{kb,f}\cdot \left(\gamma_{kb,f}-\frac{u_{kb,f}|\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}|^2}{\sum\limits_{\substack{b'=1\\(b',k')\neq(b,k)}}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2}\right)\Bigg] \tag{3}$$

where the $\lambda_{kb,f}$ represent the Lagrange multipliers for each of the equality constraints in (2e). For notational clarity, the SINR auxiliary variables and Lagrange multipliers are collected in the matrices $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ respectively. Consequently, in order to satisfy the first-order condition with respect to the $\gamma_{kb,f}$ values, we set the partial derivative with respect to the Lagrangian equal to zero, i.e.,

$$\frac{\partial\mathcal{L}(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{\Lambda})}{\partial\gamma_{kb,f}}=0. \tag{4}$$

Now substituting (4) into the equality constraint (2e), we then obtain the optimal Lagrange multipliers as

$$\lambda_{kb,f,opt}=\frac{w_{kb}\left(\sum\limits_{\substack{b'=1\\(b',k')\neq(b,k)}}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2\right)}{\sum\limits_{b'=1}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2} \tag{5}$$

Substituting the optimal Lagrange multipliers from (5) into the expression for the Lagrangian in (3), we obtain the following reformulated objective function, which we denote by $f_r(\mathbf{U},\mathbf{V},\mathbf{\Gamma})$.

$$\begin{aligned}f_r(\mathbf{U},\mathbf{V},\mathbf{\Gamma})=&\sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}w_{kb}\log\left(1+\gamma_{kb,f}\right)\\&-\sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}w_{kb}\gamma_{kb,f}\\&+\sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}\frac{w_{kb}(1+\gamma_{kb,f})u_{kb,f}|\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}|^2}{\sum\limits_{b'=1}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2}\end{aligned} \tag{6}$$

Thus, it follows that the original optimization problem in (2) can be expressed as the following reformulated optimization problem

$$\underset{\mathbf{U},\mathbf{V},\mathbf{\Gamma}}{\text{maximize}} \quad f_r(\mathbf{U},\mathbf{V},\mathbf{\Gamma}) \tag{7a}$$

$$\text{subject to} \quad u_{kb,f}\in\{0,1\},\ \begin{matrix}b=1,\ldots,|\mathcal{B}|;\\k=1,\ldots,K_b;\\f=1,\ldots,F\end{matrix} \tag{7b}$$

$$\sum_{k=1}^{K_b}u_{kb,f}\leq M,\quad \begin{matrix}b=1,\ldots,|\mathcal{B}|;\\f=1,\ldots,F\end{matrix} \tag{7c}$$

$$\sum_{f=1}^{F}\sum_{k=1}^{K_b}\|\mathbf{v}_{kb,f}\|_2^2\leq FP_T,\qquad b=1,\ldots,\ |\mathcal{B}| \tag{7d}$$

(7e)

We note that the reformulated optimization problem in (7) is equivalent to the original optimization problem in (2) in the sense that the optimal objective function value and the associated primal optimization variables, $\mathbf{U}$ and $\mathbf{V}$, for both problems are identical.

A key point to note before we proceed further is that the ratio terms which were present inside the logarithm function in problem (2) have now been moved outside as the sum-of-ratios term in the reformulated objective function. This is the first step in allowing us to develop an iterative optimization strategy. In order to proceed further, we make use of the following theorem, as a vector-valued version of that derived by Shen and Yu in [18]:

Theorem 1. *Let $n_i(\mathbf{x})$: $\mathbb{C}^m\longmapsto\mathbb{C}$ and $d_i(\mathbf{x})$: $\mathbb{C}^m\longmapsto\mathbb{R}_{++}$, where $i=1,\ldots,N$ be two functions of the optimization variables $\mathbf{x}$ and $m\in\mathbb{N}$. Furthermore, let $\mathcal{X}\subseteq\mathbb{C}^m$ be a constraint set. Then the sum-of-ratios optimization problem*

$$\begin{aligned}\underset{\mathbf{x}}{\text{maximize}}\quad&\sum_{i=1}^{N}\frac{|n_i(\mathbf{x})|^2}{d_i(\mathbf{x})}\\\text{subject to}\quad&\mathbf{x}\in\mathcal{X}\end{aligned} \tag{8}$$

is equivalent to the following reformulated optimization problem

$$\begin{aligned}\underset{\mathbf{x},\mathbf{y}}{\text{maximize}}\quad&\sum_{i=1}^{N}\left[2\text{Re}\{y_i^*n_i(\mathbf{x})\}-|y_i|^2d_i(\mathbf{x})\right]\\\text{subject to}\quad&\mathbf{x}\in\mathcal{X},\ \mathbf{y}\in\mathbb{C}^N\end{aligned} \tag{9}$$

in the sense that the optimal values of the objective function and primal optimization variables are identical. Note that the vector $\mathbf{y}=[y_1\ y_2\ \cdots\ y_N]$ is an auxiliary variable.

Proof. We first observe that the reformulated objective function in (9) is concave in $\mathbf{y}$. Thus, setting the partial derivative of this objective with respect to y_i^*, we obtain

$$y_{i,opt}=\frac{n_i(\mathbf{x})}{d_i(\mathbf{x})}$$

Substituting these values back into the objective function in (9) yields the objective function in (8). □

Applying Theorem 1 to the sum-of-ratios term in $f_r(\mathbf{U},\mathbf{V},\mathbf{\Gamma})$ in (6), we obtain the following new objective

function

$$f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}) = \sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F}\Big(w_{kb}\left[\log\left(1+\gamma_{kb,f}\right)-\gamma_{kb,f}\right] - \sum_{b'=1}^{|\mathcal{B}|}\sum_{k'=1}^{K_{b'}} u_{k'b',f}\left|y_{kb}\right|^2\left(\left|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}\right|^2+\sigma_{kb,f}^2\right)\Big) + \sum_{b=1}^{|\mathcal{B}|}\sum_{k=1}^{K_b}\sum_{f=1}^{F} 2Re\{y_{kb,f}^*\sqrt{w_{kb}(1+\gamma_{kb,f})}u_{kb,f}\mathbf{v}_{kb,f}^H\mathbf{h}_{kb,b,f}\} \tag{10}$$

and accordingly, (7) can be expressed as the equivalent optimization problem below

$$\underset{\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}}{\text{maximize}} \quad f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}) \tag{11a}$$

$$\text{subject to} \quad u_{kb,f}\in\{0,1\}, \quad \begin{matrix} b=1,\ldots,|\mathcal{B}|\,; \\ k=1,\ldots,K_b; \\ f=1,\ldots,F \end{matrix} \tag{11b}$$

$$\sum_{k=1}^{K_b} u_{kb,f} \le M, \quad \begin{matrix} b=1,\ldots,|\mathcal{B}|\,; \\ f=1,\ldots,F \end{matrix} \tag{11c}$$

$$\sum_{f=1}^{F}\sum_{k=1}^{K_b}\|\mathbf{v}_{kb,f}\|_2^2 \le FP_T, \quad b=1,\ldots,|\mathcal{B}| \tag{11d}$$

Once again, to avoid excessive notational clutter, we collect the $y_{kb,f}$ values in the matrix $\mathbf{Y}$.

