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

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

***Abstract*—We consider a downlink multi-cell multiple-input multiple-output (MIMO) interference broadcast channel (IBC) scenario using orthogonal frequency division multiplexing (OFDM) with multiple-user contending for space-frequency resources in a given scheduling instant. The problem is to determine the transmit precoders by the base stations (BSs) in a coordinated approach to minimize the total number of backlogged packets in the BSs, which are destined for the users in the system. Traditionally, it is solved using weighted sum rate maximization (WSRM) objective with the number of backlogged packets as the corresponding weights, *i.e*, longer the queue size, higher the priority. In contrast, we design the precoders jointly across the space-frequency resources by minimizing the total user queue deviations. The problem is nonconvex and therefore we employ successive convex approximation (SCA) technique to solve the problem by a sequence of convex subproblems using first order Taylor approximations. At first, we propose a centralized joint space-frequency resource allocation (JSFRA) solution using two different formulations by employing SCA technique, namely the sum rate formulation and the mean squared error (MSE) reformulation. We then introduce distributed precoder designs using primal and alternating directions method of multipliers method for the JSFRA solutions. Finally, we propose a practical distributed iterative precoder design based on MSE reformulation approach by solving the Karush-Kuhn-Tucker conditions with closed form expressions. Numerical results are used to compare the proposed algorithms with 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 (MUs), the main driving factor for the transmission are the packets waiting at each BS corresponding to the different users present in the network. These available packets are transmitted over the shared wireless resources subject to certain system limitations and constraints. 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. Since the space-frequency resources are shared by multiple users associated with different BSs, it can be viewed as a resource allocation problem.

In general, the resource allocation problems are formulated by assigning a binary variable for each user to indicate the presence or the absence in a particular resource [1]. In contrast, the linear transmit precoders, which are complex vectors, are implicitly used as decision variables, thereby avoiding the explicit modeling of binary decision variables. It solves two purposes. First, the formulation determines the transmission rate on a certain resource, and, secondly, by making the transmit beamformer of a particular user to be a zero vector, the corresponding user will not be scheduled on a certain 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 determined by the number of backlogged packets for each user in the system. 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, the successive convex approximation (SCA) method has become a powerful tool to deal with these problems [2]. For example, in [3], the nonconvex part of the objective has been linearized around an operating point in order to solve the WSRM problem in an iterative manner. Similar approach of solving the WSRM problem by using arithmetic-geometric inequality has been proposed in [4].

The connection between the achievable capacity and the mean squared error (MSE) for the received symbol by using the fixed minimum mean squared error (MMSE) receivers as shown in [5], [6] can also be used to solve the WSRM problem. In [6], [7], 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 decoders. Then the problem is solved using an alternating optimization method, i.e., finding a subset of variables while the remaining others are fixed. The MSE reformulation for the WSRM problem has also been studied in [8] by using the SCA to solve the problem in an iterative manner. Additional rate constraints based on the quality of service (QoS) requirements were included in the WSRM problem and solved via MSE reformulation in [9], [10].

The problem of precoder design for the MIMO IBC system are solved either by using a centralized controller or

This work has been supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia Solutions Networks, Xilinx Ireland, Academy of Finland. Part of this work has been published in ICASSP 2014 conference.

The authors are with the Centre for Wireless Communications (CWC), Department of Communications Engineering (DCE), University of Oulu, Oulu, FI-90014, (e-mail: {ganesh.venkatraman, antti.tolli, le.nam.tran, markku.juntti}@ee.oulu.fi).

by using decentralized algorithms where each BS handles the corresponding subproblem independently with the limited information exchange with the other BSs via back-haul. The distributed approaches are based on primal, dual or alternating directions method of multipliers (ADMM) decomposition, which has been discussed in [11], [12]. 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 are then updated by using the subgradient method as discussed in [13]. The dual and ADMM approaches control the distributed subproblems by fixing the *'interference price'* for each BS as detailed in [14].

