Traffic Aware Resource Allocation Schemes for Multi-Cell MIMO-OFDM Systems

Ganesh Venkatraman *Student Member, IEEE*, Antti Tölli *Senior Member, IEEE*, Markku Juntti *Senior Member, IEEE*, and Le-Nam Tran *Member, IEEE*

***Abstract*—We consider a downlink multi-cell multiple-input multiple-output (MIMO) interference broadcast channel (IBC) using orthogonal frequency division multiplexing (OFDM) with multiple users contending for space-frequency resources in a given scheduling instant. The problem is to design precoders efficiently to minimize the number of backlogged packets queuing in the coordinating base stations (BSs). Conventionally, the queue weighted sum rate maximization (Q-WSRM) formulation with the number of backlogged packets as the corresponding weights is used to design the precoders. In contrast, we propose joint space-frequency resource allocation (JSFRA) formulation, in which the precoders are designed jointly across the space-frequency resources for all users by minimizing the total number of backlogged packets in each transmission instant, thereby performing user scheduling implicitly. Since the problem is nonconvex, we use the combination of successive convex approximation (SCA) and alternating optimization (AO) to handle nonconvex constraints in the JSFRA formulation. In the first method, we approximate the signal-to-interference-plus-noise ratio (SINR) by convex relaxations, while in the second approach, the equivalence between the SINR and the mean squared error (MSE) is exploited. We then discuss the distributed approaches for the centralized algorithms using primal decomposition and alternating directions method of multipliers. Finally, we propose a more practical iterative precoder design by solving the Karush-Kuhn-Tucker expressions for the MSE reformulation that requires minimal information exchange for each update. Numerical results are used to compare the proposed algorithms to the existing solutions.**

***Index Terms*—Convex approximations, MIMO-IBC, MIMO-OFDM, precoder design, SCA, WSRM.**

I. Introduction

In a network with multiple base stations (BSs) serving multiple users, the main driving factor for the transmission is the packets waiting at each BS corresponding to the different users existing in the network. We consider the problem of transmit precoder design over the space-frequency resources provided by the multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) framework in the downlink interference broadcast channel (IBC) to minimize the number of queued packets in the BSs. Since the resources are shared by multiple users associated with different BSs, the problem of interest can be viewed as a resource allocation one.

This work has been supported by the Finnish Funding Agency for Innovation (Tekes), Nokia Networks, Xilinx, Elektrobit and the Academy of Finland. Part of this work is presented in ICASSP 2014.

G. Venkatraman, A. Tölli and M. Juntti are with Centre for Wireless Communications (CWC), Department of Communications Engineering (DCE), University of Oulu, Oulu, FI-90014, (e-mail: {ganesh.venkatraman, antti.tolli, markku.juntti}@ee.oulu.fi)

L.-N. Tran was with Centre for Wireless Communications (CWC), Oulu, FI-90014. He is now with the Department of Electronic Engineering, Maynooth University, Maynooth, Co. Kildare, Ireland, (e-mail:ltran@eeng.nuim.ie)

In general, the resource allocation problems such as admission control ones can be formulated by assigning a binary variable for each user to indicate the presence or absence of a particular resource [1]. Alternatively, linear transmit precoders, which are complex vectors, can be implicitly modeled as decision variables, thereby avoiding the use of binary decision variables. After the design stage, the non-zero precoders are used to determine the transmission rates of users on a space-frequency resource, and the zero transmit precoder indicates the absence of the user on a given resource. In this way, the soft decisions are used in the optimization problem and the hard decisions are made after the algorithm convergence.

The queue minimizing precoder designs are closely related to the weighted sum rate maximization (WSRM) problem with additional rate constraints to limit the throughput beyond the number of backlogged packets associated with the users. The topics on MIMO IBC precoder design have been studied extensively with different performance criteria in the literature. Due to the nonconvex nature of the MIMO IBC precoder design problems, successive convex approximation (SCA) approach has become a powerful tool to solve these problems. For example, in [2], the nonconvex part of the objective is linearized around a fixed operating point to solve the WSRM problem in a iterative manner. A similar approach using arithmetic-geometric inequality was proposed in [3].

The relation between the achievable sum rate and the mean squared error (MSE) of the received symbol by using fixed minimum mean squared error (MMSE) receivers can also be used to solve the WSRM problem [4]. In [5], [6], the WSRM problem is reformulated via MSE, casting the problem as a convex one for fixed linearization coefficients. In this way, the original problem is expressed in terms of the MSE weight, precoders, and receivers. Then the problem is solved using an alternating optimization method, i.e., finding a subset of variables while the other variables are fixed. The MSE reformulation for the WSRM problem was also studied in [7] by using SCA to solve the problem in an iterative manner. Moreover, distributed precoder designs with quality of service (QoS) requirements as additional rate constraints are studied for the MSE reformulated WSRM problem in [8].

The problem of precoder design for the MIMO IBC system can be solved either by using a centralized controller or by using decentralized algorithms, where each BS handles the corresponding subproblem independently with the limited information exchange with other BSs via back-haul. The distributed approaches are usually based on the primal, or dual decompositions or the alternating directions method of

multipliers (ADMM), as discussed in [9], [10]. In the primal decomposition, the so-called coupling interference variables are fixed for the subproblem at each BS to find the optimal precoders. The fixed interference is then updated using the subgradients as discussed in [11]. The dual and the ADMM approaches control the distributed subproblems by fixing the *'interference price'* for each BS as detailed in [12].

By adjusting the weights in the WSRM objective, we can find an arbitrary rate-tuple in the rate region that maximizes suitable objective measures. For example, if the weight of each user is set to be inversely proportional to its average data rate, the corresponding problem guarantees fairness on the average among the users. To reduce the number of backlogged packets, we can assign weights based on the current queue size of the users. Specifically, the queue states can be incorporated in the WSRM objective $\sum_k w_k R_k$ by replacing the weight w_k with the corresponding queue state Q_k or its function, which is the outcome of minimizing the Lyapunov drift between the current and the future queue states [13], where R_k denotes the achievable data rate of user k. In the *backpressure algorithm*, the differential queues between the source and the receiver nodes are used to scale the transmission rate [13].

Earlier studies on the queue minimization problem are summarized in the survey papers [14], [15]. In particular, the problem of power allocation to minimize the number of backlogged packets was considered in [16] using geometric programming. Since the problem addressed in [16] assumed single antenna transmitters and receivers, the queue minimizing problem reduces to the optimal power allocation problem. In the context of wireless networks, the *backpressure algorithm* mentioned above was extended in [17] by formulating the corresponding user queues as the weights in the WSRM problem. Recently, the precoder design for the video transmission over MIMO system was considered in [18]. In this design, the MU-MIMO precoders are designed by the MSE reformulation as in [5] with the higher layer performance objectives such as playback interruptions and buffer overflow probabilities.

Main Contributions: In this paper, we design precoders jointly over space-frequency resources to reduce the number of backlogged packets waiting at each BSs. The proposed formulation also limits the allocations beyond the number of backlogged packets without explicit rate constraints. Initially, we propose a centralized joint space-frequency resource allocation (JSFRA) formulation, which is solved by two iterative algorithms based on the combination of SCA and alternating optimization (AO) due to the nonconvex nature of the problem. The proposed algorithms solve a sequence of convex problems obtained by fixing a subset of optimization variables or by approximating the nonconvex constraints by the convex ones. The first approach is performed by directly relaxing the signal-to-interference-plus-noise ratio (SINR) expression, while in the second method, the equivalence between the MSE and the SINR is exploited. We then discuss the distributed implementation of the JSFRA methods using primal decomposition and the ADMM. Finally, we also propose a more practical iterative precoder design by directly solving the Karush-Kuhn-Tucker (KKT) system of equations for the MSE reformulation that is numerically shown to require minimal information exchange for each update. Note that the joint space-frequency channel matrix can be formed by stacking the channel of each sub-channel in a block-diagonal form for all users.

The rest of the paper is organized as follows. Section II introduces the system model and the problem formulation. The existing and the proposed centralized designs are presented in Section III. The distributed solutions for the proposed problem are provided in Section IV followed by the simulation results in Section V. Finally, conclusions are drawn in Section VI.

II. System Model and Problem Formulation

A. System Model

We consider a downlink MIMO IBC scenario in an OFDM framework with N sub-channels and N_B BSs each equipped with N_T transmit antennas, serving in total K users each with N_R receive antennas. The set of users associated with BS b is denoted by $\mathcal{U}_b$ and the set $\mathcal{U}$ represents all users in the system, i.e., $\mathcal{U} = \underset{b\in\mathcal{B}}{\cup}\mathcal{U}_b$, where $\mathcal{B}$ is the set of indices of all coordinating BSs. Data for user k is transmitted from only one BS which is denoted by $b_k \in \mathcal{B}$. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of all sub-channel indices available in the system.

We adopt linear transmit beamforming technique at BSs. Specifically, the data symbols $d_{l,k,n}$ for user k on the lth spatial stream over sub-channel n is multiplied with beamformer $\mathbf{m}_{l,k,n} \in \mathbb{C}^{N_T\times 1}$ before being transmitted. In order to detect multiple spatial streams at the user terminal, receive beamforming vector $\mathbf{w}_{l,k,n}$ is employed for each user. Consequently, the received data estimate corresponding to the lth spatial stream over sub-channel n at user k is given by

$$\hat{d}_{l,k,n} = \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}d_{l,k,n} + \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{n}_{l,k,n} + \mathbf{w}_{l,k,n}^{\mathrm{H}}\sum_{i\in\mathcal{U}\setminus\{k\}}\mathbf{H}_{b_i,k,n}\sum_{j=1}^{L}\mathbf{m}_{j,i,n}d_{j,i,n} \quad (1)$$

where $\mathbf{H}_{b,k,n} \in \mathbb{C}^{N_R\times N_T}$ is the channel between BS b and user k on sub-channel n, and $\mathbf{n}_{l,k,n} \sim \mathcal{CN}(0, N_0)$ is the additive noise vector for user k on the nth sub-channel and lth spatial stream. In (1), $L = \mathrm{rank}(\mathbf{H}_{b,k,n}) = \min(N_T, N_R)$ is the maximum number of spatial streams.[1] Assuming independent detection of data streams, we can write the signal-to-interference-plus-noise ratio (SINR) as

$$\gamma_{l,k,n} = \frac{\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\,\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}\right|^2}{\grave{N}_0 + \sum_{(j,i)\neq(l,k)}|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2} \quad (2)$$

where $\grave{N}_0 = N_0\,\mathrm{tr}(\mathbf{w}_{l,k,n}\mathbf{w}_{l,k,n}^{\mathrm{H}})$ denotes the equivalent noise variance. In order to avoid the overhead involved in feeding back the user channels, we consider time division duplexing (TDD) system in which BSs can estimate the downlink channels from uplink pilots by using channel reciprocity property.

Let Q_k be the number of backlogged packets destined for user k at a given scheduling instant. The queue dynamics of user k are modeled using the Poisson arrival process with the

[1]It can be easily extended for user specific streams L_k instead of using common L streams for all users. L streams are initialized but after solving the problem, only $L_{k,n} \leq L$ non-zero data streams are transmitted.

average number of packet arrivals of $A_k = \mathbf{E}_i\{\lambda_k\}$ packets or bits, where $\lambda_k(i) \sim \text{Pois}(A_k)$ represents the instantaneous number of packets arriving for user k at the ith time instant.[2] The total number of queued packets at the $(i+1)$th instant for user k, denoted as $Q_k(i+1)$, is given by

$$Q_k(i+1) = \Big[Q_k(i) - t_k(i)\Big]^+ + \lambda_k(i) \tag{3}$$

where $[x]^+ \equiv \max\{x, 0\}$ and t_k denotes the number of transmitted packets or bits for user k. At the ith instant, the transmission rate of user k is given by

$$t_k(i) = \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}(i) \tag{4}$$

where $t_{l,k,n}$ denotes the number of transmitted packets or bits over the lth spatial stream on the nth sub-channel. The maximum rate achieved over the space-frequency resource (l,n) is given by $t_{l,k,n} \leq \log_2(1+\gamma_{l,k,n})$ for the SINR $\gamma_{l,k,n}$.[3] Note that t_k and Q_k are represented by the same units, *i.e.*, in bits defined per channel use.

B. Problem Formulation

To minimize the total number of backlogged packets, we consider minimizing the weighted ℓ_q-norm of the queue deviation objective as

$$v_k = Q_k - t_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n}) \tag{5}$$

where $\gamma_{l,k,n}$ is given by (2) and the optimization variables are the transmit precoders $\mathbf{m}_{l,k,n}$ and the receivers $\mathbf{w}_{l,k,n}$.

Explicitly, the objective of the problem considered is given as $\sum_{k\in\mathcal{U}} a_k|v_k|^q$. Thus the formulation becomes

$$\underset{\mathbf{m}_{l,k,n},\mathbf{w}_{l,k,n}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{6a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\text{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\text{H}}\right) \leq P_{\max}\forall b \tag{6b}$$

where $\tilde{v}_k \triangleq a_k^{1/q}v_k$ is the element of vector $\tilde{\mathbf{v}}$, and a_k is the weighting factor, which is used to alter the user priority based on the QoS constraints such as packet delay requirements and packet waiting time, since they are proportional to the corresponding number of backlogged packets. The BS specific power constraint for all sub-channels is considered in (6b).