From Theorem 1, it follows that the optimization problem in (11) is also equivalent to the original optimization problem in (2) in the sense that the optimal objective function and the associated primal optimization variables, $\mathbf{U}$ and $\mathbf{V}$, for both problems are identical.

We emphasize that both of the reformulated problems in (7) and (11) remain nonconvex and NP-hard (like the original problem) since our reformulation steps result in equivalent problems. Crucially, however, the new objective functions are now in a form amenable to an iterative optimization strategy leading to an effective solution to our original optimization problem in (2).

The continuous variables: To develop the iterative approach, we first observe that for fixed $\mathbf{U}, \mathbf{V}$ and $\mathbf{Y}$, the optimal $\mathbf{\Gamma}$ can be found by setting

$$\frac{\partial f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y})}{\partial\gamma_{kb,f}}=0$$

$$\Rightarrow\gamma_{kb,f,opt}=\frac{u_{kb,f}|\mathbf{h}_{kb,b,f}^H\mathbf{v}_{kb,f}|^2}{\sum\limits_{\substack{b'=1\\(b',k')\neq(b,k)}}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2} \tag{12}$$

as $f_r\left(\mathbf{U},\mathbf{V},\mathbf{\Gamma}\right)$ is concave in $\mathbf{\Gamma}$. Next, we note that holding the $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Gamma}$ values fixed, $f_q\left(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}\right)$ is concave in $\mathbf{Y}$, so the optimal $\mathbf{Y}$ values can be found by setting

$$\frac{\partial f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y})}{\partial y_{kb,f}^*}=0$$

$$\Rightarrow y_{kb,f,opt}=\frac{\sqrt{w_{kb}(1+\gamma_{kb,f})}u_{kb,f}\mathbf{v}_{kb,f}^H\mathbf{h}_{kb,b,f}}{\sum\limits_{b'=1}^{|\mathcal{B}|}\sum\limits_{k'=1}^{K_{b'}}u_{k'b',f}|\mathbf{h}_{kb,b',f}^H\mathbf{v}_{k'b',f}|^2+\sigma_{kb,f}^2} \tag{13}$$

In a similar fashion, when $\mathbf{U}$, $\mathbf{\Gamma}$ and $\mathbf{Y}$ are fixed, we can find the optimal $\mathbf{V}$ values (i.e., the beamforming weight vectors). Note that due to the sum-power constraint (11d), taking the derivative of $f_q\left(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}\right)$ directly with respect to $\mathbf{V}$ to find the optimal beamforming weight vectors is not valid. To simplify the derivation, we recall that with these variables fixed, we can write the problem of finding the optimal beamforming weight vectors as:

$$\underset{\mathbf{V}}{\text{maximize}} \quad f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y}) \tag{14a}$$

$$\text{subject to} \quad \sum_{f=1}^{F}\sum_{k=1}^{K_b}\|\mathbf{v}_{kb,f}\|_2^2 \le FP_T, \quad b=1,\ldots,|\mathcal{B}| \tag{14b}$$

We note that this optimization problem is concave and thus readily solvable; in fact, we can derive a closed-form expression for the weights by introducing Lagrange multipliers μ_b for the sum-power constraint at each BS in (14b). This yields the Lagrangian $\hat{\mathcal{L}}\left(\mathbf{V},\boldsymbol{\mu}\right)$ (not to be confused with the Lagrangian $\mathcal{L}(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{\Lambda})$ we derived earlier) as:

$$\hat{\mathcal{L}}\left(\mathbf{V},\boldsymbol{\mu}\right)=f_q(\mathbf{U},\mathbf{V},\mathbf{\Gamma},\mathbf{Y})-\sum_{b=1}^{|\mathcal{B}|}\mu_b\left(\sum_{f=1}^{F}\sum_{k=1}^{K_b}\|\mathbf{v}_{kb,f}\|_2^2-FP_T\right) \tag{15}$$

where, to keep our notation uncluttered, we collect the multipliers μ_b in the vector $\boldsymbol{\mu}$. Then the first-order optimality condition of $\hat{\mathcal{L}}\left(\mathbf{V},\boldsymbol{\mu}\right)$ with respect to each $\mathbf{v}_{kb,f}$ yields:

$$\frac{\partial\hat{\mathcal{L}}\left(\mathbf{V},\boldsymbol{\mu}\right)}{\partial\mathbf{v}_{kb,f}}=\mathbf{0}$$

$$\Rightarrow\mathbf{v}_{kb,f,opt}=\sqrt{w_{kb}(1+\gamma_{kb,f})}u_{kb,f}y_{kb,f}^*\cdot\left(\sum_{b'=1}^{B}\sum_{k'=1}^{K_b}u_{kb,f}\left|y_{k'b',f}\right|^2\mathbf{h}_{k'b',b,f}\mathbf{h}_{k'b',b'f}^H+\mu_b\mathbf{I}_M\right)^{-1}\mathbf{h}_{kb,b,f} \tag{16}$$

The dual variable μ_b should be chosen to satisfy complementary slackness in the total power constraint at BS b; observing (16), it is clear that the magnitude of $\mathbf{v}_{kb,f}$ is a decreasing function of μ_b. Thus, we can obtain μ_b easily through a bisection search, which in turn can be used to obtain the optimal beamforming weight vectors.

At this juncture, we note an important point for future reference: the beamforming step involves inversion of a $M\times M$ matrix *for each user*, which is computationally costly, especially when we have to perform it for a large number of users. Furthermore, if the scheduling variable for a user is zero, the beamforming weight for that user is automatically zero; there is no need to perform any computation in this case.

The binary scheduling variables: The final step in the iterative approach is to optimize the user scheduling variables $\mathbf{U}$ when the continuous variables, $\mathbf{\Gamma}$, $\mathbf{V}$ and $\mathbf{Y}$, are held fixed. To do so, we first observe that since the frequency bands are assumed to be orthogonal, the scheduling decisions are decoupled across the different frequency bands. Furthermore, we make use of

an intuitive yet powerful insight first suggested in [34] and also observed in [15]: provided that the beamforming weight vectors $\mathbf{V}$ are held fixed, the interference value experienced by a user in the downlink scheduled on a particular frequency band *depends only on the beam used to serve that user and remains fixed regardless of which other users are scheduled on the remaining beams*. We note that this is different from the uplink setting, in which the interference pattern in the network changes when a new set of users is scheduled on a given set of beams.

The fact that the interference pattern changes in the uplink creates significant challenges in terms of scheduling: as the authors of [18] observe, even with beamforming weights *fixed*, the problem of optimal scheduling remains NP-hard. This substantially affects the quality of solutions obtained since, as emphasized earlier, we do not know which users are suitable to schedule *a priori*. One solution, as mentioned previously, is to schedule all users in the network and transform the original problem to an unconstrained problem in terms of scheduling; however, since we need a matrix inversion per user, this results in an undesirable increase in computational complexity to levels identical with the WMMSE algorithm [18]. In the downlink, however, we are not bound by this constraint, i.e., changing scheduling decisions does not affect the interference pattern. Thus, we can schedule only a subset of users in the entire network, reducing complexity as only the beamforming weights for a small number of users need to be calculated.