By adjusting the weights in the WSRM objective properly, we can find an arbitrary rate-tuple in the rate region that maximizes the 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 an average among the users. As an approximation, we may assign weights based on the current queue size of the users. More specifically, the queue states can be incorporated to traditional weighted sum rate 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 [15], [16]. In backpressure algorithm, the differential queues between the source and the destination nodes are used as the weights scaling the transmission rate [17].

Earlier studies on the queue minimization problem were summarized in the survey paper [18], [19]. In particular, the problem of power allocation to minimize the number of backlogged packets was considered in [20] using geometric programming. Since the problem addressed in [20] 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 [21] by formulating the corresponding user queues as the weights in the WSRM problem. Recently, the precoder design for the video transmission over MIMO system is considered in [22]. In this design, the MU-MIMO precoders are designed by the MSE reformulation as in [6] with the higher layer performance objective such as playback interruptions and buffer overflow probabilities.

Main Contributions: In this paper, we design the precoders jointly across space-frequency resources by minimizing the total number of backlogged packets waiting at the BSs. Since the problem is nonconvex, we adopt the SCA technique to solve by a sequence of convex subproblems using first order approximations. Initially, we propose centralized joint space-frequency resource allocation (JSFRA) algorithms, which employs the SCA technique for the nonconvex constraints. First method is by using the direct formulation and the second one is by using the MSE equivalence with the rate expression to solve for an optimal precoders. Then we propose distributed precoder designs based on the primal and the ADMM methods. Finally, we propose a iterative practical algorithm to decouple the precoder design across the coordinating BSs with limited information exchange by solving the Karush-Kuhn-Tucker (KKT) conditions for the MSE reformulation solution.

The paper is organized as follows. In Section II, we introduce the system model and the problem formulation for the queue minimizing precoder design. The existing and the proposed centralized precoder designs are presented in Section III. The distributed solutions are provided in Section IV followed by the simulation results in Section V. 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 $b_k \in \mathcal{B}$. We denote by $\mathcal{N} = \{1, 2, \ldots, N\}$ 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 l^{th} spatial stream over the 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 symbol estimate corresponding to the l^{th} spatial stream over sub-channel n at user k is given by

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

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 $\mathbf{n}_{k,n} \sim \mathcal{CN}(0, N_0)$ is the additive noise vector for the user k on the n^{th} sub-channel and l^{th} spatial stream. In (1), $L = \text{rank}(\mathbf{H}_{b,k,n}) = \min(N_T, N_R)$ 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

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

where $\widetilde{N}_0 = N_0 \|\mathbf{w}_{l,k,n}\|^2$ denotes the equivalent noise variance. In order to minimize the feedback overhead on the user channels, we assume time division duplexing (TDD) system.

Let Q_k be the number of backlogged packets destined for the user k at a given scheduling instant. The queue dynamics of the user k are modeled using the Poisson arrival process with the average number of packet arrivals of $A_k = \mathbf{E}_i\{\lambda_k\}$ packets/bits, where $\lambda_k(i) \sim \text{Pois}(A_k)$ represents the instantaneous number of packets arriving for the user k at the i^{th} time

[1] L streams are initialized but after solving the problem, only $L_{k,n} \leq L$ non-zero data streams are transmitted

instant[2]. The total number of queued packets at the $(i+1)^{\text{th}}$ instant for the 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 i^{th} instant, transmission rate of the 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 l^{th} spatial stream on the n^{th} sub-channel. The maximum rate achieved over the (l, n) space-frequency resource is given by $t_{l,k,n} \le \log_2(1+\gamma_{l,k,n})$ for the signal-to-interference-plus-noise ratio (SINR) of $\gamma_{l,k,n}$[3]. Note that the units of t_k and Q_k are 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 given by

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

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

Explicitly, the objective of the problem considered is given by $\sum_{k\in\mathcal{U}} a_k |v_k|^q$. With this objective function, the weighted queued packet minimization formulation is given by