For practical reasons, we impose a constraint on the maximum number of transmitted bits for the user k, since it is limited by the total number of backlogged packets available at the transmitter. As a result, the number of backlogged packets v_k for user k remaining in the system is given by

$$v_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n}) \geq 0. \tag{7}$$

The above positivity constraint needs to be satisfied by v_k to avoid the excessive allocation of the resources.

[2]The unit can either be packets or bits as long as the arrival and the transmission units are similar.

[3]The upper bound can be achieved by using Gaussian signaling.

Before proceeding further, we show that the constraint in (7) is handled implicitly by the definition of ℓ_q norm in the objective of (6). Suppose that $t_k > Q_k$ for a certain k at the optimum, i.e., $-v_k = t_k - Q_k > 0$. Then there exists $\delta_k > 0$ such that $-v_k' = t_k' - Q_k < -v_k$ where $t_k' = t_k - \delta_k$. Since $\|\tilde{\mathbf{v}}\|_q = \||\tilde{\mathbf{v}}|\|_q = \||-\tilde{\mathbf{v}}|\|_q$, this means that the newly created vector $\mathbf{t}'$ achieves a strictly smaller objective which contradicts with the fact that the optimal solution has been obtained. The choice of ℓ_q norm used in the objective function [14], [16] alters the priorities for the queue deviation function as follows.

- ℓ_1 norm results in greedy allocation *i.e*, emptying the queue of users with good channel states before considering the users with worse channel conditions. As a special case, it is easy to see that (6) reduces to the WSRM problem when the queue size is large enough for all users.
- ℓ_2 norm prioritizes users with a higher number of queued packets before considering the users with a smaller number of backlogged packets. For example, it could be more ideal for the delay limited scenario when the packet arrival rates of the users are similar, since the number of backlogged packets is proportional to the delay in the transmission following the Little's law [13].
- ℓ_∞ norm minimizes the maximum number of queued packets among users with the current transmission, thereby providing queue fairness by allocating resources proportional to the number of backlogged packets.

III. Proposed Queue Minimizing Precoder Designs

In general, the precoder design for the MIMO OFDM problem is difficult due to its nonconvex nature. In addition, the objective of minimizing the number of the queued packets over space-frequency dimensions adds further complexity. Since the scheduling of users in each sub-channel is achieved by allocating zero transmit power over certain sub-channels, our solutions perform joint precoder design and user scheduling. Before discussing the proposed solutions, we consider an existing algorithm to minimize the number of backlogged packets with additional constraints required by the problem.

A. Queue Weighted Sum Rate Maximization (Q-WSRM)

The queue minimizing algorithms have been studied extensively in the networking context to provide congestion-free routing between any two nodes in the network. One such is the *backpressure algorithm* [13]. It finds an optimal control policy in the form of rate or resource allocation by considering differential backlogged packets between any two entities.

The queue weighted sum rate maximization (Q-WSRM) formulation extends the *backpressure algorithm* to the downlink MIMO-OFDM framework, in which the BSs act as the source nodes and the users as the receivers. The control policy in the form of transmit precoders aims at minimizing the number of queued packets waiting in the BSs. To find an optimal strategy, we resort to the Lyapunov theory, which is predominantly used in control theory to achieve system stability. Since at each time slot, the system is described by the channel conditions and the number of backlogged packets for each user, the Lyapunov function is used to provide a scalar measure, which grows

large when the system moves towards an undesirable state [13]. The scalar measure for queue stability is given by

$$\mathrm{L}\left[\mathbf{Q}(i)\right] = \frac{1}{2}\sum_{k\in\mathcal{U}} Q_k^2(i) \tag{8}$$

where $\mathbf{Q}(i) = [Q_1(i), Q_2(i), \ldots, Q_K(i)]^{\mathrm{T}}$. It provides a scalar measure of congestion in the system [13, Ch. 3].

To minimize the total number of backlogged packets for time instant i, the optimal transmission rate of all users is obtained by minimizing the Lyapunov drift expressed as

$$\mathrm{L}\left[\mathbf{Q}(i+1)\right] - \mathrm{L}\left[\mathbf{Q}(i)\right] = \frac{1}{2}\Big[\sum_{k\in\mathcal{U}}\Big([Q_k(i)-t_k(i)]^+ + \lambda_k(i)\Big)^2 - Q_k^2(i)\Big]. \tag{9}$$

To eliminate the nonlinear operator $[x]^+$, we bound (9) as

$$\leq \sum_{k\in\mathcal{U}} \frac{\lambda_k^2(i)+t_k^2(i)}{2} + \sum_{k\in\mathcal{U}} Q_k(i)\left\{\lambda_k(i)-t_k(i)\right\} \tag{10}$$

by using the following inequality

$$\left[\max(Q-t,0)+\lambda\right]^2 \leq Q^2+t^2+\lambda^2+2Q(\lambda-t). \tag{11}$$

The total number of backlogged packets at any given instant i is reduced by minimizing the conditional expectation of the Lyapunov drift expression (10) given the current number of queued packets $\boldsymbol{Q}(i)$ waiting in the system. The expectation is taken over all possible arrival and transmission rates of the users to obtain the optimal rate allocation strategy.

Now, the conditional Lyapunov drift, denoted by $\Delta(\boldsymbol{Q}(i))$, is given by the infimum over the transmission rate as

$$\inf_{\mathbf{t}} \quad \mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\left\{\mathrm{L}\left[\mathbf{Q}(i+1)\right] - \mathrm{L}\left[\mathbf{Q}(i)\right]|\mathbf{Q}(i)\right\} \tag{12a}$$

$$\leq \underbrace{\mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\Big\{\sum_{k\in\mathcal{U}}\frac{\lambda_k^2(i)+t_k^2(i)}{2}|\mathbf{Q}(i)\Big\}}_{\leq B} + \sum_{k\in\mathcal{U}} Q_k(i)A_k(i) - \mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\Big\{\sum_{k\in\mathcal{U}} Q_k(i)t_k(i)|\mathbf{Q}(i)\Big\}, \tag{12b}$$

where $\mathbf{t}$ and $\boldsymbol{\lambda}$ are the vectors formed by stacking the transmission and the arrival rate of all users. Since they are bounded, the second order moments on the first term in (12b) can be bounded by a constant B without affecting the optimal solution [13]. The second term in (12b) follows the Poisson arrivals.

The expression in (12) looks similar to the WSRM formulation if the weights in the WSRM problem are replaced by the numbers of backlogged packets of the corresponding users. The above approach was extended to the wireless networks in [17], in which the queues were used as weights in the WSRM formulation to determine the transmit precoders. Since the expectation is minimized by minimizing the function inside, the Q-WSRM formulation is given by

$$\underset{\substack{\mathbf{m}_{l,k,n},\\ \mathbf{w}_{l,k,n}}}{\text{maximize}} \quad \sum_{k\in\mathcal{U}} Q_k\Big(\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n})\Big) \tag{13a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}\forall b. \tag{13b}$$

To avoid excessive allocation of the resources, we include an additional rate constraint $t_k \leq Q_k$ to address $[x]^+$ operation in (3). The rate constrained version of the Q-WSRM, denoted by Q-WSRM extended (Q-WSRME) problem for a cellular system, is given by with the additional constraints as

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n}) \leq Q_k\ \forall k\in\mathcal{U} \tag{14}$$

where the precoders are associated with $\gamma_{l,k,n}$ defined in (2). By using the number of queued packets as the weights, the resources can be allocated to the user with more backlogged packets; this essentially results in greedy allocation.

B. JSFRA Scheme via SINR Relaxation

The problem defined in (13) ignores the second order term arising from the Lyapunov drift minimization objective by limiting it to a constant value. In fact, by using $\ell_{q=2}$ norm in (5), we obtain the objective function, similar to (13) as

$$\underset{t_k}{\text{minimize}}\ \sum_k v_k^2 = \underset{t_k}{\text{minimize}}\ \sum_k Q_k^2 - 2\,Q_k t_k + t_k^2 \tag{15}$$

The equivalence is achieved by either removing t_k^2 from (15) or when the number of queued packets is large enough.

By ignoring t_k^2 from (15), the Q-WSRM scheme requires an explicit rate constraint (14) to avoid over-allocation of the resources. In the proposed queue deviation approach, explicit rate constraints are not needed, since they are handled by the objective function (5) itself. In contrast to the WSRM formulation, the JSFRA and the Q-WSRME problems handle the subchannels jointly to obtain an efficient allocation by identifying the optimal space-frequency resources for the users.

We present iterative algorithms to solve (6) using alternating optimization technique in conjunction with SCA approach presented [19]. The problem is to determine the transmit precoders $\mathbf{m}_{l,k,n}$ and the receive beamformers $\mathbf{w}_{l,k,n}$ to minimize the total number of backlogged packets in the system. Note that the SINR expression in (2) cannot be used to formulate the problem directly due to the equality constraint. However, by using additional variables, we can relax the SINR expression in (2) by inequality constraints to solve the problem (6) as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\\ \beta_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{16a}$$

$$\text{subject to} \quad \gamma_{l,k,n} \leq \frac{|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}|^2}{\beta_{l,k,n}} \tag{16b}$$

$$\beta_{l,k,n} \geq \grave{N}_0 + \sum_{(j,i)\neq(l,k)}|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{16c}$$

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}. \tag{16d}$$

The SINR expression in (2) is relaxed by the inequalities (16b) and (16c). Note that (16b) is an under-estimator for SINR $\gamma_{l,k,n}$, and (16c) provides an upper bound for the total interference seen by user $k\in\mathcal{U}_b$, denoted by variable $\beta_{l,k,n}$. Therefore, the problem formulation in (16) is an equivalent approximation for the problem presented in (6). Note that the JSFRA formulation in (16) can be reformulated as a WSRM

problem, which is known to be NP-hard [20], and therefore it belongs to the class of NP-hard problems.

In order to find a tractable solution for (16), we note that (16d) is the only convex constraint with the involved variables. Thus, we need to deal with (16b) and (16c). To this end, we resort to the AO technique by fixing the linear receivers to solve for the transmit beamformers. For fixed receivers $\mathbf{w}_{l,k,n}$, *i.e.*, by fixing the receive beamformers of all users in the system, the problem now is to find optimal transmit precoders $\mathbf{m}_{l,k,n}$ which is still a challenging task. Now, by fixing $\mathbf{w}_{l,k,n}$, (16c) can be written as a second-order cone (SOC) constraint. Thus, the difficulty is due to the non-convexity of the constraint in (16b). Let

$$g(\mathbf{u}_{l,k,n}) \triangleq \frac{|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}|^2}{\beta_{l,k,n}} \tag{17}$$

be the r.h.s of (16b), where $\mathbf{u}_{l,k,n} \triangleq \{\mathbf{m}_{l,k,n}, \mathbf{w}_{l,k,n}, \beta_{l,k,n}\}$. Note that the function $g(\mathbf{u}_{l,k,n})$ is convex for a fixed $\mathbf{w}_{l,k,n}$, since it is in fact the ratio between a quadratic form of $\mathbf{m}_{l,k,n}$ over an affine function of $\beta_{l,k,n}$ as in [21]. The nonconvex set defined by (16b) can be decomposed as a series of convex subsets by linearizing the convex function $g(\mathbf{u}_{l,k,n})$ with the first order Taylor approximation around a fixed operating point $\tilde{\mathbf{u}}_{l,k,n}$ [22], [23], also referred to as SCA in [19]. By using the reduced convex subset for (16b), the problem in (16) is solved iteratively by updating the operating point in each iteration.

For this purpose, let the real and imaginary components of the complex number $\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}$ be represented by

$$p_{l,k,n} \triangleq \Re\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{18a}$$

$$q_{l,k,n} \triangleq \Im\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{18b}$$

and hence $g(\mathbf{u}_{l,k,n}) = (p_{l,k,n}^2 + q_{l,k,n}^2)/\beta_{l,k,n}$.[4] Let $\tilde{\mathbf{u}}_{l,k,n} \triangleq \{\tilde{\mathbf{m}}_{l,k,n}, \tilde{\mathbf{w}}_{l,k,n}, \tilde{\beta}_{l,k,n}\}$ be a minimizer from the previous SCA iteration. Now, by using the first order Taylor approximation around the operating point $\tilde{\mathbf{u}}_{l,k,n}$, we can approximate (16b) as

$$2\frac{\tilde{p}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\left(p_{l,k,n} - \tilde{p}_{l,k,n}\right) + 2\frac{\tilde{q}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\left(q_{l,k,n} - \tilde{q}_{l,k,n}\right) + \frac{\tilde{p}_{l,k,n}^2 + \tilde{q}_{l,k,n}^2}{\tilde{\beta}_{l,k,n}}\left(1 - \frac{\beta_{l,k,n} - \tilde{\beta}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\right) \geq \gamma_{l,k,n}. \tag{19}$$

In summary, for fixed receivers $\tilde{\mathbf{w}}_{l,k,n}$ and operating point $\tilde{\mathbf{u}}_{l,k,n}$ as in (19), obtained by using (18), the relaxed convex subproblem for finding transmit precoders is given by

$$\underset{\substack{\mathbf{m}_{l,k,n},\\ \gamma_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{20a}$$

$$\text{subject to} \quad \beta_{l,k,n} \geq \grave{N}_0 + \sum_{(j,i)\neq(l,k)} |\tilde{\mathbf{w}}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{20b}$$

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}\ \forall b \tag{20c}$$

$$\text{and (19).} \tag{20d}$$

[4]Note that $p_{l,k,n}$ and $q_{l,k,n}$ are just symbolic notations. In CVX [24], for example, we declare $p_{l,k,n}$ and $q_{l,k,n}$ with the '*expression*' qualifier.