To illustrate this point, let us consider the b^{th} BS in a multicell network. The scheduling decisions for different frequency bands are decoupled; thus, we consider the f^{th} frequency band without loss of generality. In addition to this, each BS can schedule at most M users in a single time slot, whereas it has $K_b{>}M$ users associated with it. Suppose the BS is serving users on the indicated frequency band using a fixed set of $N_{b,f}{\leq}M$ nonzero beamforming weights which we denote by $\check{\mathcal{V}}_{b,f}$, i.e.

$$\check{\mathcal{V}}_{b,f} = \{\check{\mathbf{v}}_{nb,f}{\in}\mathbb{C}^M |\ n = 1,\ldots,N_{b,f};\ \check{\mathbf{v}}_{nb,f}{\neq}\mathbf{0}\}$$

It follows that the k^{th} user associated with this BS can find itself in one of two scenarios with regards to the given frequency band: either it is scheduled for transmission on the one of the $N_{b,f}$ beams from the set $\check{\mathcal{V}}_{b,f}$ or it is not being served by the BS. For the former setting, suppose the user is scheduled on the n^{th} nonzero beam; then the power received on this beam is the signal power. If the user is not scheduled, however, all the power received is interference power. For notational convenience, let us denote by $\zeta_{kb,f}$ the combined received signal, interference and noise power by the user in question, i.e.

$$\zeta_{kb,f}{=}\sum_{b'=1}^{|\mathcal{B}|}\sum_{k'=1}^{K} u_{k'b',f}\left|\mathbf{h}^{\mathrm{H}}_{kb,b',f}\mathbf{v}_{k'b',f}\right|^2{+}\sigma^2_{kb,f} \tag{17}$$

Then the total interference power received by this user $I_{kb,f}$ is given by

$$I_{kb,f} = \begin{cases} \zeta_{kb,f} & \text{if } u_{kb,f} = 0 \\ \zeta_{kb,f} - \left|\mathbf{h}^H_{kb,b,f}\check{\mathbf{v}}_{nb,f}\right|^2 & \text{if } u_{kb,f} = 1 \end{cases}$$

Importantly, if this user is scheduled on the n^{th} nonzero beam, the interference power it experiences does not depend on which users are scheduled on the remaining $N_{b,f}{-}1$ beams within its own cell. The same holds true when the user in question is not scheduled on any beam; the interference power experienced by the user is the same regardless of which set of users in its own cell is scheduled on the given set of beams. In addition, we also make the following critical observation: for the b^{th} BS, changing the set of users scheduled on the fixed set of beams does not change the interference pattern experienced by users *outside* its own cell. This can be seen from the fact that in (17), the inter-cell interference power received by user k associated with BS b from BS b' on frequency band f depends only on the interference channel $\mathbf{h}_{kb,b',f}$, and the beamforming weight vectors $\mathbf{v}_{k'b',f}$ can be permuted over any of the users served by BS b' without affecting $\zeta_{kb,f}$. Meanwhile, the intracell interference power recieved by user k associated with BS b on frequency band f depends only upon the information-bearing channel $\mathbf{h}_{kb,b,f}$ and the beamforming weight the user is scheduled on $\mathbf{v}_{k'b',f}$. Taken together, these observations imply that if the beamforming weights throughout the network are held fixed, we can *locally* optimize the scheduling at each BS in order to maximize the *network-wide* sum weighted rate *for the given set of beamforming weights*.

Accordingly, we can formulate a strategy to help us find the best set of $N_{b,f}$ users to be served by the b^{th} BS on its set of nonzero beamforming weights $\check{\mathcal{V}}_{b,f}$ for the f^{th} frequency band. In other words, our goal is to find the set of $N_{b,f}$ users out of the K_b total users in the cell that will yield the maximum weighted sum rate on the given set of beamforming weight vectors. Our choice of users should satisfy the constraint that a user can only be served on a single beam by a BS in keeping with our original system model. A greedy strategy of assigning the user capable of achieving the highest weighted rate on each beam is not guaranteed to solve the combinatorial problem of selecting the best subset of $N_{b,f}$ users of the K_b available.

Our first step in matching the users to the fixed beams to maximize the WSR is to define the $K_b{\times}N_{b,f}$ combined weighted rates matrix $\hat{\mathbf{R}}_{b,f}$ for the given set of beams and users on the f^{th} frequency band. The $(i,j)^{\text{th}}$ entry in this matrix, denoted by $\widehat{r}_{ib,j,f}$, indicates the weighted rate that would be achieved by the i^{th} user if it is scheduled on the j^{th} nonzero beam on the f^{th} frequency band, i.e.

$$\left[\hat{\mathbf{R}}_{b,f}\right]_{ij} = \widehat{r}_{ib,j,f} = w_{ib}\log\left(1{+}\frac{\left|\mathbf{h}^H_{ib,b,f}\check{\mathbf{v}}_{jb,f}\right|^2}{\zeta_{ib,f}{-}\left|\mathbf{h}^H_{ib,b,f}\check{\mathbf{v}}_{jb,f}\right|^2}\right)$$

Note that we compute the total interference received by every user in the network, regardless of whether it is scheduled or not; thus, every user is considered eligible for possible scheduling on a non-zero beamforming weight.

It follows that our goal of scheduling the users on the appropriate beams can be formulated as the following binary

integer optimization problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{maximize}} \quad & \sum_{k=1}^{K_b}\sum_{n=1}^{N_{b,f}} \widehat{r}_{kb,n,f} x_{kb,n,f} \\ \text{subject to} \quad & \sum_{k=1}^{K_b} x_{kb,n,f} = 1, \qquad n=1,\ldots,N_{b,f} \\ & \sum_{n=1}^{N_{b,f}} x_{kb,n,f} \leq 1, \qquad k=1,\ldots,K_b \\ & x_{kb,n,f}\in\{0,1\}, \qquad \begin{array}{l} n=1,\ldots,N_{b,f}; \\ k=1,\ldots,K_b \end{array} \end{aligned} \tag{18}$$

where the binary variables $x_{kb,n,f}$ indicate whether or not the k^{th} user is scheduled on the n^{th} nonzero beam by the b^{th} BS on the f^{th} frequency band. The objective function maximizes the WSR. The first constraint requires at least one user to be scheduled on every beam while the second ensures that a user is scheduled in one beam only.

Note that these binary variables are not the same as the optimization variables $u_{kb,f}$ which refer to whether the user is scheduled or not. These variables have the additional index n which denotes the n^{th} beam. Specifically, $u_{kb,f} = \sum_{n=1}^{N_{b,f}} x_{kb,n,f}$, i.e., if the user is scheduled on any one of the beams associated with the BS, it is scheduled by that BS.

The problem in (18) is, in fact, a linear sum assignment problem, and can also be viewed as a maximum weighted bipartite matching problem, which has been extensively studied in the literature and can be solved in polynomial time using techniques like the Hungarian algorithm [35] (more formally known as the Kuhn-Munkres algorithm) or the auction algorithm [36]. Specifically, using the Hungarian algorithm to solve the linear sum assignment problem for an $m\times n$ matrix, where $m>n$, has a complexity of $\mathcal{O}(n^2m)$ [35]. In our case, we have $N_{b,f}<K_b$; hence the complexity of solving problem (18) is $\mathcal{O}(N_{b,f}^2 K_b)$. Solving this optimization problem for each BS allows us to *optimally* schedule the users to maximize the network weighted sum rate on the fixed set of beamforming weight vectors. We remark that this scheduling scheme using fractional programming and Hungarian algorithm is different from the uplink setting [18], where the scheduling of one user would have changed the interference pattern; consequently the only way to solve the uplink problem to global optimality is by extensive search as mentioned earlier.