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

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

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 incorporated to control user priority based on their respective QoS. Let $\mathbf{M}_{k,n} \triangleq [\,\mathbf{m}_{1,k,n}\,\mathbf{m}_{2,k,n} \ldots \mathbf{m}_{L,k,n}\,]$ be the matrix formed by stacking the beamformers associated with user k for n^{th} sub-channel transmission. Similarly, let $\mathbf{W}_{k,n} \triangleq [\,\mathbf{w}_{1,k,n}\,\mathbf{w}_{2,k,n} \ldots \mathbf{w}_{L,k,n}\,]$ denotes the matrix formed by stacking the receive beamformers respectively[4]. In (6b), we consider a BS specific sum power constraint for each BS across all sub-channels.

For practical reasons, we may impose a constraint that the maximum number of transmitted bits for the user k 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

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

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

Before proceeding further, we show that the constraint in (7) is handled implicitly by the definition of norm ℓ_q in the objective of (6). Suppose that $t_k > Q_k$ for 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 smaller objective which contradicts with the fact that the optimal solution has been obtained. The choice of the norm ℓ_q used in the objective function [18], [20] alters the priorities for the queue deviation function as follows

- ℓ_1 results in greedy allocation *i.e*, emptying the queue of users with good channel conditions 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 prioritizes users with 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 [16].
- ℓ_∞ minimizes the maximum number of queued packets among users with the current transmission, thereby providing queue fairness by allocating the 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 highly difficult due to the nonconvex nature of the problem. In addition, the objective of minimizing the number of the queued packets over the spatial and the sub-channel dimensions adds further complexity to the existing problem. Since the scheduling of users in each sub-channel can be made 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 the existing algorithm to minimize the number of backlogged packets with additional constraints required by the problem.

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

The queue minimizing algorithms are discussed extensively in the networking literature to provide congestion-free routing between any two nodes in the network. One such algorithm is the *backpressure algorithm* [15]–[17]. It determines an optimal control policy in the form of rate or resource allocation for the nodes in the network by considering the differential backlogged packets between the source and the destination

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

[3]Upper bound is achieved by using Gaussian signaling

[4]It can be easily extended for user specific streams $L_{k,n}$ instead of using the common L streams for all users

nodes. Even though the algorithm is primarily designed for the wired infrastructure, it can be extended to the wireless networks by designing the user rate variable t_k in accordance to the wireless network.

The queue weighted sum rate maximization (Q-WSRM) formulation extends the *backpressure algorithm* to the downlink MIMO-OFDM framework, in which the multiple BSs act as the source nodes and the user terminals as the receiver nodes. The control policy in the form of transmit precoders aims at minimizing the number of queued packets waiting in the BSs. In order to find the optimal strategy, we resort to the Lyapunov theory, which is predominantly used in the 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 of each user, the Lyapunov function is used to provide a scalar measure, which grows large when the system moves toward the undesirable state. Following similar approach as in [16], the scalar measure for the 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 $\mathbf{Q}(i) = [Q_1(i), Q_2(i), \ldots, Q_K(i)]^{\mathrm{T}}$ and $\frac{1}{2}$ is used for the convenience. It provides a scalar measure of congestion present in the system [16, Ch. 3].

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

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

In order to eliminate the nonlinear operator $[x]^+$, we bound the expression in (9) as

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

by using the following inequality

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

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

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

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

where the subscripts $\mathbf{t}$ and $\boldsymbol{\lambda}$ represents the vector formed by stacking the transmission and the arrival rate of all users in the system. Since the transmission and the arrival rates are bounded, the second order moments in the first term of (12b) can be bounded by a constant B without affecting the optimal solution of the problem [16]. The second term in (12b) follows from the Poisson arrival process.