Now, the optimal receivers for fixed transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ are obtained by minimizing (16) w.r.t. $\mathbf{w}_{l,k,n}$ as

$$\underset{\substack{\gamma_{l,k,n},\\ \mathbf{w}_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{21a}$$

$$\text{subject to} \quad \beta_{l,k,n} \geq \grave{N}_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\tilde{\mathbf{m}}_{j,i,n}|^2 \tag{21b}$$

$$\text{and (19).} \tag{21c}$$

Solving (21) using the KKT conditions, we obtain the following iterative expression for an optimal receiver $\mathbf{w}_{l,k,n}^o$ as

$$\mathbf{A}_{l,k,n} = \sum_{(j,i)\neq(l,k)}\mathbf{H}_{b_i,k,n}\tilde{\mathbf{m}}_{j,i,n}\tilde{\mathbf{m}}_{j,i,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\,\mathbf{I}_{N_R} \tag{22a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \left(\frac{\tilde{\beta}_{l,k,n}\tilde{\mathbf{m}}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)}}{\|\mathbf{w}_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\|^2}\right)\mathbf{A}_{l,k,n}^{-1}\mathbf{H}_{b_k,k,n}\tilde{\mathbf{m}}_{l,k,n} \tag{22b}$$

where $\mathbf{w}_{l,k,n}^{(i-1)}$ is the receive beamformer from the previous iteration, upon which the linear relaxation is performed for the nonconvex constraint in (16b), as used in the formulation (21). The optimal receiver $\mathbf{w}_{l,k,n}^o$ is obtained by either iterating (22b) until convergence or for a fixed number of iterations. Note that the receiver has no explicit relation with the choice of ℓ_q norm used in the objective. The dependency is implied by the precoders $\mathbf{m}_{l,k,n}$, which depends on the exponent q.

It can be seen that the optimal receiver in (22b) is in fact a scaled version of the MMSE receiver, which is given by

$$\mathbf{R}_{l,k,n} = \sum_{i\in\mathcal{U}}\sum_{j=1}^{L}\mathbf{H}_{b_i,k,n}\tilde{\mathbf{m}}_{j,i,n}\tilde{\mathbf{m}}_{j,i,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\,\mathbf{I}_{N_R} \tag{23a}$$

$$\mathbf{w}_{l,k,n} = \mathbf{R}_{l,k,n}^{-1}\,\mathbf{H}_{b_k,k,n}\,\tilde{\mathbf{m}}_{l,k,n}. \tag{23b}$$

Note that the scaling present in the optimal receiver (22b) has no impact on the received SINRs, and therefore the MMSE receiver in (23b) can also be used. However, the convergence speed is different between the two receiver implementations.

The proposed solution involves two nested iterations, *i.e.*, one for the outer AO loop and the second for the inner SCA loop. Each AO iteration involves two steps, one for finding the transmit precoders by solving (20) iteratively until convergence for fixed receivers, and the other for updating the receive beamformers with the previously found fixed transmit precoders by either solving (22b) recursively or by using (23b).

Let us consider the jth SCA iteration in the ith AO step to find the transmit precoders by solving the subproblem (20). For brevity, let us consider $\mathbf{m}$ as the collection of all transmit precoders in the system as $\mathbf{m} = \{\mathbf{m}_{l,k,n}\}_{\forall l,\forall k,\forall n}$. Similarly, we denote $\mathbf{w}$ and $\boldsymbol{\beta}$ as $\{\mathbf{w}_{l,k,n}\}_{\forall l,\forall k,\forall n}$ and $\{\beta_{l,k,n}\}_{\forall l,\forall k,\forall n}$ respectively. Let $\{\mathbf{m}_j^{(i)}, \boldsymbol{\beta}_j^{(i)}\}$ be the solution of (20) in the $(j-1)$th SCA step. Since SCA is an iterative procedure, the next operating point $\tilde{\mathbf{u}}_{l,k,n}$ for the $(j+1)$th step is updated as $\mathbf{z}_j^{(i)} \triangleq \{\mathbf{m}_j^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\beta}_j^{(i)}\}$. Note that $\mathbf{w}_*^{(i-1)}$ is the receiver, which is obtained by solving either (22b) recursively or by using the MMSE receiver (23b) from the $(i-1)$th AO step.

Upon convergence of the objective sequence, the receivers are updated using fixed transmit precoders $\mathbf{m}_*^{(i)}$ obtained from the previous SCA step. Once the receive beamformers are updated, the transmit precoders are again evaluated with the

Algorithm 1: Algorithm of JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0, j = 0$ and $\tilde{\mathbf{m}}_{l,k,n}$ using single user beamformer satisfying power constraint (20c)
update $\tilde{\mathbf{w}}_{l,k,n}, \tilde{\beta}_{l,k,n}$ using (23b) and (20b) with $\tilde{\mathbf{m}}_{l,k,n}$
repeat
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (20)
    update the constraint set (19) with $\mathbf{u}_{l,k,n}$ and $\mathbf{m}_{l,k,n}$ using (18) and increment $j = j + 1$
  until *SCA convergence or* $j \geq K_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (21) or (23b) with the updated precoders $\mathbf{m}_{l,k,n}$
  $i = i + 1$, $j = 0$
until *Queue convergence or* $i \geq I_{\max}$

newly found receivers by solving (20) recursively or by using (23b). The above procedure is repeated until $i \to \infty$ or for $I_{\max}$ number of iterations as outlined in Algorithm 1.

The initial feasible point $\tilde{\mathbf{u}}_{l,k,n}$ is obtained by fixing $\tilde{\mathbf{m}}_{l,k,n}$ with the respective single-user beamformers satisfying the total power constraint (20c) and $\tilde{\mathbf{w}}_{l,k,n}$'s are obtained with the corresponding MMSE receivers in (23b). Now, by using the fixed transmit and receive beamformers, $\tilde{\beta}_{l,k,n}$'s are evaluated using (20b) for all the users in the system.

Note that in all our numerical simulations, the objective sequence generated by Algorithm 1 converges. However, to analyze the convergence of the sequence of beamformer iterates, we consider a modified objective (46b) with a quadratic term instead of (16a). Using the regularized objective in (46b), the convergence of the sequence of iterates is discussed for the centralized solutions in Appendix A.

C. JSFRA Scheme via MSE Reformulation

In the second method, we solve the JSFRA problem by exploiting the relation between the MSE and the achievable SINR when the MMSE receivers are used at the user terminals [4], [5]. The MSE $\epsilon_{l,k,n}$, for a data symbol $d_{l,k,n}$ is given by

$$\mathbb{E}\big[(d_{l,k,n} - \hat{d}_{l,k,n})^2\big] = \big|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\big|^2 + \sum_{(j,i)\neq(l,k)} \big|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}\big|^2 + \grave{N}_0 = \epsilon_{l,k,n} \quad (24)$$

where $\hat{d}_{l,k,n}$ is the estimate of the transmitted symbol. Plugging the MMSE receivers in (23b) into the MSE expression in (24) and into the SINR expression in (2), we arrive at the following relation between the MSE and the SINR as

$$\epsilon_{l,k,n} = (1 + \gamma_{l,k,n})^{-1}. \quad (25)$$

The above equivalence is valid only if the receivers are based on the MMSE criterion. Using the equivalence in (25), the WSRM objective can be reformulated as the weighted minimum mean squared error (WMMSE) to obtain the precoders for the MU-MIMO scenario as discussed in [5]–[7]. Note that the receive beamformers based on the MMSE criterion are independent of the choice of the ℓ_q norm used in the objective function to obtain the optimal transmit precoders $\mathbf{m}_{l,k,n}$.

Before proceeding futher, let $v'_k = Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}$ denotes the queue deviation corresponding to user k and $\tilde{v}'_k \triangleq a_k^{1/q} v'_k$ be the corresponding weighted objective.By using the relaxed MSE expression in (24), we can reformulate (6) as

$$\underset{\substack{t_{l,k,n}, \mathbf{m}_{l,k,n}, \\ \epsilon_{l,k,n}, \mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \quad (26a)$$

$$\text{subject to } t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n}) \quad (26b)$$

$$\sum_{(j,i)\neq(l,k)} \big|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}\big|^2 + \grave{N}_0 + \big|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\big|^2 \leq \epsilon_{l,k,n} \quad (26c)$$

$$\sum_{n=1}^{N} \sum_{k\in\mathcal{U}_b} \sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n} \mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max} \; \forall b. \quad (26d)$$

The alternative MSE formulation given by (26) is non-convex even for the fixed $\mathbf{w}_{l,k,n}$ due to the constraint (26b). Again we resort to the SCA approach [19] by relaxing the constraint by a sequence of convex subsets using the first order Taylor approximation around a fixed MSE point $\tilde{\epsilon}_{l,k,n}$ as

$$-\log_2(\tilde{\epsilon}_{l,k,n}) - \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\, \tilde{\epsilon}_{l,k,n}} \geq t_{l,k,n} \quad (27)$$

Using the above approximation for the rate constraint, the problem defined in (26) is solved for optimal transmit precoders $\mathbf{m}_{l,k,n}$, MSEs $\epsilon_{l,k,n}$, and the user rates over each sub-channel $t_{l,k,n}$ for fixed receive beamformers. The optimization subproblem to find the transmit precoders for fixed receive beamformers $\mathbf{w}_{l,k,n}$ is given by

$$\underset{t_{l,k,n}, \mathbf{m}_{l,k,n}, \epsilon_{l,k,n}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \quad (28a)$$

$$\text{subject to} \quad (26c), (26d), \text{ and } (27). \quad (28b)$$

The optimal transmit precoders for fixed receivers are obtained by solving the subproblem (28) iteratively and by updating the fixed MSE point $\tilde{\epsilon}_{l,k,n}$ with $\epsilon_{l,k,n}$ from the previous iteration until termination as discussed in Section III-B. The convergence analysis follows the discussions in Appendix A.

D. Reduced Complexity Spatial Resource Allocation (SRA)

The complexity of the JSFRA algorithm scales quickly with the number of sub-channels, since the complexity of an interior point method, which is used to solve the problem, increases with the problem size. Thus, we can use the decomposition methods presented in [9], [10] to overcome this complexity by designing precoders for each sub-channels independently with minimal information exchange.

As an alternative sub-optimal solution, we present a queue minimizing spatial resource allocation (SRA), which solves for the precoders using JSFRA formulation for a specific sub-channel i with a fixed transmit power $P_{\max,i}$. For each sub-channel, the power sharing can either be equal or based on some predetermined limits as in partial frequency reuse satisfying

$$\sum_{i=1}^{N} P_{\max,i} = P_{\max}. \quad (29)$$

Even though N sub-channels are present at any given scheduling instant, precoders are computed for each sub-channel sequentially with $P_{\max,i}$ and the residual number of backlogged packets. Let $Q_{k,i}$ be the number of backlogged packets associated with user k while designing the precoders for the ith sub-channel. Since the precoder design is sequential, *i.e*, the precoders are designed for sub-channels $[0, i-1]$ before the ith sub-channel, the number of queued packets for the first chosen sub-channel is given by $Q_{k,1} = Q_k$. The queues associated with the consecutive sub-channels are updated as

$$Q_{k,i+1} = \max\left(Q_k - \sum_{j=1}^{i}\sum_{l=1}^{L} t_{l,k,j}, 0\right) \; \forall \; k \in \mathcal{U} \tag{30}$$

where $t_{l,k,j}$ is the kth user rate on sub-channel j.

For simplicity we use random sub-channel ordering in our paper, *i.e.*, after finding the precoders for a current sub-channel, we can choose any previously unselected sub-channels as the next candidate sub-channel for which the precoders are identified using the updated backlogged packets. Alternatively, we can also use greedy ordering to select the sub-channels by sorting the norm of the channel seen between the users and the respective serving BSs. However, it comes at the cost of increased complexity. Nevertheless, the SRA scheme will be insensitive to the sub-channel ordering as the number of users in the system increases, due to the available multi-user diversity in the system.

IV. Distributed Solutions

In this section, we discuss the distributed designs for the formulations proposed in Sections III-B and III-C for problem (16). Note that the convex subproblems in (20) or (28) requires a centralized controller to design the precoders for all users in the system. However, the amount of channel state information (CSI) that needs to be exchanged between the BSs and the controller grows significantly as the network size increases (e.g., the number of BSs and users, and the number of associated antennas). The overhead involved in the CSI exchanges can be avoided, if we design the respective precoders independently at each BS with the minimal information exchange, which is determined by the distributed schemes discussed herein.

Let us consider the convex subproblem with fixed receivers $\mathbf{w}_{l,k,n}$ as presented in (20) based on the SINR relaxation for the nonconvex constraint (16b). The following discussions are equally valid for the MSE based solution outlined in (28) as well. Since the objective of (20) can be decoupled across each BS, the centralized problem can be equivalently written as

$$\underset{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\beta_{l,k,n}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{31a}$$

$$\text{subject to} \quad \text{(20b)} - \text{(20d)} \tag{31b}$$

where $\tilde{\mathbf{v}}_b$ denotes the stacked vector of weighted queue deviations corresponding to all users in BS b as $\forall k \in \mathcal{U}_b$.