This scheduling setup also reduces complexity as compared to the unconstrained scheduling setting, as we now only have to calculate the beamforming weight vectors for a maximum of M rather than K_b users at each iteration of the algorithm. Importantly, this set of M users can change from iteration to iteration as the beamforming weights are matched to the best set of users; this is unlike the uplink setting where a user not scheduled during the initialization remains unscheduled throughout all subsequent iterations of fractional programming algorithm [18].

With a fixed set of beamforming weight vectors, therefore, the proposed scheduling scheme finds a per-cell optimal selection of scheduling decisions; hence, from iteration to iteration, the user assignment to beamforming weight vectors is set according to which combination yields the greatest network WSR. Here, we only consider scheduling each user on a single beam per frequency band for two reasons: first, this is in keeping with our original system model in which each user is scheduled on a maximum of one data stream per frequency band and the standard assumption for coordinated resource allocation algorithms including multicell WMMSE [10] and uplink fractional programming [18], and second, the framework of the Hungarian algorithm does not allow for scheduling on more than a single beamforming weight vector.

Combining all these steps together, the proposed technique for coordinated resource allocation in the downlink of multiuser MISO networks is summarized in Algorithm 1. The algorithm optimizes one of the optimization variables keeping the others fixed, iterating till convergence.

Algorithm 1 Coordinated Resource Allocation for Downlink of LS-MIMO Networks

1: Initialize $\mathbf{U}$, $\mathbf{V}$, $N_{iterations}$.
2: Set $i = 1$.
3: **repeat**
4: Update $\mathbf{\Gamma}$ using (12).
5: Update $\mathbf{Y}$ using (13).
6: Update $\mathbf{V}$ using (16).
7: Update $\mathbf{U}$ and $\mathbf{V}$ jointly by solving (18) for each BS on each frequency band.
8: Increment i.
9: **until** convergence or $i = N_{iterations}$.

Theorem 2. *The proposed algorithm described in Algorithm 1 is non-decreasing in the objective function $f_0(\mathbf{U},\mathbf{V})$ after each iteration.*

Proof: We refer to the objective function in (2a) as $f_0(\mathbf{U},\mathbf{V})$. The non-decreasing convergence can be proven by considering the following chain of reasoning going from iteration i to $i+1$:

$$\begin{aligned} f_0(\mathbf{U}^{(i)},\mathbf{V}^{(i)}) &= f_r(\mathbf{U}^{(i)},\mathbf{V}^{(i)},\mathbf{\Gamma}^{(i)}) && \text{(19a)} \\ &\leq f_r(\mathbf{U}^{(i)},\mathbf{V}^{(i)},\mathbf{\Gamma}^{(i+1)}) && \text{(19b)} \\ &= f_q(\mathbf{U}^{(i)},\mathbf{V}^{(i)},\mathbf{\Gamma}^{(i+1)},\mathbf{Y}^{(i)}) && \text{(19c)} \\ &\leq f_q(\mathbf{U}^{(i)},\mathbf{V}^{(i)},\mathbf{\Gamma}^{(i+1)},\mathbf{Y}^{(i+1)}) && \text{(19d)} \\ &\leq f_q(\mathbf{U}^{(i)},\mathbf{V}^{(i+1)},\mathbf{\Gamma}^{(i+1)},\mathbf{Y}^{(i+1)}) && \text{(19e)} \\ &\leq f_q(\mathbf{U}^{(i+1)},\tilde{\mathbf{V}}^{(i+1)},\mathbf{\Gamma}^{(i+1)},\mathbf{Y}^{(i+1)}) && \text{(19f)} \\ &= f_r\left(\mathbf{U}^{(i+1)},\tilde{\mathbf{V}}^{(i+1)},\mathbf{\Gamma}^{(i+1)}\right) && \text{(19g)} \\ &= f_0(\mathbf{U}^{(i+1)},\tilde{\mathbf{V}}^{(i+1)}) && \text{(19h)} \end{aligned}$$

where (19a) follows from the fact that the reformulated objective function f_r equals the original when the optimal $\mathbf{\Gamma}$ values are substituted; (19b) follows from the fact that the update of $\mathbf{\Gamma}$ when all other variables are fixed maximizes f_r; (19c) follows from Theorem 1; (19d) follows from the fact that the update of $\mathbf{Y}$ when all other variables are fixed maximizes f_q; (19e) follows from the fact that the update of $\mathbf{V}$ when all other variables are fixed maximizes f_q; (19f) follows from the

fact that the joint update of $\mathbf{U}$ and $\mathbf{V}$ using (18) maximizes f_q when all other variables are fixed; (19g) follows from Theorem 1; and (19h) from similar reasoning to (19a). Note that we use $\tilde{\mathbf{V}}$ to denote the set of permuted beamforming weights obtained from $\mathbf{V}$ by solving (18). ■

Coupled with the fact that the objective function has a finite maximum, we can state that the algorithm converges. However, since the scheduling variables are binary, we cannot call this a local optimum. Furthermore, we observe that the proposed scheme is not exactly a block coordinate ascent scheme, since we use the partial derivative of f_r in (12); nonetheless, the algorithm converges in a non-decreasing fashion to an effective solution of the original WSR maximization problem as described above.

IV. Performance Evaluation of Proposed Scheme

In order to evaluate the performance of the proposed algorithm, we compare it with the following different coordinated and uncoordinated resource allocation schemes:

1) *Matched filtering transmission with equal power allocation and round-robin scheduling*: This is the simplest uncoordinated resource allocation strategy which can be implemented and as such it provides a useful benchmark with which to compare the performance of the multiuser algorithm.
2) *Zero-forcing with equal power allocation and round-robin scheduling*: Zero-forcing eliminates intracell interference and thus provides improved performance compared to matched-filtering. Zero-forcing does involve increased computational complexity compared to matched filtering, requiring a matrix inversion to determine the beamforming weight for each of the scheduled users.
3) *WMMSE with greedy scheduling*: The WMMSE algorithm has been well studied in the literature as a coordinated beamforming scheme; adaptive power allocation is implicitly included in the beamformer design. However, as noted in [18], WMMSE is intended for use as a beamforming algorithm; the question of which users to schedule remains to be answered. Accordingly, we adopt the greedy proportionally fair scheduling scheme introduced in [15].
   In this scheduling scheme, we first initialize the algorithm with a random set of M users. The beamforming weights are held fixed and we sequentially determine which user will maximize the weighted rate on each beamforming weight from the BS, i.e., the k^{th} user associated with the b^{th} BS is scheduled on the j^{th} beam on the f^{th} frequency band if
   $$k = \underset{i=1,\ldots K}{\arg\max}\ \hat{r}_{ib,j,f}$$
   This approach to scheduling is distinct from solving the linear sum assignment problem in the proposed algorithm, as the users are selected greedily for each beam rather than jointly across the set of all given beams.
   Importantly, this algorithm is not necessarily monotonically non-decreasing.
4) *Multicell WMMSE*: In multicell WMMSE [10], each BS initializes the algorithm by simultaneously scheduling *all the users* in the network. The idea is to let the algorithm iterate on the beamformer design for all these users; eventually the beamforming weights for the majority of users will converge to zero, and these users are then implicitly not scheduled by the BS. This multicell WMMSE scheme is the state-of-the-art in the literature and has the same convergence properties as our algorithm. However, this comes at a cost: in order to determine the beamforming weights for each user, WMMSE performs a matrix inversion and bisection search. With the multicell WMMSE scheme, since all the users in the network are scheduled, the number of matrix inversions becomes extremely high. This is especially inefficient as the number of users ultimately assigned beamforming weights with nonzero power is very small. As we will show in the analysis of the results, our proposed algorithm is capable of outperforming multicell WMMSE, while simultaneously providing significant savings in computational complexity.