The expression in (12) looks similar to the WSRM formulation if the weights in the WSRM problem are replaced by the number of backlogged packets corresponding to the users. The above discussed approach is extended for the wireless networks in [21], where the queue weighted sum rate maximization is considered as the objective function to determine the transmit precoders. Since the expectation can be minimized by minimizing the function inside, the Q-WSRM formulation is given by

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

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

In order to avoid the excessive allocation of the resources, we include an additional rate constraint $t_k \le Q_k$ 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 constraint

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

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 the more backlogged packets, which essentially does the allocation in a greedy manner.

As a special case of the problem defined in (13), we can formulate the sum rate maximization problem by setting the weights in (13a) as unity, leading to the problem as in (13) with $Q_k = 1, \forall k \in \mathcal{U}$. This approach provides a greedy queue minimizing allocation as compared to Q-WSRME, since the resource allocation is driven by the channel conditions in comparison to the number of queued packets as in Q-WSRME. Note that in both formulations, the resources allocated to the users are limited by the number of backlogged packets with an explicit maximum rate constraint defined by (14).

B. JSFRA Scheme via SCA approach

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, using $\ell_{q=2}$ in (5), we obtain the following objective

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

which is equivalent to the objective defined in (13). Note that the equivalence can be seen either by discarding t_k^2 from (15) or when the total number of queued packets is significantly greater than the maximum allowable transmission.

It is evident that (15) is equivalent to (12) if the second order terms are ignored. Limiting t_k^2 by a constant value, the Q-WSRM formulation requires the explicit rate constraint (14) to avoid the resource wastage in the form of over-allocation. In the proposed queue deviation formulation, the explicit rate constraint is not needed, since it is handled by the objective function itself. This makes the problem simpler and allows us to employ efficient algorithms to distribute the precoder design problem across each BSs independently by exchanging minimal information exchange [12]. In contrast to the WSRM formulation, the JSFRA and the Q-WSRM problems include the sub-channels jointly to achieve an efficient allocation by identifying the optimal space-frequency resource for each user. The queue deviation objective provides an alternative approach to perform the resource allocation without the additional rate constraints as in the Q-WSRME formulation.

We now present an algorithm to solve problem (6) by using alternating optimization technique in conjunction with successive convex approximation (SCA) [23]. To do this, first by using the SINR expression in (2), we equivalently reformulate problem (6) as

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

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

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

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

In this formulation, we relaxed the SINR expression (2) by the inequalities in (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 $\beta_{l,k,n}$. The constraints are tight when the per BS objective is non zero at the optimal point, as discussed in Appendix A. Note that the problem in (16) is similar to the WSRM formulation, which is known to be NP-hard even for the single antenna case [24].

In order to find a tractable solution for problem (16), we note that the constraints (16d) are convex with involved variables. Thus, we only need to deal with (16b) and (16c). Towards this end, we resort to the traditional coordinate descent technique by fixing the linear receivers, and finding the optimal transmit beamformers. For a fixed receivers $\mathbf{w}_{l,k,n}$, the problem now is to find the optimal transmit beamformers $\mathbf{m}_{l,k,n}$ which is still a challenging task. We note that for fixed $\mathbf{w}_{l,k,n}$, (16c) can be written as a second-order cone (SOC) constraint. Thus, the difficulty is due to the non-convexity in (16b). To arrive at a tractable formulation, we adopt SCA to handle (16b) by replacing the original non-convex constraint by a series of convex constraints [23]. Let us define a function,

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

where $\mathbf{u}_{l,k,n} \triangleq \{\mathbf{w}_{l,k,n}, \mathbf{m}_{l,k,n}, \beta_{l,k,n}\}$ is the vector which needs to be identified for the optimal allocation. Note that the function $f(\mathbf{u}_{l,k,n})$ is convex for 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 $\beta_{l,k,n}$ [25]. Eq. (16b) can be viewed as a difference of convex (DC) constraint, in which the convex function given by $f(\mathbf{u}_{l,k,n})$ is upper bounded by first order approximation around an operating point $\tilde{\mathbf{u}}_{l,k,n}$.