To begin with, let $\bar{\mathcal{B}}_b$ be the set $\mathcal{B}\backslash\{b\}$ and $\bar{\mathcal{U}}_b$ represent the set $\mathcal{U}\backslash\mathcal{U}_b$. Following an approach similar to the one presented in [11], [12], the coupling constraint (20b) or (26c) can be expressed by stacking the interference from all BSs in $\bar{\mathcal{B}}_{b_k}$ as

$$\grave{N}_0 + \sum_{j=1,j\neq l}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} + \sum_{i\in\mathcal{U}_{b_k}\backslash\{k\}}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 \leq \beta_{l,k,n} \tag{32}$$

where $\zeta_{l,k,n,b}$ is the total interference caused by the transmission of BS b to user $k \in \mathcal{U}_{b_k}$ in spatial stream l and sub-channel n. To ensure (20b), we impose

$$\zeta_{l,k,n,b} \geq \sum_{i\in\mathcal{U}_b}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{j,i,n}|^2 \; \forall b \in \bar{\mathcal{B}}_{b_k}. \tag{33}$$

The decentralization is achieved by decomposing the original convex problem in (31) to a parallel subproblems coordinated by either primal or dual decomposition update. The coupling variables are updated in each iteration by exchanging limited information among the BSs. Before proceeding further, let $\bar{\boldsymbol{\zeta}}_b$ be the vector formed by stacking interference terms (33) from the neighboring BSs to the users of BS b and $\hat{\boldsymbol{\zeta}}_b$ be the stacked interference terms caused by BS b to all users in the neighboring BSs $\bar{\mathcal{B}}_b$, represented as

$$\bar{\boldsymbol{\zeta}}_b = \left[\zeta_{l,k,n,\bar{\mathcal{B}}_b(1)}, \ldots, \zeta_{l,k,n,\bar{\mathcal{B}}_b(|\bar{\mathcal{B}}_b|)}\right]^{\mathrm{T}}, \forall k \in \mathcal{U}_b \tag{34a}$$

$$\hat{\boldsymbol{\zeta}}_b = \left[\zeta_{l,\bar{\mathcal{U}}_b(1),n,b}, \zeta_{l,\bar{\mathcal{U}}_b(2),n,b}, \ldots, \zeta_{l,\bar{\mathcal{U}}_b(|\bar{\mathcal{U}}_b|),n,b}\right]^{\mathrm{T}}. \tag{34b}$$

Let us define the vector $\boldsymbol{\zeta}_b$, formed by stacking the interference terms corresponding to the BS b as

$$\boldsymbol{\zeta}_b = \left[\hat{\boldsymbol{\zeta}}_b^{\mathrm{T}}, \bar{\boldsymbol{\zeta}}_b^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{35}$$

Since the decentralization solution is an iterative procedure, we represent the jth iteration index as $x^{(j)}$. Let $\boldsymbol{\zeta}_b(b_k)$ denote the interference terms corresponding to BS b_k in BS b as

$$\boldsymbol{\zeta}_b(b_k) = \left[\zeta_{l,\mathcal{U}_b(1),n,b_k}, \ldots, \zeta_{l,\mathcal{U}_b(|\mathcal{U}_b|),n,b_k}\right]. \tag{36}$$

To decentralize the problem in (31), the BS specific vector $\boldsymbol{\zeta}_b$ in (35), which are relevant for the BS b, can either be fixed or treated as a variable in accordance to the decomposition method. To decouple the precoder design across BSs, the equivalent downlink channels $\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}, \forall k \in \mathcal{U}$ are to be known at each BS b through the precoded uplink pilots from all the users in the system, where the linear precoders are evaluated at the user using MMSE formulation in (28). Similarly, to update the MMSE receivers at each user k, the equivalent channels $\mathbf{H}_{b,k,n}\mathbf{m}_{l,k',n}, \forall k' \in \mathcal{U}_b, \forall b \in \mathcal{B}$ need to be known. It is obtained through the user specific downlink pilots precoded with the newly obtained transmit beamformers $\mathbf{m}_{l,k,n}, \forall k \in \mathcal{B}$ evaluated at the BS b using the equivalent downlink channels as discussed in [25].

A. Primal Decomposition

In the primal decomposition, the convex problem in (31) is solved for the transmit precoders in an iterative manner by fixing the BS specific interference terms $\boldsymbol{\zeta}_{b_k}$ using master-slave model [11]. The slave subproblem is solved in each BS for the optimal transmit precoders only for the associated users by assuming fixed interference terms $\boldsymbol{\zeta}_{b_k}^{(i)}$ in each ith iteration.

Upon finding the associated transmit precoders by each slave subproblems, the master problem is used to update the BS specific interference terms $\boldsymbol{\zeta}_{b_k}^{(i+1)}$ for the next iteration by using dual variables corresponding to the interference constraint (32) as discussed in [11]. In this manner, the interference variables are updated until the global consensus is reached. The master problem treats $\boldsymbol{\zeta}_b$ as a variable and the slave subproblems assumes it to be a constant in each iteration.

B. Alternating Directions Method of Multipliers (ADMM)

The ADMM approach can also be used to decouple the precoder design across multiple BSs to solve the convex subproblem in (31). Generally, the ADMM is preferred over the dual decomposition in [12] for its robustness and improved convergence behavior [10]. In contrast to the primal decomposition, the ADMM relaxes the interference constraints by including in the objective function of each subproblem with a penalty pricing [9], [10]. Similar approach for the precoder design in the minimum power context was considered in [26].

Using the formulation presented in [10], [26], we can write the BS b specific ADMM subproblem for the jth iteration as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n}\\ \beta_{l,k,n},\boldsymbol{\zeta}_b}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}_b^{(j)\,\mathrm{T}}(\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(j)}) + \tfrac{\rho}{2}\|\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(j)}\|^2 \quad (37a)$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \le P_{\max} \quad (37b)$$

$$\sum_{\bar{b}\in\bar{\mathcal{B}}_b} \zeta_{l,k,n,\bar{b}} + \sum_{\{\bar{l},\bar{k}\}\neq\{l,k\}} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{\bar{l},\bar{k},n}|^2 + \dot{N}_0 \le \beta_{l,k,n} \quad (37c)$$

$$\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} |\mathbf{w}_{\bar{l},\bar{k},n}^{\mathrm{H}}\mathbf{H}_{b,\bar{k},n}\mathbf{m}_{l,k,n}|^2 \le \zeta_{\bar{l},\bar{k},n,b} \;\forall \bar{k}\in\bar{\mathcal{U}}_b \;\forall n \quad (37d)$$

$$\text{and (19)} \quad (37e)$$

where $\boldsymbol{\zeta}_b^{(j)}$ denotes the interference vector updated from the earlier iteration and $\boldsymbol{\nu}_b^{(j)}$ represents the dual vector corresponding to the equality constraint at the jth iteration as

$$\boldsymbol{\zeta}_b = \boldsymbol{\zeta}_b^{(j)}. \quad (38)$$

Upon solving (37) for $\boldsymbol{\zeta}_b \forall b$ in the jth iteration, the next iterate is updated by exchanging the corresponding interference terms between two BSs b and b_k as

$$\boldsymbol{\zeta}_{b_k}(b)^{(j+1)} = \boldsymbol{\zeta}_b(b_k)^{(j+1)} = \frac{\boldsymbol{\zeta}_b(b_k) + \boldsymbol{\zeta}_{b_k}(b)}{2}. \quad (39)$$

The dual vector for the next iteration is updated by using the subgradient search to maximize the dual objective as

$$\boldsymbol{\nu}_b^{(j+1)} = \boldsymbol{\nu}_b^{(j)} + \rho\left(\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(j+1)}\right) \quad (40)$$

where step size parameter ρ is chosen in accordance with [10] to depend on the system model. The convergence rate is susceptible to the choice of step size parameter ρ. In numerical simulations, we consider step size $\rho = 2$. The above iterative procedure is performed until convergence or terminated when exceeding a predetermined number of steps, say, $J_{\max}$. Algorithm 2 outlines a practical way of implementing the ADMM based precoder design with minimal signaling overhead.

Algorithm 2: Distributed JSFRA scheme using ADMM

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall users$
Initialize: $i = 0, j = 0$ and $\mathbf{m}_{l,k,n}$ satisfying (37b)
update $\mathbf{w}_{l,k,n}$ with (23b) and $\tilde{\mathbf{u}}_{l,k,n}$ using (16c) and (18)
initialize the vectors $\boldsymbol{\nu}_b^{(0)} = \mathbf{0}^{\mathrm{T}}, \boldsymbol{\zeta}_b^{(0)} = \mathbf{0}^{\mathrm{T}}, \forall b \in \mathcal{B}$
foreach *BS* $b \in \mathcal{B}$ **do**
  repeat
    repeat
      solve for $\mathbf{m}_{l,k,n}$ and $\boldsymbol{\zeta}_b$ with (37) using $\boldsymbol{\zeta}_b^{(j)}$
      exchange $\boldsymbol{\zeta}_b$ among BSs in $\mathcal{B}$
      update $\boldsymbol{\zeta}_b^{(j+1)}$ and $\boldsymbol{\nu}_b^{(j+1)}$ using (39) and (40)
      update $j = j + 1$
    until *convergence or* $j \ge J_{\max}$, $\forall b \in \mathcal{B}$
    send precoded downlink pilot with $\mathbf{m}_{l,k,n}$
    evaluate $\mathbf{w}_{l,k,n}$ using (23b) at each user
    update $\tilde{\mathbf{u}}_{l,k,n}$ using (16c) and (18) for SCA point
    or $\tilde{\epsilon}_{l,k,n}$ using (26c) for MSE operating point
    $i = i + 1$, $j = 0$
  until *convergence or* $i \ge I_{\max}$
end

The convergence analysis for the distributed algorithms is provided in Appendix B. To reduce the amount of information exchange among the coupling BSs, often distributed schemes are iterated for few steps, say, $J_{\max}$. If so, the convergence of the iterates is not guaranteed, since it is difficult to show the monotonicity of objective in each SCA step. The reason is that in each primal or ADMM update, the global objective need not be decreasing monotonically, despite reaching the centralized solution upon convergence. However, if the following conditions are satisfied in each SCA step, *i.e.*, (i) by ensuring uniqueness of the minimizer, and (ii) by choosing $J_{\max}$ to be sufficiently large in order to ensure strict monotonicity of the objective, then the convergence discussions follows Appendices A and B. Moreover, if the receivers are updated together with the transmit precoders as in Algorithm 2, then strict monotonic decrease of the objective sequence is still ensured. The reason is that the MMSE receiver is optimal for the fixed transmit precoders obtained after each SCA update.

C. Decomposition via KKT Conditions for MSE Formulation

In this section, we discuss an alternative way to decentralize the precoder design across the coordinating BSs in $\mathcal{B}$ based on the MSE reformulation method discussed in Section III-C. In contrast to Sections IV-A and IV-B, the problem is solved using the KKT conditions in which the transmit precoders, receive beamformers and the subgradient updates are performed at the same time to minimize the queue deviation with few number of iterations. The alternative way is motivated by the fact that the distributed approaches presented in the preceeding sections may not be efficient for large systems in terms of signaling overhead involved in exchanging the coupling variables and the receivers.

In this section, we provide an algorithm that can be of practical importance owing to lower signaling requirements.

We consider an idealized TDD system due to the knowledge of complete CSI at the transmitter. Similar work has been considered for the WSRM problem with minimum rate constraints in [8]. Since the work in [8] are similar to that of the Q-WSRME scheme with an additional maximum rate constraint (14), the formulations require explicit dual variables to handle the maximum rate constraint, thereby making the problem difficult to solve in an iterative manner.

In the proposed scheme, the maximum rate constraints are implicitly handled by the objective function without any need for explicit constraints. However, due to the non-differentiability of the objective, the KKT conditions are not computationally useful to find the optimization variables. In order to make the objective function differentiable, we consider the following two cases for which the absolute operator can be ignored without affecting the optimal solution, namely,

- when the exponent q is even, or
- when the number of backlogged packets waiting to be transmitted for each user is large enough, *i.e*, $Q_k \gg \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ to ignore the absolute operator in the objective and queues in the first place as well.

With the assumption of either one of the above conditions, the problem in (28) can be written as

$$\underset{\substack{t_{l,k,n},\mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{U}_b} a_k \left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^q \tag{41a}$$

subject to

$$\alpha_{l,k,n} : \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 + \grave{N}_0 + \sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\right|^2 \leq \epsilon_{l,k,n} \tag{41b}$$

$$\sigma_{l,k,n} : \log_2(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\tilde{\epsilon}_{l,k,n}} \leq -t_{l,k,n} \tag{41c}$$

$$\delta_b : \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}\ \forall b \tag{41d}$$

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (41b), (41c) and (41d).