We consider a network partitioned into identical hexagonal cells, with BSs located at the center of each cell. The users are distributed randomly with uniform density over the entire network area. Furthermore, we also assume that the number of users associated with each BS significantly exceeds the number of transmit antennas available at the BS (i.e., $K_b \gg M$ for all b).

To compare the performance of the aforementioned resource allocation schemes, we simulate a 7-cell network with wraparound. To ensure a fair comparison, all iterative optimization schemes were run for 15 iterations. The rest of the simulation parameters are as listed in Table I.

We begin by comparing the performance of the proposed algorithm against the benchmark schemes listed, as well as against standard interior-point and sequential quadratic programming algorithms utilized in the literature [37], [38]. The latter were implemented using standard available optimization software; all users were scheduled in a similar fashion to the multicell WMMSE algorithm. Figure 2 shows the convergence of network sum-rate for M=2 transmit antennas and K_b=5 users per cell. As we can observe, the proposed algorithm achieves a higher network sum-rate, converging smoothly in a monotonically non-decreasing fashion; this is as expected from Theorem 2. At the same time, it is also clear that the uncoordinated resource allocation strategies perform substantially worse than the coordinated strategies, with matched

TABLE I
THE NUMERICAL VALUES OF PARAMETERS USED IN THE SYSTEM MODEL.

Total bandwidth	W = 20 MHz
BS maximum transmit power per frequency band	P_T = 43 dBm
Noise figure	N_f = 9 dB
Path-loss exponent	$\alpha = 3.76$
Reference distance	0.3920 m

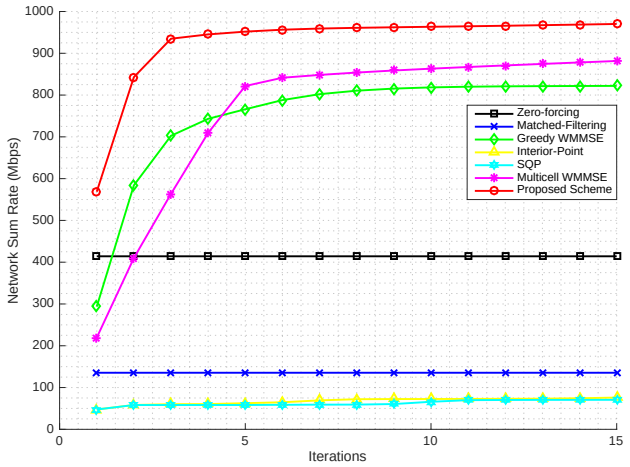

Fig. 2. Convergence of Network Sum Rate for Different Resource Allocation Schemes for M=2, K_b=5.

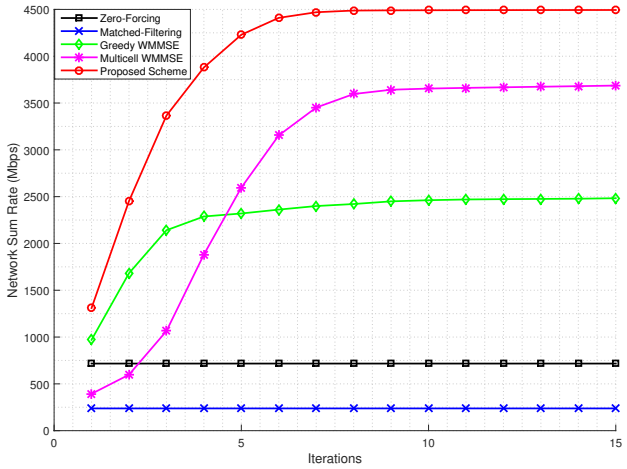

Fig. 3. Convergence of Network Sum Rate for Different Resource Allocation Schemes for M=8, K_b=80.

filtering being the worst beamforming strategy. In addition, we observe that there is a gap in performance between greedy and multicell WMMSE. This is readily understood since the greedy scheduling approach is *not* guaranteed to increase the network WSR after the scheduling reassignment. The multicell WMMSE algorithm is guaranteed to converge in a monotonically non-decreasing fashion, and as such provides good performance. Nevertheless, it is outperformed by the proposed scheme, which converges to the highest network weighted sum rate of all resource allocation schemes. In contrast, the sequential quadratic programming and interior-point algorithms show improving objective values as the number of iterations increase; however, they provide the worst performance. Due to the highly non-convex nature of the WSR-max optimization problem, these methods have demonstrated inferior performance in prior works in this area [18], [39]; hence these results are not unexpected.

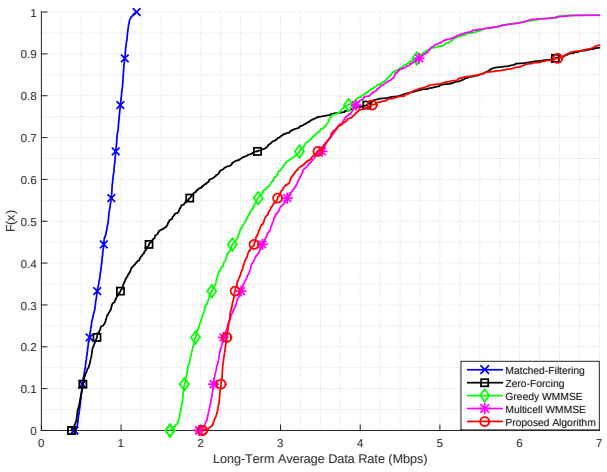

Fig. 4. CDFs of average data rates achieved with different resource allocation schemes for a fixed number of iterations.

Convergence: Figure 3 shows the sum-rate convergence of the various resource allocation schemes for a single time slot with identical channel sets but with a much larger network size of $M = 8$ and $K_b = 80$. We observe similar trends to those in Figure 2, with two notable differences. First, the greedy WMMSE algorithm is substantially outperformed by the multicell WMMSE algorithm, and the performance gap between the proposed scheme and the benchmarks grows larger as well. Secondly, due to the large number of optimization variables involved, the SQP and interior-point algorithms failed to converge for this setting. In particular, for the SQP approach, taking a direct step involves computing the Hessian of the optimization variables, which is extremely computationally demanding for a problem of this given size [37].

It is important to emphasize that none of the schemes result in the globally optimal solution and, if the number of potential users to be scheduled is very large, the WMMSE algorithm can get stuck in a poor solution and often takes longer to converge. In contrast, the proposed scheme restricts the BS to serve at most M users, thereby narrowing the pool of potential users and ensuring faster convergence to a higher-quality local optimum.