The problem in (41) is solved using the KKT conditions which include stationarity, complementary slackness, and primal and dual feasibility requirement as shown in Appendix C. In particular, we propose an iterative algorithm to compute a solution to the system of equations (62) in Appendix C as

$$\mathbf{m}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\alpha_{y,x,n}^{(i-1)}\mathbf{H}_{b_k,x,n}^{\mathrm{H}}\mathbf{w}_{y,x,n}^{(i-1)}\mathbf{w}_{y,x,n}^{\mathrm{H}\,(i-1)}\mathbf{H}_{b_k,x,n} + \delta_b\mathbf{I}_{N_T}\Big)^{-1}\alpha_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)} \tag{42a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\mathbf{H}_{b_x,k,n}\mathbf{m}_{y,x,n}^{(i)}\mathbf{m}_{y,x,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_x,k,n}^{\mathrm{H}} + N_0\mathbf{I}_{N_R}\Big)^{-1}\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}^{(i)} \tag{42b}$$

$$\epsilon_{l,k,n}^{(i)} = \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}^{(i)}\right|^2 + \grave{N}_0 + \sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}^{(i)}\right|^2 \tag{42c}$$

$$t_{l,k,n}^{(i)} = -\log_2(\epsilon_{l,k,n}^{(i-1)}) - \frac{\left(\epsilon_{l,k,n}^{(i)} - \epsilon_{l,k,n}^{(i-1)}\right)}{\log(2)\,\epsilon_{l,k,n}^{(i-1)}} \tag{42d}$$

$$\sigma_{l,k,n}^{(i)} = \left[\tfrac{a_k\,q}{\log(2)}\left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}^{(i)}\right)^{(q-1)}\right]^+ \tag{42e}$$

$$\alpha_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i-1)} + \rho^{(i)}\left(\tfrac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i)}} - \alpha_{l,k,n}^{(i-1)}\right) \tag{42f}$$

Since the dual variables $\alpha^{(i)}$ and $\sigma^{(i)}$ are interdependent in (42), one has to be fixed to optimize for the other. So, $\alpha^{(i)}$ is fixed to evaluate $\sigma^{(i)}$ using (42). At iteration i, the dual variables $\alpha^{(i)}$ is a point in the line segment between $\alpha^{(i-1)}$ and $\frac{\sigma^{(i)}}{\epsilon^{(i)}}$ determined by using a diminishing or a fixed step size $\rho^{(i)} \in (0, 1)$. The choice of $\rho^{(i)}$, which depends on the system, affects the convergence behavior and also controls the oscillations in the users' rate when $\sigma^{(i)}$ is negative (before projection) due to over-allocation. However, in all numerical simulations, the step size $\rho^{(i)}$ is fixed to 0.1 irrespectively.

For a physical interpretation, when the allocated rate $t_k^{(i-1)}$ is greater than Q_k for a user k, the corresponding dual variable $\sigma^{(i)}$ will be negative and due to the projection operator $[x]^+$ in (42e), it will be zero, thereby forcing $\alpha_k^{(i)} < \alpha_k^{(i-1)}$ as in (42f). Once $\alpha_k^{(i)}$ is reduced, the precoder weight in (42a) is lowered to make the rate $t_k^{(i)} < t_k^{(i-1)}$ eventually.

The KKT expressions in (42) are solved in an iterative manner by initializing the transmit and the receive beamformers $\mathbf{m}_{l,k,n}, \mathbf{w}_{l,k,n}$ with the single user beamforming and the MMSE vectors. The dual variable α's are initialized with ones to have equal priorities to all the users in the system. Then the transmit and the receive beamformers are evaluated using the expressions in (42). The transmit precoder in (42a) depends on the BS specific dual variable δ_b, which can be found by bisection search satisfying the total power constraint (41d). Note that the fixed SCA operating point is given by $\tilde{\epsilon}_{l,k,n} = \epsilon_{l,k,n}^{(i-1)}$, which is considered in the expression (42).

To obtain a practical distributed precoder design, we assume that each BS b knows the corresponding equivalent channels $\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}, \forall k \in \mathcal{U}$, which embeds the receivers $\mathbf{w}_{l,k,n}$, through precoded uplink pilot signaling. We extend the decentralization methods discussed in [25], for the current problem as follows. Upon receiving the updated transmit precoders from all BSs in $\mathcal{B}$, each user will evaluate the MMSE receiver (42b) and notify to all BSs by using precoded uplink pilots. On receiving the pilots, each BS updates the MSE in (24) as

$$\epsilon_{l,k,n}^{(i)} = 1 - \mathbf{w}_{l,k,n}^{(i)\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}^{(i)}. \tag{43}$$

Using $\epsilon_{l,k,n}^{(i)}$, the variables $t_{l,k,n}^{(i)}, \sigma_{l,k,n}^{(i)}$, and $\alpha_{l,k,n}^{(i)}$ are updated using (42d), (42e) and (42f) respectively, and the obtained dual variables $\alpha_{l,k,n}$ are exchanged between the BSs to evaluate the transmit precoders $\mathbf{m}_{l,k,n}^{(i+1)}$ for the next iteration. The SCA operating point is also updated with the current MSE value.

To avoid the backhaul exchanges between the BSs, as an alternative approach, users can perform all the required processing and BSs will update the precoders based on the

Algorithm 3: KKT approach for the JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall l \in \{1, 2, \ldots, L\}$
Initialize: $i = 1$, $\mathbf{w}_{l,k,n}^{(0)}, \tilde{\epsilon}_{l,k,n}$ randomly, dual variables $\alpha_{l,k,n}^{(0)} = 1$, and $I_{\max}$ for certain value
foreach *BS* $b \in \mathcal{B}$ **do**
  repeat
    update $\mathbf{m}_{l,k,n}^{(i)}$ using (42a), and perform precoded downlink pilot transmission
    find $\mathbf{w}_{l,k,n}^{(i)}$ using (42b) at each user
    evaluate $\epsilon_{l,k,n}^{(i)}$, $t_{l,k,n}^{(i)}$, $\sigma_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ using (42c) and (42d), (42e) and (42f) at each user with the updated $\mathbf{w}_{l,k,n}^{(i)}$
    using precoded uplink pilots, $\mathbf{m}_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ are notified to all BSs in $\mathcal{B}$
    $i = i + 1$
  until *until convergence or* $i \geq I_{\max}$
end

feedback from all the users. Upon receiving the transmit precoders from the BSs, each user will update the receive beamformer $\mathbf{w}_{l,k,n}$, the MSE $\epsilon_{l,k,n}$, and the dual variables $\sigma_{l,k,n}$ and $\alpha_{l,k,n}$. Then, the updated $\alpha_{l,k,n}$ and $\mathbf{w}_{l,k,n}$ are notified to the BSs using two separate precoded uplink pilots with $\tilde{\mathbf{w}}_{l,k,n}^{(i)} = \left(\alpha_{l,k,n}^{(i)}\right)^{1/2} \mathbf{w}_{l,k,n}^{*(i)}$ and $\bar{\mathbf{w}}_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i)} \mathbf{w}_{l,k,n}^{*(i)}$ as the precoders, where $\mathbf{x}^*$ is the complex conjugate of $\mathbf{x}$. Upon receiving the precoded uplink pilots, each BS evaluates the equivalent channels $\mathbf{H}_{b,k,n}^{\mathrm{T}} \tilde{\mathbf{w}}_{l,k,n}^{(i)}$ and $\mathbf{H}_{b,k,n}^{\mathrm{T}} \bar{\mathbf{w}}_{l,k,n}^{(i)}$ to update the transmit precoders using (42a). The algorithmic description of the above scheme is presented in Algorithm 3.

We conclude this section by providing some remarks on selecting an algorithm within centralized and distributed approaches for certain scenarios. Among the centralized algorithms proposed in Section III, both the SINR relaxation and the MSE reformulation schemes are equally attractive when $N_R > 1$. However, when $N_R = 1$ the SINR relaxation is preferable, since the receiver update is not needed, unlike the MSE reformulation which requires the scalar MMSE receiver update. Similarly, for the distributed approaches with $N_R = 1$, the SINR relaxation is preferred to the MSE reformulation schemes, since the variables to be exchanged among BSs are only the scalar interference values. However, when $N_R > 1$ the KKT scheme in Section IV-C is favorable, since the overhead required to achieve a certain throughput improvement is less, as compared to the primal decomposition or ADMM.

V. Simulation Results

The simulations carried out in this work consider the path loss (PL) varying uniformly across all users in the system with the small-scale fading drawn from the *i.i.d.* samples. The queues are generated based on the Poisson process with the average values specified for each numerical experiments.

A. Centralized Solutions

We discuss the performance of the centralized algorithms in Section III for some system configurations. To begin with, we consider a single cell single-input single-output (SISO) model operating at 10 dB signal-to-noise ratio (SNR) with $K = 3$ users sharing $N = 3$ sub-channel resources. The total number of backlogged packets waiting at the transmitter for each user is given by $Q_k = 4, 8$ and 4 bits, respectively.

Table I tabulates the channels of the users over each sub-channel followed by the rates assigned by three different algorithms, Q-WSRME allocation, JSFRA approach and the sub-channel wise Q-WSRM scheme using the WMMSE design [6]. The metric used for the comparison is the total number of backlogged bits left over after each transmission, which is denoted as $\chi = \sum_{k=1}^{K} [Q_k - t_k]^+$. Even though $\mathcal{U}(1)$ and $\mathcal{U}(3)$ have equal number of backlogged packets of $Q_1 = Q_3 = 4$ bits, user $\mathcal{U}(3)$ is scheduled in the first sub-channel due to the better channel condition. In contrast, the JSFRA approach assigns the first user on the first sub-channel, which reduces the total number of backlogged packets. The rate allocated for $\mathcal{U}(2)$ on the second sub-channel is higher in JSFRA scheme compared to the others. It is due to the efficient allocation of the total power shared across the sub-channels.

For a MIMO setup, we consider a system with $N = 4$ and $N = 2$ sub-channels with $N_B = 3$ BSs, each equipped with $N_T = 4$ transmit antennas operating at 10dB SNR, serving $|\mathcal{U}_b| = 3$ users each. The PL between the BSs and the users are uniformly generated from $[0, -3]$ dB and the BS-user associations are made by selecting the BS with the lowest PL component. Fig. 1(a) and Fig. 1(b) compares the performance of the centralized schemes for $N_R = 1$ and $N_R = 2$ receive antenna cases respectively.

The comparison in Fig. 1 is made in terms of the total number of residual bits remaining in the system after each SCA update. The Q-WSRM scheme is not optimal due to the problem of over-allocation when the number of queued packets is small. In contrast, the Q-WSRME algorithm provides better allocation with the explicit rate constraint to avoid the over-allocation. For both scenarios in Fig. 1, the Q-WSRME performs marginally inferior to the JSFRA algorithms due to the weights used in the algorithm, since the Q-WSRME scheme favors the users with the large number of backlogged packets as compared to the users with better channel conditions. Note that the receivers are updated along with the SCA update instants *i.e*, $K_{\max} = 1$ for both Fig. 1(a) and Fig. 1(b). However, the performance loss incurred by the combined update of the SCA operating point and the receiver is marginal.

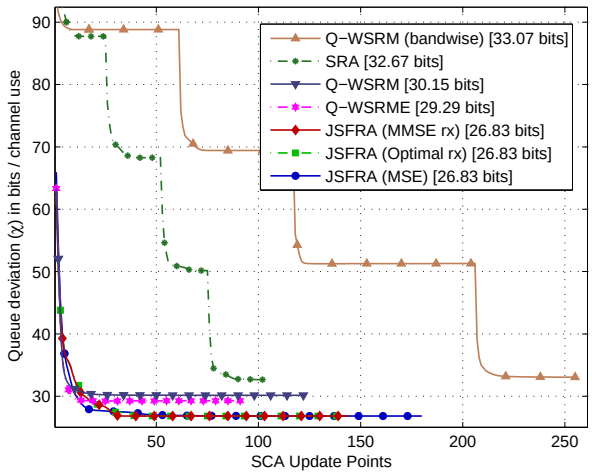
The behavior of the JSFRA algorithm for different exponents q is outlined in the Table II for the users located at the cell-edge of the system employing $N_T = 4$ transmit antennas. It is evident that the JSFRA algorithm minimizes the total number of queued bits for the ℓ_1 norm compared to the ℓ_2 norm, which is shown in the column displaying the total number of left over packets χ in bits. The ℓ_∞ norm provides fair allocation of the resources by making the left over packets to be equal for all users to $\chi_k = 3.58$ bits.

B. Distributed Solutions

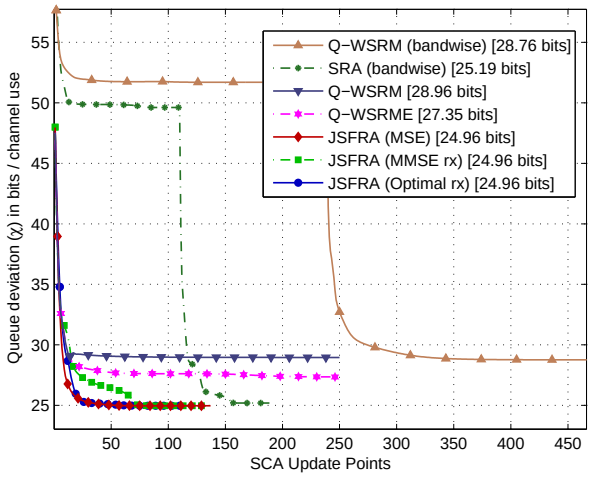
The distributed algorithms are compared using the total number of backlogged packets after each SCA update. Fig.