PF Rates: In Figure 3, we present the cumulative distribution functions (CDFs) of the long-term user average data rates achieved with the different resource allocation schemes (with $N_{iterations} = 15$, $M = 8$ and $K_b = 80$). In order to compare these, we consider two metrics: the sum of the logarithm of the long-term average data rates (in megabits per second) and the 10^{th} percentile user rates, which are logged in Table II. We choose to maximize the WSR in each time slot when the weights are chosen according to the proportionally fair metric described earlier (with the forgetting factor α chosen as 0.05), as this leads to maximization of the sum of the logarithm of the average data rates achieved by the users. Thus, comparing the average sum-log utility of the different resource allocation schemes allows us to directly compare them in terms of our original objective. We note that comparing the absolute

TABLE II
NETWORK SUM-LOG UTILITY AND EDGE USER RATES FOR DIFFERENT RESOURCE ALLOCATION SCHEMES WITH FIXED NUMBER OF ITERATIONS.

Resource Allocation Strategy	Average Network Sum-Log Utility	Edge User Rate (Mbps)
Matched-Filtering	-203	0.51
Zero-Forcing	446	0.51
Greedy WMMSE	821	1.78
Multicell WMMSE	908	2.15
Proposed Algorithm	952	2.25

difference (rather than the relative gain) in the sum-log-utility of different schemes illustrates the improvements made to the average rates achieved by the users.

Comparing the 10^{th} percentile user rates allows us to compare the quality of service for the cell-edge users for the different resource allocation schemes. It is important to note, however, that the algorithms do not optimize the cell-edge rate; comparing edge user rates merely allows us to understand the quality of service that these different resource allocation schemes provide to the lower-percentile users in the network.

As we can observe from Figure 3 and Table II, the uncoordinated resource allocation schemes have the worst performance in terms of both the average sum-log-utility and edge user rates. This is unsurprising, since the benefits of coordinated resource allocation schemes over uncoordinated schemes are well-known. We note that employing zero-forcing results in significantly higher average sum log utility than matched filtering; this is also to be expected, since the former scheme eliminates intracell interference. All three coordinated resource allocation schemes achieve significantly higher performance than the uncoordinated schemes. However, of the two WMMSE resource allocation schemes, greedy scheduling has the worst performance; this is because greedy scheduling is sub-optimal and the associated algorithm is not guaranteed to be monotonic.

Utilizing multicell WMMSE results in a further significant gain to both the average sum-log-utility and the edge user rates. The proposed scheme performs even better than the multicell WMMSE scheme with considerably higher sum-log-utility and slightly better edge rates. The majority of the performance gain comes at higher percentiles, where the proposed approach achieves much better data rates than multicell WMMSE. Indeed, compared to the uncoordinated resource allocation schemes, there is a fourfold increase in the 10^{th} percentiles, with an increase of almost 30% compared to the greedy WMMSE scheme. The sum-log-utility of the proposed scheme is also considerably higher than that achieved by the greedy WMMSE scheme.

A key point related to Table II is that we initialize the proposed algorithm by scheduling the set of users that achieves the highest interference-free weighted sum rate with an equal power allocation. For a worst-case initialization (i.e., if we start with the set of users that achieves the lowest interference-free weighted sum rate), the sum-log utility function is 909 for the proposed algorithm, still higher than that achieved by the multicell WMMSE algorithm. We emphasize that there is no known optimal initialization for the coordinated resource allocation schemes.

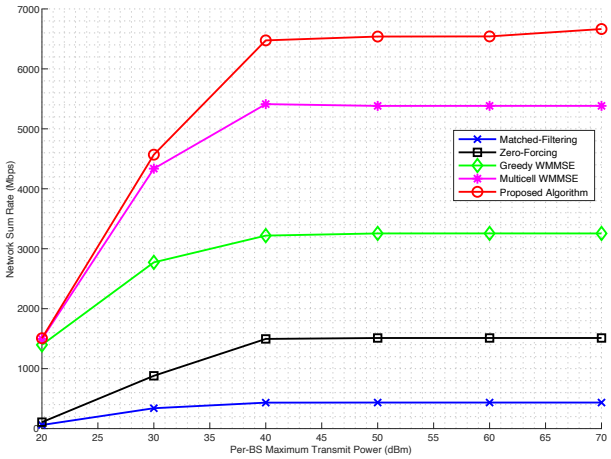

Fig. 5. Network sum-rate as a function of per-BS transmit power for different resource allocation schemes.

Sum Rate: To change the optimization objective function, we compare the performance of the aforementioned resource allocation schemes in terms of network sum-rate when the BS maximum transmit power, P_T is varied from 20 to 70 dBm for $M = 8$ and $K_b = 80$. When maximizing the sum-rate, all users' weights are set to unity and this assignment does not change across time-slots. The corresponding results are shown in Figure 5.

As the figure clearly shows, the proposed algorithm substantially outperforms the competing WMMSE approaches. In particular, there is a gap of approximately 22% in network sum-rate across transmit powers above 40dBm between the proposed approach and the benchmark multicell WMMSE algorithm. Both algorithms substantially outperform the greedy WMMSE algorithm, delivering sum-rates that are more than twice as high for large BS transmit powers. As expected, the performance the matched-filtering and zero-forcing schemes lags behind those of the coordinated schemes.

This result in particular demonstrates the massive performance advantage our optimal scheduling approach based on the Hungarian algorithm delivers over greedy scheduling. One interesting phenomenon to note is that the network-sum rate of the proposed algorithm strictly increases as a function of BS transmit power; however, this is not always the case for the multicell WMMSE algorithm. This result highlights the tendency of the multicell WMMSE algorithm to get stuck in low-quality local optima.

Optimization Across Frequency Bands: Finally, we compare the performance of the joint power allocation approach versus the per-band power allocation method. In the former setting, we assign a total power of FP_T to be distributed over the F frequency bands at each BS. This means that the BS is free to use as much or as little power in each of the frequency bands, provided that the total power consumed across all frequency bands is less than FP_T. In contrast, with the per-band power

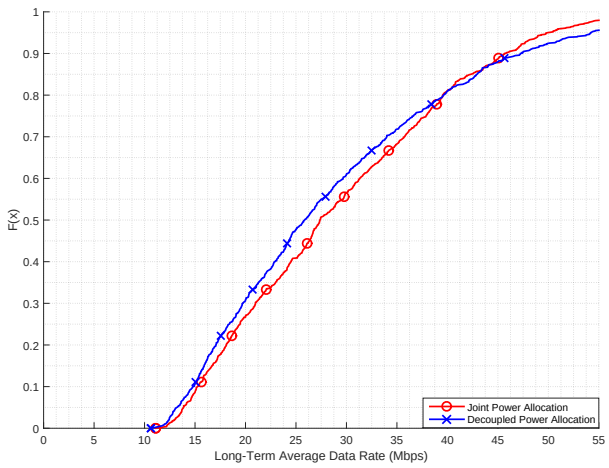

Fig. 6. CDFs of average data rates achieved with joint and decoupled power allocation schemes.

allocation strategy, each BS can only utilize a maximum power of P_T per individual frequency band. As we can observe in Fig. 6; there is no significant performance benefit in terms of either edge rates or overall utility to choosing the joint power allocation strategy over the per-band strategy. Furthermore, the number of iterations to calculate the beamforming weights is more than the decoupled setting, since the bisection search step is now being performed across all frequency bands.

A. Complexity Analysis

In comparing the performance of these various resource allocation strategies, a critical point is the computational complexity involved. From [10], the computational complexity of the beamforming step in WMMSE can be derived as $\mathcal{O}(\kappa^2 M^2 + \kappa M^3)$, where κ represents the total number of users scheduled in the network. For our setting, we have $|\mathcal{B}|$ cooperating cells. For simplicity of analysis, we assume that each cell has K users associated with it; thus we have $\kappa = K\,|\mathcal{B}|$ for the multicell WMMSE scheme and $\kappa = M\,|\mathcal{B}|$ for the greedy WMMSE scheme and proposed.

Accordingly, the computational complexity of the WMMSE algorithm with greedy scheduling can be found as $\mathcal{O}(M^4|\mathcal{B}|^2 + M^2K|\mathcal{B}|^2)$ whereas the computational complexity of the multicell WMMSE algorithm is $\mathcal{O}(M^3K\,|\mathcal{B}| + M^2K^2|\mathcal{B}|^2)$. The proposed algorithm has the same computational complexity as the greedy WMMSE scheme (since we schedule at most M users in a single time slot). It follows that the computational complexity of the fractional programming strategy is at most given by $\mathcal{O}(M^4|\mathcal{B}|^2 + M^2K|\mathcal{B}|^2)$. A comparison of the per-iteration computational complexity of the various resource allocation schemes discussed is provided in Table III.

To understand these results, we revisit some of the assumptions made earlier in the system model. Since we deal

[3]Note that these uncoordinated schemes require only a single iteration to determine the network resource allocation strategy.

TABLE III
PER-ITERATION COMPUTATIONAL COMPLEXITY OF DIFFERENT RESOURCE ALLOCATION SCHEMES.

Resource Allocation Strategy	Complexity Per Iteration
Matched-Filtering[3]	$\mathcal{O}(M^2\,\lvert\mathcal{B}\rvert\,)$
Zero-Forcing[3]	$\mathcal{O}(M^4\,\lvert\mathcal{B}\rvert\,)$
Greedy WMMSE	$\mathcal{O}(M^4\lvert\mathcal{B}\rvert + M^2K\lvert\mathcal{B}\rvert^2)$
Multicell WMMSE	$\mathcal{O}(M^3K\lvert\mathcal{B}\rvert + M^2K^2\lvert\mathcal{B}\rvert^2)$
Proposed Algorithm	$\mathcal{O}(M^4\lvert\mathcal{B}\rvert + M^2K\lvert\mathcal{B}\rvert^2)$

with a large-scale MIMO system, the number of users in each cell (K) is assumed to be significantly larger than the number of antennas at the BS. Critically, as K becomes asymptotically large, for multicell WMMSE, the $M^2K^2|\mathcal{B}|^2$ term will dominate the complexity expression. On the other hand, for the proposed algorithm, the only term dependent upon the number of users per cell is $M^2K|\mathcal{B}|^2$; thus, if M is fixed and $K \gg M$ as per the assumption in our system model earlier, then the $M^4|\mathcal{B}|^2$ term becomes negligible in comparison. It follows that in this case, the complexity of the multicell WMMSE algorithm will be roughly K/M times higher than the proposed algorithm. Our algorithm achieves this noteworthy reduction in complexity while exceeding the performance of the multicell WMMSE algorithm.

It is worth noting that even with large computation resources in a CRAN, a fully centralized globally optimal solution is infeasible since the problem at hand is NP-hard. Furthermore, even in this case, the reduced computational complexity of our algorithm, compared to say the multicell WMMSE approach, is important for implementation with a large number of users and BS antennas. We emphasize that, furthermore, this gain in computational complexity is accompanied by improved performance. Compared to the generic solvers we have considered, our proposed algorithm is more convenient from an implementation perspective since the updates for the optimization variables are expressed in closed-form. This is also in contrast to globally optimal strategies such as outer polyblock approximation, in which the updates are not expressible in closed-form [24]. We also note that the SQP and interior-point algorithms do not have closed-form updates; calculating Hessians for the latter approach in particular is computationally taxing for large network sizes [37].

Prior to proceeding further, we consider the reason for this behavior in greater detail. Recall that in the multicell WMMSE scheme, we schedule all users simultaneously and let the beamformer design iterate. Thus, this is equivalent to considering the original optimization problem but with *no scheduling constraints*. As the algorithm converges, most users are assigned beamformers with zero power; the number of users ultimately assigned nonzero power is very close to M. Nonetheless, the beamforming weights still have to be calculated for the users who will ultimately be dropped since they are not known *a priori*. This requires computationally costly matrix inversions, thus leading to a higher overall complexity. With the proposed algorithm, since we schedule the best M

TABLE IV
AVERAGE EXECUTION TIME OF DIFFERENT RESOURCE ALLOCATION SCHEMES.

Resource Allocation Scheme	$M = 2$, $K_b = 5$	$M = 8$, $K_b = 80$
Matched-Filtering	0.01	1.1
Zero-Forcing	0.05	1.2
Greedy WMMSE	1.3	27.2
Multicell WMMSE	1.3	48.8
Proposed Algorithm	1.3	19.5
Interior-Point	40.6	N/A
SQP	53.6	N/A

users in a single time slot, the number of matrix inversions required is identical to the greedy WMMSE strategy.

A second consideration is that the set of users scheduled in each iteration of the algorithm has the potential to change. Unlike the greedy WMMSE strategy, however, the proposed scheme ensures that the network WSR increases after each scheduling step as we find the *best network-wide scheduling pattern* for the fixed set of beamforming weight vectors. Hence, we conclude that the proposed strategy of intelligently scheduling the smaller set of users (which is close to the number of users implicitly scheduled by the multicell WMMSE scheme) nets an improvement in terms of computational complexity while providing superior performance. Also, it is worth pointing out that scheduling all users, as the multicell WMMSE algorithm does, requires a much greater overhead in terms of communication between the coordinated BSs in the network.

Finally, we consider the actual execution time of the various resource allocation schemes. Although the complexity analysis provides a formal characterization of how the running time of each resource allocation scheme scales as the network parameters change, it is nonetheless useful for us to compare the average execution time of each scheme. To ensure a fair comparison, we measure the time taken from initialization until the per-iteration increase in the network weighted sum rate is less than 10% for the given time slot. For the simulation parameters detailed in Table I, the average execution times on the desktop computer used to generate these results are logged in Table IV. As we can see, the uncoordinated schemes require a much lower execution time on average than the coordinated schemes; however, this comes at the expense of compromised performance as discussed earlier. Both the proposed algorithm and greedy WMMSE approach perform similarly in terms of average execution time. The multicell WMMSE scheme has the highest average execution time among coordinated schemes; as discussed earlier, this is due to the fact that all users in the network are scheduled simultaneously, so the number of matrix inversions needed is very large compared to both greedy WMMSE and the proposed algorithm. Finally, the interior-point and SQP algorithms have extremely long execution times for the $M = 2$ and $K_b = 5$ setting and do not converge within a reasonable time for the $M = 8$ and $K_b = 80$ setting.

[3]These execution times are obtained using a desktop computer with a 3.6 GHz Intel® Core™ i7-4790 CPU and 24 GB of RAM.