TABLE I
SUB-CHANNEL-WISE LISTING OF CHANNEL GAINS AND RATE ALLOCATIONS BY DIFFERENT ALGORITHMS FOR A SCHEDULING INSTANT

Users	Queued Packets	Channel Gains			Q-WSRME approach (modified *backpressure*)			JSFRA Scheme			Q-WSRM band Alloc Scheme		
		SC-1	SC-2	SC-3	SC-1	SC-2	SC-3	SC-1	SC-2	SC-3	SC-1	SC-2	SC-3
1	4	1.71	0.53	0.56	0	0	0	4.0	0	0	0	0	0
2	8	0.39	1.41	1.03	0	4.88	3.11	0	5.49	0	0	4.39	3.53
3	4	2.34	1.26	2.32	4.0	0	0	0	0	4.0	5.81	0	0
Remaining backlogged packets (χ)					3.92 bits			2.51 bits			5.89 bits		



(a). System model $\{N, N_B, K, N_T, N_R\} = \{4, 3, 9, 4, 1\}$, number of backlogged bits for each user $Q_k = [14, 15, 14, 8, 12, 9, 12, 11, 11]$ bits



(b). System model $\{N, N_B, K, N_T, N_R\} = \{2, 3, 9, 4, 2\}$, number of backlogged bits for each user $Q_k = [9, 12, 8, 12, 5, 4, 10, 8, 5]$ bits

Fig. 1. Total number of backlogged packets χ present in the system after each SCA updates using $\ell_1(q = 1)$ norm for JSFRA schemes

TABLE II
NUMBER OF BACKLOGGED BITS ASSOCIATED WITH EACH USER FOR A SYSTEM $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$.

q	user indices								χ
	1	2	3	4	5	6	7	8	
1	15.0	3.95	5.26	8.95	7.0	11.9	12.0	9.7	25.15
2	11.2	3.9	10.76	10.65	10.27	9.68	8.77	5.9	27.77
∞	11.4	4.4	10.4	10.4	10.4	8.4	8.4	6.4	28.68
Q_k	15.0	8.0	14.0	14.0	14.0	12.0	12.0	10.0	

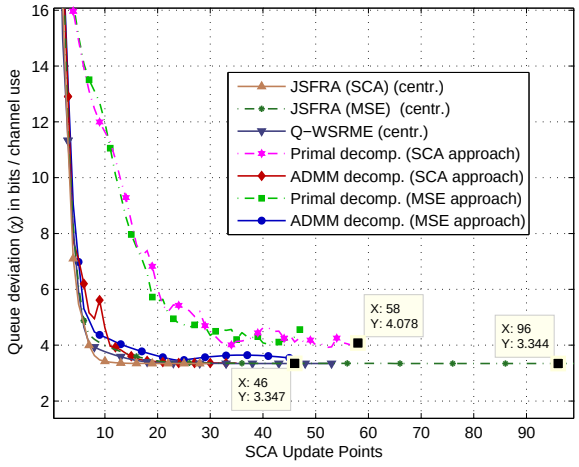


Fig. 2. Convergence of the centralized and the distributed algorithms for $\{N, N_B, K, N_T, N_R\} = \{3, 2, 8, 4, 1\}$ using ℓ_1 norm for JSFRA schemes with $Q_k = [5, 7, 9, 11, 8, 12, 5, 4]$ bits

2 compares the performances of the algorithms with the PL varies uniformly between $[0, -6]$ dB and each BS serves $|\mathcal{U}_b| = 4$ users. As discussed in Section IV, the performance and the convergence speed of the distributed algorithms are susceptible to the step size $\rho^{(i)}$. Due to the fixed interference levels in the primal approach, it may lead to infeasible solutions if the initial or any intermediate update is not feasible.

Fig. 2 compares the performance of the JSFRA schemes discussed in Sections III-B and III-C using primal and ADMM approaches. For each SCA update, the primal or the ADMM scheme is performed for $J_{\max} = 20$ iterations to exchange the respective coupling variables. The number of backlogged packets only at the SCA points are marked in the figure. The performance of the distributed approaches is similar to that of the centralized schemes if the distributed algorithms are allowed converge. However, in our simulations, we observe that $J_{\max} = 20$ is sufficient for the ADMM to converge.

Fig. 3 compares the performances of the centralized and the KKT algorithm in Section IV-C for different exponents with PL chosen uniformly between $[0, -3]$ dB. The ℓ_1 norm JSFRA scheme performs better over other schemes due to the greedy nature of the objective. The KKT approach for ℓ_1 norm is not defined due to the non-differentiability of the objective as discussed in Section IV-C. If used for ℓ_1 norm, the over-allocation will not affect the dual variables $\sigma_{l,k,n}$ and $\alpha_{l,k,n}$ since the queue deviation is raised to the power zero in (42e). A heuristic method is proposed in Fig. 3 by assigning zero for $\sigma_{l,k,n}$ when $Q_k - t_k < 0$ to addresses the over-allocation.

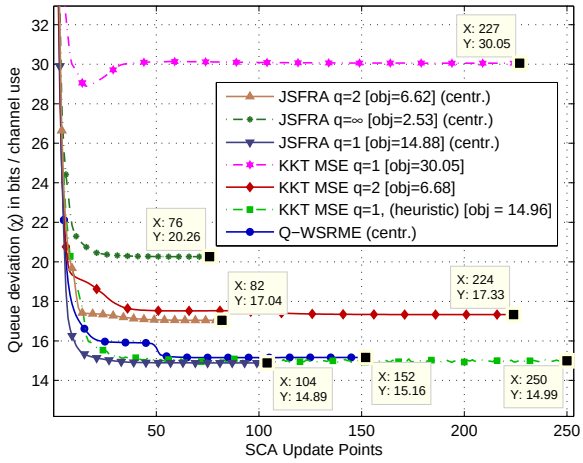


Fig. 3. Impact of varying q on the total number of backlogged packets after each SCA update for a system $\{N, N_B, K, N_T, N_R\} = \{5, 2, 8, 4, 1\}$ and $Q_k = [9, 16, 14, 16, 9, 13, 11, 12]$ bits

The heuristic approach oscillates near the converging point with the deviation determined by the factor ρ used in (42f). The objective values are mentioned in the legend for all the schemes and the ℓ_1 norm is used for comparison.

C. Queuing Analysis over Multiple Transmission Slots

In this section, we numerically study the performance of the centralized algorithms with different ℓ_q values over multiple transmission slots. The system model examined for the illustrations is provided in Fig. 4. For all users in the system, the average arrivals A_k's are fixed and varied equally for the model considered in Fig. 4(a), and for Fig. 4(b), the average arrival is fixed to be $A_k = 6$ bits. Note that the instantaneous arrivals $\lambda_k(i)$ are all different and it follows the Poisson process. The PL is modeled as a uniform random variable $[0, -6]$ dB.

Fig. 4(a) plots the average of the total number of backlogged packets left out in the system after each transmission instant, *i.e*, $\mathbb{E}_i\left[\sum_k \left[Q_k(i) - t_k(i)\right]^+\right]$. Unlike the Q-WSRM scheme, the average backlogged packets of the ℓ_2 JSFRA scheme is comparable to the Q-WSRME approach for all average arrival rates due to the explicit rate constraints (14). However, when $A_k \geq 7$ bits in Fig. 4(a), both Q-WSRM and Q-WSRME schemes perform the same since the problem of over-allocation is negligible. The performance of the ℓ_1 JSFRA scheme outperforms all other schemes in terms of the average number of residual packets due to the greedy allocation at each instant.

Fig. 4(a) also includes the uncoordinated ℓ_1 JSFRA scheme and the time division multiplexing (TDM) mode, which ignores the inter-cell interference terms in the SINR expressions while designing the precoders. The performance of the TDM scheme with the total power constraint is inferior to the uncoordinated transmission due to the diverse user PL variations in the system model. Fig. 4(b) compares the number of backlogged packets left in the system after each transmission slot by different centralized algorithms. The total number of residual packets for the Q-WSRM scheme is noticeably large in comparison with the other schemes in Fig. 4(b). This performance loss is due to the inability in controlling the over-allocations at each instant. The instantaneous fairness constraint imposed by the ℓ_∞ JSFRA scheme is effective in reducing the number of backlogged packets for the average arrival rate considered in Fig. 4(b). However, in Fig. 4(a), the performance of the ℓ_∞ JSFRA is inferior to the Q-WSRM scheme, since the fairness is not effective when the system is unstable, *i.e*, when $A_k \geq 7$ in the current model.

VI. Conclusions

In this paper, we have addressed the problem of allocating downlink space-frequency resources to the users in a multi-cell MIMO IBC system using OFDM. The resource allocation is considered as a joint space-frequency precoder design problem since the allocation of a resource to a user is obtained by a non-zero precoding vector. We have proposed the JSFRA scheme by relaxing the nonconvex constraint by a sequence of convex subsets using the SCA for designing the precoders to minimize the total number of user queued packets. Additionally, an alternative MSE reformulation approach has been proposed by using the SCA to address the nonconvex constraints for a fixed MMSE receivers. We also proposed various methods to decentralize the precoder designs for the JSFRA problem using primal and ADMM methods. Finally, we proposed a practical iterative algorithm to obtain the precoders in a decentralized manner by solving the KKT expressions of the MSE reformulated problem. The proposed iterative algorithm requires few iterations and limited signaling exchange between the coordinating BSs to obtain the efficient precoders for a given number of iterations. Numerical results are used to compare the performances of the proposed schemes.

Appendix A
Convergence Proof for Centralized Algorithm

The following conditions are required to show the convergence of Algorithm 1, which is used to solve problems (16) and (26) in an iterative manner.

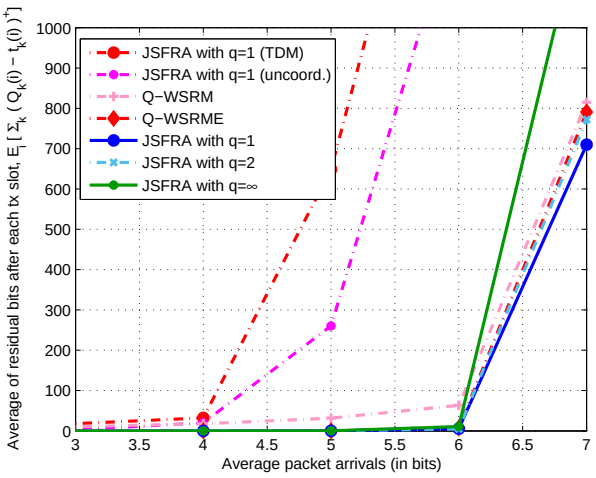
(a) Objective function should be bounded below
(b) Feasible set should be compact
(c) Sequence of objective values should be strictly decreasing
(d) Uniqueness of the minimizer in each step.

The conditions (a), (b), and (c) are required to prove the convergence of objective sequence generated by Algorithm 1. However, to ensure strict monotonicity of the objective sequence, we require condition (d). Therefore, by assuming (a), (b), and (c) are satisfied by Algorithm 1, the convergence of the objective sequence can be shown by using [27, Th. 3.14]. Finally, by using the discussions in [19], [22], [23], we show that every limit point of the sequence of iterates generated by Algorithm 1 is a stationary point of nonconvex problems (16) and (26). The term iterates denotes the stacked vector of solution variables obtained in each iteration.

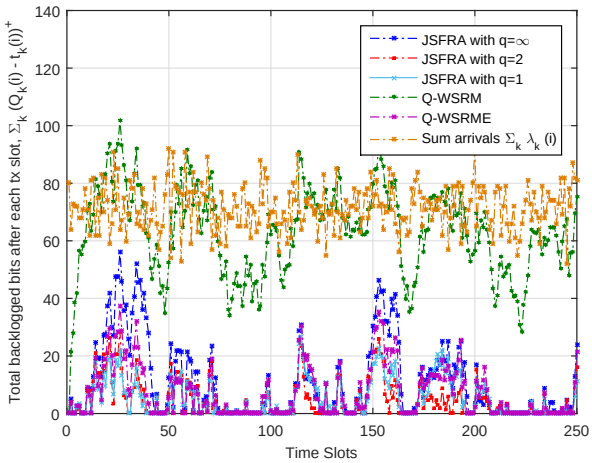
Before discussing the convergence analysis, let us consider a generalized formulation for the problems in (16) and (26) as

$$\underset{\mathbf{m},\mathbf{w},\boldsymbol{\gamma}}{\text{minimize}} \quad \hat{f}(\mathbf{m}, \mathbf{w}, \boldsymbol{\gamma}) \tag{44a}$$

$$\text{subject to} \quad h(\boldsymbol{\gamma}) - g_0(\mathbf{m}, \mathbf{w}) \leq 0 \tag{44b}$$



(a). Average backlogged packets in the system after 250 transmission instants



(b). Total backlogged packets at each transmission slot for $A_k = 6$ bits

Fig. 4. Time analysis of the Queue dynamics for a system $\{N, N_B, K, N_T, N_R\} = \{4, 2, 12, 4, 1\}$

$$g_1(\mathbf{m}, \mathbf{w}) \leq 0 \tag{44c}$$
$$g_2(\mathbf{m}) \leq 0 \tag{44d}$$

where the functions g_2 and $\hat{f}$ are convex, and the function h is linear. Let functions g_0 and g_1 be convex w.r.t either $\mathbf{m}$ or $\mathbf{w}$, but not jointly convex on both variables. The constraint (44b) corresponds to (16b) or (26b) and the constraint (44c) corresponds to (16c) or (26c) respectively. Other convex constraints are handled by (44d) and the feasible set of (44) is given by

$$\mathcal{F} = \{\, \mathbf{m}, \mathbf{w}, \boldsymbol{\gamma} \;\big|\; h(\boldsymbol{\gamma}) - g_0(\mathbf{m}, \mathbf{w}) \leq 0, \; g_1(\mathbf{m}, \mathbf{w}) \leq 0, g_2(\mathbf{m}) \leq 0\}. \tag{45}$$

We use the following notations to discuss the convergence proof. Since Algorithm 1 involves two nested iterations, *i.e.*, one for SCA and another one for AO, we denote AO step by a superscript (i) and SCA iteration by a subscript k. Let $\mathbf{m}$, $\mathbf{w}$, and $\boldsymbol{\gamma}$ be the stacked vector all transmit precoders, receive beamformers, and other optimization variables present in formulations (16) and (26) respectively. Since Algorithm 1 includes two SCA iterations performed until convergence in each AO iteration, let us now consider AO step i for fixed $\mathbf{w}$. The solution obtained at the kth SCA step is represented by $\mathbf{m} = \mathbf{m}_k^{(i)}$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_k^{(i)}$. As $k \to \infty$, the objective sequence converges, and the corresponding solution is denoted as $\mathbf{m}_*^{(i)}$ and $\boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}$. Similarly, for fixed $\mathbf{m}$ and as $k \to \infty$, the corresponding solution obtained is represented as $\mathbf{w} = \mathbf{w}_*^{(i)}$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{*|\mathbf{m}}^{(i)}$ respectively.