V. CONCLUSIONS

In this paper, we developed a coordinated resource allocation scheme for the downlink of multiuser MIMO networks with multiple orthogonal frequency bands. The proposed scheme outperforms uncoordinated schemes like zero-forcing and matched-filtering, as well as the coordinated greedy and state-of-the-art multicell WMMSE schemes in terms of the average sum-log-utility function and network sum-rate. Furthermore, the proposed scheme offers significant computational complexity savings over the state-of-the-art multicell WMMSE scheme and also has a much lower average execution time. By intelligently scheduling the best subset of M users for a fixed given set of beamforming weights, the proposed approach is able to reduce the computational complexity as well as providing a higher weighted sum-rate in a single time slot and higher long-term average sum-log-utility. Thus, we conclude that the proposed approach offers an effective high-performance and low-complexity solution to the nonconvex NP-hard weighted sum-rate maximization problem.

REFERENCES

[1] A. A. Khan, R. Adve, and W. Yu, "Optimizing Multicell Scheduling and Beamforming via Fractional Programming and Hungarian Algorithm," in *IEEE Globecom Workshops (GC Wkshps)*, Abu Dhabi, Dec. 2018, pp. 1–6.

[2] H. Bolcskei, "MIMO-OFDM Wireless Systems: Basics, Perspectives, and Challenges," *IEEE Wireless Commun.*, vol. 13, no. 4, pp. 31–37, Aug. 2006.

[3] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[4] E. Björnson, E. Jorswieck *et al.*, "Optimal Resource Allocation in Coordinated Multi-Cell Systems," *Found. Trends Commun. Inf. Theory*, vol. 9, no. 2–3, pp. 113–381, 2013.

[5] H. Dahrouj and W. Yu, "Coordinated Beamforming for the Multicell Multi-Antenna Wireless System," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[6] Z.-Q. Luo and S. Zhang, "Dynamic Spectrum Management: Complexity and Duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[7] L. Liu, R. Zhang, and K.-C. Chua, "Achieving Global Optimality for Weighted Sum-Rate Maximization in the K-user Gaussian Interference Channel with Multiple Antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1933–1945, May 2012.

[8] C. Suh, M. Ho, and D. N. Tse, "Downlink Interference Alignment," *IEEE Trans. Commun.*, vol. 59, no. 9, pp. 2616–2626, Sep. 2011.

[9] K. Hosseini, C. Zhu, A. A. Khan, R. S. Adve, and W. Yu, "Optimizing the MIMO Cellular Downlink: Multiplexing, Diversity, or Interference Nulling?" *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6068–6080, Dec. 2018.

[10] 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 Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[11] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving Sustainable Ultra-dense Heterogeneous Networks for 5G," *IEEE Comm. Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[12] P. C. Weeraddana, M. Codreanu, M. Latva-aho, A. Ephremides, C. Fischione *et al.*, "Weighted Sum-Rate Maximization in Wireless Networks: A Review," *Found. Trends Netw.*, vol. 6, no. 1–2, pp. 1–163, Oct. 2012.

[13] D. Park, "Iterative Waterfilling With User Selection in Gaussian MIMO Broadcast Channels," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 1902–1911, Jan. 2018.

[14] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Coordinated Scheduling and Power Control in Cloud-Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2523–2536, Dec. 2016.

[15] W. Yu, T. Kwon, and C. Shin, "Multicell Coordination via Joint Scheduling, Beamforming, and Power Spectrum Adaptation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 1–14, June 2013.

[16] S. Chinnadurai, P. Selvaprabhu, X. Jiang, H. Hai, and M. H. Lee, "Worst-Case Weighted Sum-Rate Maximization in Multicell Massive MIMO Downlink System for 5G Communications," *Physical Commun.*, vol. 27, pp. 116–124, Apr. 2018.

[17] K. Shen and W. Yu, "Fractional Programming for Communication Systems-Part I: Power Control and Beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, Mar. 2018.

[18] ——, "Fractional Programming for Communication Systems-Part II: Uplink Scheduling via Matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, Mar. 2018.

[19] J. Kaleva, A. Tölli, and M. Juntti, "Decentralized Sum Rate Maximization with QoS Constraints for Interfering Broadcast Channel via Successive Convex Approximation," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2788–2802, Feb. 2016.

[20] X. Li, S. You, L. Chen, A. Liu, and Y. E. Liu, "A New Algorithm for the Weighted Sum Rate Maximization in MIMO Interference Networks," in *IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, Mar. 2018, pp. 147–152.

[21] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Joint Scheduling and Beamforming via Cloud-Radio Access Networks Coordination," in *IEEE Vehicular Tech. Conf. (VTC)*, Chicago, IL, Aug. 2018, pp. 1–5, Aug. 2018, pp. 1–5.

[22] C. Zhang, Y. Huang, Y. Jing, S. Jin, and L. Yang, "Sum-Rate Analysis for Massive MIMO Downlink With Joint Statistical Beamforming and User Scheduling," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2181–2194, Jan. 2017.

[23] J. Brehmer, *Utility Maximization in Nonconvex Wireless Systems*. Springer Science & Business Media, 2012, vol. 5.

[24] W. Utschick and J. Brehmer, "Monotonic Optimization Framework for Coordinated Beamforming in Multicell Networks," *IEEE Trans. Signal Proc.*, vol. 60, no. 4, pp. 1899–1909, 2012.

[25] K. P. Roshandeh, M. Ardakani, and C. Tellambura, "Exact Solutions for Certain Weighted Sum-Rate and Common-Rate Maximization Problems," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1026–1029, Feb. 2018.

[26] T. Q. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud radio access networks: Principles, technologies, and applications*. Cambridge University Press, 2017.

[27] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent Advances in Cloud Radio Access Networks: System Architectures, key Techniques, and Open Issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 2016.

[28] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, 2014.

[29] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband Processing Units Virtualization for Cloud Radio Access Networks," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 189–192, 2015.

[30] M. Gerasimenko, D. Moltchanov, R. Florea, S. Andreev, Y. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar, "Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks," *IEEE Access*, vol. 3, pp. 397–406, 2015.

[31] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, "Optimization of MIMO Device-to-Device Networks via Matrix Fractional Programming: A Minorization-Maximization Approach," Aug. 2018. [Online]. Available: https://arxiv.org/pdf/1808.05678.pdf

[32] Y. Zhang, J. An, K. Yang, X. Gao, and J. Wu, "Energy-Efficient User Scheduling and Power Control for Multi-Cell OFDMA Networks Based on Channel Distribution Information," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5848–5861, 2018.

[33] W. Yu, T. Kwon, C. Shin, and V. K. Bhargava, *Adaptive resource allocation in cooperative cellular networks*. Cambridge University Press, 2011, pp. 233—-258.

[34] A. L. Stolyar and H. Viswanathan, "Self-Organizing Dynamic Fractional Frequency Reuse for Best-Effort Traffic Through Distributed Inter-Cell Coordination," in *IEEE INFOCOM*, Rio de Janeiro, Apr. 2009, pp. 1287–1295.

[35] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*. Springer Science & Business Media, 2012, vol. 2.

[36] D. P. Bertsekas, "The auction algorithm for assignment and other network flow problems: A tutorial," *Interfaces*, vol. 20, no. 4, pp. 133–149, Aug. 1990.

[37] J. Nocedal and S. Wright, *Numerical Optimization*. Springer Science & Business Media, 2006.

[38] R. Fletcher, *Practical Methods of Optimization*. John Wiley & Sons, 1987.

[39] K. Chitti, Q. Kuang, and J. Speidel, "Joint Base Station Association and Power Allocation for Uplink Sum-Rate Maximization," in *IEEE SPAWC*, Darmstadt, Jun. 2013, pp. 6–10.