A. Bounded Objective Function and Compact Feasible Set

The feasible sets of the problems (16) and (26) are bounded and closed, which is due to the total power constraint on the transmit precoders (16d), therefore, the sets are compact.

The minimum of the norm in the objective is $\|x\|_q > -\infty$, therefore, it is bounded below. The objective function is Lipschitz continuous over the feasible set. The objective function is bounded from above as well, since the feasible set is bounded.

B. Uniqueness of the Iterates and Strong Convexity

To ensure the uniqueness of the transmit and the receive beamformers, the objective function of the convex subproblems in (20) and (28) are regularized by a strongly convex function in each SCA iteration k and AO update step i as

$$\hat{f}(\mathbf{z}) = \|\tilde{\mathbf{v}}\|_q \tag{46a}$$
$$f(\mathbf{z}) = \hat{f}(\mathbf{z}) + \tau_k^{(i)} \|\mathbf{z} - \mathbf{z}_k^{(i)}\|_2^2 \tag{46b}$$

where $\mathbf{z}$ is a vector stacking all optimization variables, which can be either $[\mathbf{m}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}]$ or $[\mathbf{m}_*^{(i)}, \mathbf{w}, \boldsymbol{\gamma}]$ depending on the problem and $\mathbf{z}_k^{(i)} \triangleq [\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)}]$ is the solution obtained in the previous $(k-1)$th SCA step. The positive constant $\tau_k^{(i)} > 0$ ensures the strong convexity of f in each step and $c > 0$ be the constant of strong convexity, defined as $c \triangleq \min_k\{\tau_k^{(i)}\}$, as shown in [28]. Thus, due to the strong convexity of f in (46b), uniqueness of the minimizer is guaranteed in each SCA step as discussed in [28], [29].

C. Strict Monotonicity of the Objective Sequence

In the following discussions, we consider the modified objective f in (46b) instead of $\hat{f}$ due to the uniqueness of the minimizer and upon convergence of the objective, f is equal to $\hat{f}$. Therefore, the discussions are valid for the JSFRA problem in (16) by regularizing the objective (16a) as (46b).

To begin with, let us consider the variable $\mathbf{w}$ is fixed for the ith AO with the optimal value found in previous iteration $i-1$ as $\mathbf{w}_*^{(i-1)}$. In order to solve for $\mathbf{m}$ in SCA iteration k, we linearize the nonconvex function g_0 using previous SCA iterate $\mathbf{m}_k^{(i)}$ of $\mathbf{m}$ as

$$\hat{g}_o(\mathbf{m}, \mathbf{w}_*^{(i-1)}; \mathbf{m}_k^{(i)}) = g_0(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}) + \nabla g_0(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)})^{\mathrm{T}}(\mathbf{m} - \mathbf{m}_k^{(i)}). \tag{47}$$

Let $\mathcal{X}_k^{(i)}$ be the feasible set for the ith AO iteration and the kth SCA point for a fixed $\mathbf{w}_*^{(i-1)}$ and $\mathbf{m}_k^{(i)}$. Similarly, $\mathcal{Y}_k^{(i)}$

denotes the feasible set for a fixed $\mathbf{m}_*^{(i)}$ and $\mathbf{w}_k^{(i)}$. Using (47), the convex subproblem for the ith AO iteration and the kth SCA point for variables $\mathbf{m}$ and $\boldsymbol{\gamma}$ is given by

$$\underset{\mathbf{m},\boldsymbol{\gamma}}{\text{minimize}} \quad f(\mathbf{m}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}) \tag{48a}$$

$$\text{subject to} \quad h(\boldsymbol{\gamma}) - \hat{g}_0(\mathbf{m}, \mathbf{w}_*^{(i-1)}; \mathbf{m}_k^{(i)}) \leq 0 \tag{48b}$$

$$g_1(\mathbf{m}, \mathbf{w}_*^{(i-1)}) \leq 0, \quad g_2(\mathbf{m}) \leq 0 \tag{48c}$$

The feasible set defined by the problem (48) is denoted by $\mathcal{X}_k^{(i)} \subset \mathcal{F}$. To prove the convergence of the SCA updates in the ith AO iteration, let us consider that (48) yields $\mathbf{m}_{k+1}^{(i)}$ and $\boldsymbol{\gamma}_{k+1}^{(i)}$ as the solution in the kth iteration. The point $\mathbf{m}_{k+1}^{(i)}$ and $\boldsymbol{\gamma}_{k+1}^{(i)}$, which minimizes the objective function satisfies

$$h(\boldsymbol{\gamma}_{k+1}^{(i)}) - g_0(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}) \leq h(\boldsymbol{\gamma}_{k+1}^{(i)}) - \hat{g}_0(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}; \mathbf{m}_k^{(i)}) \leq 0. \tag{49}$$

Using (49), we can show that $\{\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k+1}^{(i)}\}$ is feasible, since the initial SCA operating point $\mathbf{m}_*^{(i-1)}$ was chosen to be feasible from the $(i-1)$th AO iteration. In each SCA step, the feasible set includes the solution from the previous iteration as $\{\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k+1}^{(i)}\} \in \mathcal{X}_{k+1}^{(i)} \subset \mathcal{F}$, therefore, it decreases the objective as [22], [23], [30]

$$f(\mathbf{m}_0^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_0^{(i)}) \geq f(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)}) \geq f(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k+1}^{(i)}) \geq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}). \tag{50}$$

Thus, the sequence $\{f(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)})\}$ is nonincreasing. However, to ensure strict monotonicity of the objective sequence, we rely on the modified objective in (48a), which is strongly convex with some parameter $c > 0$. Now, to show strict monotonicity, let us consider $\mathbf{z}_k^{(i)} \triangleq [\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)}]$ as the minimizer for (48) in the $(k-1)$th SCA step and let $\mathbf{z}_{k+1}^{(i)}$ be the minimizer in the kth step. At the kth SCA iteration, $\forall \mathbf{z} \in \mathcal{X}_k^{(i)}$, it follows

$$\nabla f(\mathbf{z}_{k+1}^{(i)})^{\mathrm{T}}(\mathbf{z} - \mathbf{z}_{k+1}^{(i)}) \geq 0 \tag{51a}$$

$$f(\mathbf{z}) - f(\mathbf{z}_{k+1}^{(i)}) \geq c \left\| \mathbf{z} - \mathbf{z}_{k+1}^{(i)} \right\|^2 \tag{51b}$$

since $\mathbf{z}_{k+1}^{(i)}$ is the solution. Using (51) and $\mathbf{z}_k^{(i)} \in \mathcal{X}_k^{(i)}$, we can show that $f(\mathbf{z}_k^{(i)}) > f(\mathbf{z}_{k+1}^{(i)})$ holds in each SCA step unless $\mathbf{z}_k^{(i)} \to \mathbf{z}_*^{(i)}$ as $k \to \infty$. Now using the facts that (i) $\{f(\mathbf{z}_k^{(i)})\}$ is monotonic, and (ii) the uniqueness of the minimizer (see (51)), strict monotonicity of $\{f(\mathbf{z}_k^{(i)})\}$ is ensured [31]. Now, by using the fact that the objective sequence is strictly nonincreasing and bounded, we can ensure the convergence of objective sequence in each AO step i and $\mathbf{z}_*^{(i)} \triangleq \{\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}\}$ is the solution obtained upon convergence of the objective sequence $\{f(\mathbf{z}_k^{(i)})\}$ in the ith AO iteration.

Once $\mathbf{m}_*^{(i)}$ is obtained for fixed $\mathbf{w}$, then (44) is solved for $\mathbf{w}$ with fixed $\mathbf{m}$. However, after fixing $\mathbf{m}$ as $\mathbf{m}_*^{(i)}$ in (44), the problem is still nonconvex due to (44b). Following similar approach as above, the minimizer $\{\mathbf{m}_*^{(i)}, \mathbf{w}_{k+1}^{(i)}, \boldsymbol{\gamma}_{k+1}^{(i)}\}$ can be found in each SCA step k by solving (48) iteratively. Note that $\boldsymbol{\gamma}_{k+1}^{(i)}$ is reused since the variable $\mathbf{m}$ is fixed in the ith AO iteration. The convergence and the nonincreasing behavior of the objective follow similar arguments as above. [5] The solution obtained by solving (48) iteratively until convergence of the objective value is given as $\{\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i)}, \boldsymbol{\gamma}_{*|\mathbf{m}}^{(i)}\} \in \mathcal{Y}_*^{(i)} \subset \mathcal{F}$.

Finally, to prove the global convergence of the objective sequence, we need to show that the AO updates also produce a nonincreasing sequence of objectives, *i.e.*,

$$f(\mathbf{m}_*^{(i)}, \mathbf{w}_0^{(i)}, \boldsymbol{\gamma}_0^{(i)}) \leq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}). \tag{52}$$

Let $\mathbf{m}_*^{(i)}$ and $\boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}$ be the solution obtained by solving (48) iteratively until SCA convergence in the ith AO iteration for $\mathbf{m}$ and $\boldsymbol{\gamma}$ with fixed $\mathbf{w} = \mathbf{w}_*^{(i-1)}$. To find $\mathbf{w}_0^{(i)}$, we fix $\mathbf{m}$ as $\mathbf{m}_*^{(i)}$ and optimize for $\mathbf{w}$. Since the convex function is linearized in (44b), the fixed operating point is also included in the feasible set $\{\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}\} \in \mathcal{Y}_0^{(i)}$ by the equality in (49). Using this, we can show the monotonicity of the objective as

$$f(\mathbf{m}_*^{(i)}, \mathbf{w}_0^{(i)}, \boldsymbol{\gamma}_0^{(i)}) \leq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{m}}^{(i)}). \tag{53}$$

The AO update follows $\{\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|\mathbf{w}}^{(i)}\} \in \{\mathcal{X}_*^{(i)} \cap \mathcal{Y}_0^{(i)}\}$.

Using (51), strict monotonicity of the objective (46b) while alternating the optimization variables from $\mathbf{m}, \boldsymbol{\gamma}$ to $\mathbf{w}, \boldsymbol{\gamma}$ in the ith AO update can also be ensured. It follows from that fact that $\mathbf{z}_*^{(i)}$ is included in the feasible set $\mathcal{Y}_0^i$, and therefore the objective function follows (53) strictly in each AO update.

D. Stationarity of Limit Points

In this section, we discuss the convergence of the sequence of iterates generated by the iterative algorithm. In order to do so, let us use single global index t to represent both SCA indexing k and AO index i. Using this, we can represent $\mathbf{x}^t \triangleq [\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k|\mathbf{w}}^{(i)}]$ as a stacked vector of solution obtained in the kth SCA step of the ith AO update with fixed $\mathbf{w}$. The index t is used to represent both SCA updates involving transmit and receive beamformers within each AO step.

Using the discussions in Appendices A-A and A-C, we can show that the sequence of objective values $\{f(\mathbf{x}^t)\}$ converges. Note that similar claim cannot be made on the convergence of $\{\mathbf{x}^t\}$, since the sequence need not be monotonic. However, due to the compactness of the feasible set, the sequence of iterates $\{\mathbf{x}^t\}$ is bounded. Therefore, there exists at least one subsequence that converges to a limit point in the feasible set [27, Th. 3.6]. Assuming that there exists two convergent subsequences of $\{\mathbf{x}^t\}$, say, $\{\mathbf{x}^{t_i} \,|\, i = 0, 1, \dots\}$ and $\{\mathbf{x}^{t_k} \,|\, k = 0, 1, \dots\}$ that converges to different limit points, such that

$$\lim_{t\to\infty} f(\mathbf{x}^t) = \lim_{i\to\infty} f(\mathbf{x}^{t_i}) = \lim_{k\to\infty} f(\mathbf{x}^{t_k}) \tag{54}$$

therefore, $\{\mathbf{x}^t\}$ need not be a convergent sequence.

Since $\{\mathbf{x}^t\}$ need not be a convergent sequence, we can only claim that every limit point of the sequence $\{\mathbf{x}^t\}$ is a stationary point, *i.e.,* the limit point of every convergent subsequence is a stationary point. To prove that, let us consider a subsequence $\{\mathbf{x}^{t_j} \,|\, j = 0, 1, \dots\}$ of $\{\mathbf{x}^t\}$ that converges to a limit point $\bar{\mathbf{x}}$,

[5] Note that the MMSE receiver in (23b) can also be used instead of performing the SCA updates until convergence for the optimal receiver.

where $\bar{\mathbf{x}} = \lim_{j\to\infty} \mathbf{x}^{t_j}$. It follows from (50), (51), and (53) that

$$f(\mathbf{x}^{t_{j+1}}) < f(\mathbf{x}^{t_j}) < f(\mathbf{x}^{t_{j-1}}). \tag{55}$$

Eq. (55) holds strictly while alternating the optimization variables as shown in (53). Now, by taking the limit as $j \to \infty$ and by using the continuity of objective function f, we obtain

$$\lim_{j\to\infty} f(\mathbf{x}^{t_j}) = f(\lim_{j\to\infty} \mathbf{x}^{t_j}) = f(\bar{\mathbf{x}}). \tag{56}$$

Using strict monotonicity in $\{f(\mathbf{x}^{t_j})\}$ and the fact that as $j \to \infty$, $|f(\mathbf{x}^{t_j}) - f(\mathbf{x}^{t_{j+1}})| \to 0$, we can show that limit point $\bar{\mathbf{x}}$ satisfies the optimality condition [32, Prop. 2.1.2] as

$$\nabla f(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geq 0, \ \forall \mathbf{x} \in \mathcal{Q} \subset \mathcal{F} \tag{57}$$

where $\mathcal{Q}$ is a convex subset defined by (48b) and (48c) as $k \to \infty, i \to \infty$. Hence, $\bar{\mathbf{x}}$ is a stationary point.

Now to show that $\bar{\mathbf{x}}$ is a stationary point of original nonconvex problem (44), we show the equivalence between the gradient of the subproblem (48) and (44) at stationary point $\bar{\mathbf{x}}$. In order to show that $\bar{\mathbf{x}}$ is a KKT point for (44), it must satisfy the gradient condition of (44), *i.e.*,

$$\nabla \hat{f}(\bar{\mathbf{x}}) + \mu_0 \big[\nabla h(\bar{\mathbf{x}}) - \nabla g_0(\bar{\mathbf{x}})\big] + \textstyle\sum_{i=1}^{2} \mu_i \nabla g_i(\bar{\mathbf{x}}) = 0 \tag{58}$$

for some Lagrange multipliers $\mu_i \geq 0$ and feasible as $\bar{\mathbf{x}} \in \mathcal{F}$.

To prove the feasibility of the limit point $\bar{\mathbf{x}}$, note that in each SCA step, the feasible sets $\mathcal{X}_k^{(i)}, \mathcal{Y}_k^{(i)} \subset \mathcal{F}$. Therefore, the limit point $\bar{\mathbf{x}} \in \mathcal{F}$. Since $\bar{\mathbf{x}}$ satisfies Slater's constraint qualifications, there exist Lagrange multipliers μ_i such that

$$\nabla f(\bar{\mathbf{x}}) + \mu_0 \big[\nabla h(\bar{\mathbf{x}}) - \nabla \hat{g}_0(\bar{\mathbf{x}}; \bar{\mathbf{x}})\big] + \textstyle\sum_{i=1}^{2} \mu_i \nabla g_i(\bar{\mathbf{x}}) = 0. \tag{59}$$

The relation between (58) and (59) is evident by the following facts: (i) The quadratic term $\|\mathbf{x} - \mathbf{x}^{t_j}\|^2$ in (46b) vanishes as $j \to \infty, \mathbf{x}^{t_j} \to \bar{\mathbf{x}}$, since we assume that $\bar{\mathbf{x}}$ is the limit point of convergent subsequence $\{\mathbf{x}^{t_j}\}$. Therefore, the gradient evaluated at $\bar{\mathbf{x}}$ satisfies $\nabla f(\bar{\mathbf{x}}) = \nabla \hat{f}(\bar{\mathbf{x}})$. (ii) Additionally, by using the relation in (47), the gradient of the constraint (44b) evaluated at $\bar{\mathbf{x}}$ is given as

$$\nabla h(\bar{\mathbf{x}}) - \nabla \hat{g}_o(\bar{\mathbf{x}}; \bar{\mathbf{x}}) = \nabla h(\bar{\mathbf{x}}) - \nabla g_o(\bar{\mathbf{x}}). \tag{60}$$

Now, by applying the relation (60) and $\nabla f(\bar{\mathbf{x}}) = \nabla \hat{f}(\bar{\mathbf{x}})$ in (59), we can show that the limit point $\bar{\mathbf{x}}$ satisfies (58). By using [22, Thms. 2 and 11] and [30, Prop. 3.2], we can show that every limit point of the sequence of iterates $\{\mathbf{x}^t\}$ generated by the algorithm is a stationary point of problem (44).

B Convergence Proof for Distributed Algorithms

The convergence of the distributed algorithm in Algorithm 2 follows the same discussion as the one in Appendix A, if the subproblem in (31) converges to the centralized solution. Note that (31) satisfies the Slater's constraint qualification by having a non-empty interior and a compact set, which are required for the convergence. Since in primal decomposition, the master subproblem uses subgradient to update the coupling interference vectors in consensus with the objective function, the convergence of the subproblem is guaranteed as $i \to \infty$ for a diminishing step size as shown in [22], [33, Prop 8.2.6].

To show the convergence of the ADMM, we use the discussion in [34, Prop. 4.2] and [10] by writing (37) as

$$\underset{\mathbf{x}\in\mathcal{C}_1, \mathbf{z}\in\mathcal{C}_2}{\text{minimize}} \quad G(\mathbf{x}) + H(\mathbf{z}) \tag{61a}$$

$$\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{z} \tag{61b}$$

where the constraint in (61b) is identical to that in (38) used in the ADMM subproblem (37). To show the convergence, we rely on the following conditions. (i) G, H should be convex. (ii) $\mathcal{C}_1, \mathcal{C}_2$ should be a convex set and bounded. (iii) $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ should be invertible. It is evident from the equality constraints (61b) and (38) that $\mathbf{A} = \mathbf{I}$, and therefore it is invertible. Note that the objective functions G, H include ℓ_q norm and an additional quadratic term as in (46b), thereby exhibiting strong convexity. Moreover, the feasible set defined by the constraints of (31) is convex and has a nonempty interior. Now, by using [34, Prop. 4.2], we can show that the ADMM converges to the centralized solution as $i \to \infty$ by using diminishing step size in each ADMM iteration as discussed in [10]. Therefore, if both the primal and ADMM approaches are allowed to converge to the centralized solution in each SCA step, then every limit point of the sequence generated by the distributed schemes is a stationary point of (44), by following similar arguments as in Appendix A.

Unlike the primal or the ADMM methods, the decomposition approach via KKT conditions, presented in Section IV-C, updates all the optimization variables at once, *i.e*, the SCA update of $\epsilon^{(i-1)}$, the AO update of $\mathbf{w}_{l,k,n}$ and the dual variable update of α using subgradient method. Therefore, it is difficult to theoretically prove the convergence of the algorithm to a stationary point of the nonconvex problem in (16).

The algorithm in (42) is identical to (28), if the receivers $\mathbf{w}_{l,k,n}$ and the MSE operating point $\epsilon_{l,k,n}^{(i-1)}$ are fixed to find the optimal transmit precoders $\mathbf{m}_{l,k,n}$ and the dual variable $\alpha_{l,k,n}$. Note that it requires four nested loops to obtain the centralized solution, namely, the receive beamformer loop, MSE operating point loop, the dual variable update loop and the bisection method to find the transmit precoders.

However, to avoid the nested iterations, the proposed method performs group update of all the variables at once to obtain the transmit and the receive beamformers with the limited number of iterations, thus, achieving improved speed of convergence. Since the optimization variables are updated together, it is theoretically difficult to prove the monotonicity of the objective in each block update. Therefore, the convergence of the sequence of iterates generated by the Algorithm 3 is not guaranteed. Moreover, the objective function used to obtain the iterative algorithm is not strongly convex, and therefore the uniqueness of the iterates is not guaranteed in each update. Thus, the iterative scheme outlined in Algorithm 3 cannot be guaranteed to find a fixed point to ensure the convergence of the beamformer iterates.

C KKT conditions for MSE approach

To design the transmit precoders iteratively, we solve the KKT conditions of problem (41) by assuming (41b) and (41c)

hold with equality. Hence, we obtain the following equations.

$$\nabla_{\epsilon_{l,k,n}} : -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \tag{62a}$$

$$\nabla_{t_{l,k,n}} : -q\, a_k \Big(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\Big)^{(q-1)} + \frac{\sigma_{l,k,n}}{\log_2(e)} = 0 \tag{62b}$$

$$\nabla_{\mathbf{m}_{l,k,n}} : \sum_{y\in\mathcal{U}}\sum_{x=1}^{L} \alpha_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_k,y,n}\mathbf{w}_{x,y,n}\mathbf{w}^{\mathrm{H}}_{x,y,n}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n} + \delta_b\mathbf{m}_{l,k,n} = \alpha_{l,k,n}\mathbf{H}^{\mathrm{H}}_{b_k,k,n}\mathbf{w}_{l,k,n} \tag{62c}$$

$$\nabla_{\mathbf{w}_{l,k,n}} : \sum_{(x,y)\neq(l,k)} \mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}^{\mathrm{H}}_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_y,k,n}\mathbf{w}_{l,k,n} + N_0\mathbf{I}_{N_R}\mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}. \tag{62d}$$

In addition to the primal constraints in (41b), (41c) and (41d), the complementary slackness criterion must also be satisfied for any stationary point. Upon solving the above expressions in (62) with the complementary slackness conditions, we formulate an algorithm to determine the transmit and the receive beamformers iteratively as shown in (42).

REFERENCES

[1] E. Matskani, N. Sidiropoulos, Z.-Q. Luo, and L. Tassiulas, "Convex Approximation Techniques for Joint Multiuser Downlink Beamforming and Admission Control," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2682–2693, July 2008.

[2] C. Ng and H. Huang, "Linear Precoding in Cooperative MIMO Cellular Networks with Limited Coordination Clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, December 2010.

[3] L.-N. Tran, M. Hanif, A. Tölli, and M. Juntti, "Fast Converging Algorithm for Weighted Sum Rate Maximization in Multicell MISO Downlink," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 872–875, 2012.

[4] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE Transceiver Optimization for Multiuser MIMO Systems: Duality and Sum-MSE Minimization," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5436–5446, Nov 2007.

[5] S. S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted Sum-Rate Maximization using Weighted MMSE for MIMO-BC Beamforming Design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, 2008.

[6] 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, sept. 2011.

[7] M. Hong, Q. Li, Y.-F. Liu, and Z.-Q. Luo, "Decomposition by Successive Convex Approximation: A Unifying Approach for Linear Transceiver Design in Interfering Heterogeneous Networks," 2012.

[8] J. Kaleva, A. Tölli, and M. Juntti, "Decentralized Beamforming for Weighted Sum Rate Maximization with Rate Constraints," in *24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*. IEEE, 2013, pp. 220–224.

[9] D. P. Palomar and M. Chiang, "A Tutorial on Decomposition Methods for Network Utility Maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[11] H. Pennanen, A. Tölli, and M. Latva-Aho, "Decentralized Coordinated Downlink Beamforming via Primal Decomposition," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 647–650, 2011.

[12] A. Tölli, H. Pennanen, and P. Komulainen, "Decentralized Minimum Power Multi-Cell Beamforming with Limited Backhaul Signaling," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 570–580, 2011.

[13] M. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*, ser. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2010, vol. 3, no. 1.

[14] R. A. Berry and E. M. Yeh, "Cross-Layer Wireless Resource Allocation," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 59–68, 2004.

[15] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan 2007.

[16] K. Seong, R. Narasimhan, and J. Cioffi, "Queue Proportional Scheduling via Geometric Programming in Fading Broadcast Channels," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1593–1602, 2006.

[17] P. C. Weeraddana, M. Codreanu, M. Latva-aho, and A. Ephremides, "Resource Allocation for Cross-Layer Utility Maximization in Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 6, pp. 2790–2809, 2011.

[18] F. Zhang and V. Lau, "Cross-Layer MIMO Transceiver Optimization for Multimedia Streaming in Interference Networks," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1235–1244, March 2014.

[19] B. R. Marks and G. P. Wright, "A General Inner Approximation Algorithm for Nonconvex Mathematical Programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

[20] 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.

[21] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[22] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed Methods for Constrained Nonconvex Multi-Agent Optimization – Part I: Theory." [Online]. Available: http://arxiv.org/abs/1410.4754v1

[23] G. R. Lanckriet and B. K. Sriperumbudur, "On the Convergence of the Concave-Convex Procedure," in *Advances in Neural Information Processing Systems*, 2009, pp. 1759–1767.

[24] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2013.

[25] P. Komulainen, A. Tölli, and M. Juntti, "Effective CSI Signaling and Decentralized Beam Coordination in TDD Multi-Cell MIMO Systems," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2204–2218, 2013.

[26] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed Robust Multicell Coordinated Beamforming With Imperfect CSI: An ADMM Approach," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, June 2012.

[27] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill New York, 1964, vol. 3.

[28] G. Scutari, F. Facchinei, P. Song, D. Palomar, and J.-S. Pang, "Decomposition by Partial Linearization: Parallel Optimization of Multi-Agent Systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb 2014. [Online]. Available: http://arxiv.org/abs/1302.0756v2

[29] Y. Yang, G. Scutari, P. Song, and D. Palomar, "Robust MIMO Cognitive Radio Systems Under Interference Temperature Constraints," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2465–2482, November 2013.

[30] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.

[31] G. Scutari, D. P. Palomar, F. Facchinei, and J.-S. Pang, "Convex optimization, game theory, and variational inequality theory," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 35–49, 2010.

[32] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, sep 1999.

[33] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar *et al.*, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[34] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall Englewood Cliffs, NJ, 1989, vol. 23.