For this purpose, let the real and imaginary component 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

$$p_{l,k,n} \triangleq \Re\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{17a}$$

$$q_{l,k,n} \triangleq \Im\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{17b}$$

and hence $f(\mathbf{u}_{l,k,n}) = (p_{l,k,n}^2 + q_{l,k,n}^2)/\beta_{l,k,n}$[5]. Suppose that the current value of $p_{l,k,n}$ and $q_{l,k,n}$ at a specific iteration are $\tilde{p}_{l,k,n}$ and $\tilde{q}_{l,k,n}$, respectively. Using the first order Taylor approximation around the local point $[\tilde{p}_{l,k,n}, \tilde{q}_{l,k,n}, \tilde{\beta}_{l,k,n}]^T$, we can approximate (16b) by the following linear inequality constraint as

$$2\frac{\tilde{p}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(p_{l,k,n} - \tilde{p}_{l,k,n}) + 2\frac{\tilde{q}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(q_{l,k,n} - \tilde{q}_{l,k,n}) + \frac{\tilde{p}_{l,k,n}^2 + \tilde{q}_{l,k,n}^2}{\tilde{\beta}_{l,k,n}}\left(1 - \frac{\beta_{l,k,n} - \tilde{\beta}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\right) \geq \gamma_{l,k,n}. \tag{18}$$

In summary, for the fixed linear receivers $\mathbf{w}_{l,k,n}$, the JSFRA problem to find transmit beamformers is shown by

$$\underset{\substack{\mathbf{m}_{l,k,n},\\ \gamma_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{19a}$$

$$\text{subject to} \quad \beta_{l,k,n} \geq \widetilde{N}_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{19b}$$

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}, \forall b, \tag{19c}$$

$$\text{and (18).} \tag{19d}$$

Now, the optimal linear receivers for the fixed transmit precoders $\mathbf{m}_{j,i,n}\, \forall i \in \mathcal{U},\, \forall n \in \mathcal{C}$ are obtained by minimizing (6) with respect to $\mathbf{w}_{l,k,n}$ as

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

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

$$\text{and (18).} \tag{20c}$$

Solving (20) using the KKT conditions, we obtain the following iterative expression for the receiver $\mathbf{w}_{l,k,n}^{*}$ as

$$\mathbf{A}_{l,k,n} = \sum_{(j,i)\neq(l,k)} \mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\mathbf{m}_{j,i,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\,\mathbf{I}_{N_R} \tag{21a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \left(\frac{\tilde{\beta}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)}}{\|\mathbf{w}_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\|^2}\right)\mathbf{A}_{l,k,n}^{-1}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}, \tag{21b}$$

where $\mathbf{w}_{l,k,n}^{(i-1)}$ is the receive beamformer from the previous

[5]Note that $p_{l,k,n}$ and $q_{l,k,n}$ are just symbolic notation and not the newly introduced optimization variables. In CVX [26], for example, we declare $p_{l,k,n}$ and $q_{l,k,n}$ with the '*expression*' qualifier

iteration, upon which the linear relaxation is performed for the nonconvex constraint in (20). The optimal receiver $\mathbf{w}^*_{l,k,n}$ is obtained by either iterating (21b) until convergence or for fixed number of iterations. Note that the receiver has no explicit relation with the choice of ℓ_q norm used in the objective function. The dependency is implicitly implied by the transmit precoders $\mathbf{m}_{l,k,n}$, which in deed depend on the q value.

It can be seen that the optimal receiver in (21b) is in fact a scaled version of the MMSE receiver, which is given by

$$\mathbf{R}_{l,k,n} = \sum_{i\in\mathcal{U}}\sum_{j=1}^{L}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\mathbf{m}^{\mathrm{H}}_{j,i,n}\mathbf{H}^{\mathrm{H}}_{b_i,k,n} + N_0\,\mathbf{I}_{N_R} \quad (22a)$$

$$\mathbf{w}_{l,k,n} = \mathbf{R}^{-1}_{l,k,n}\;\mathbf{H}_{b_k,k,n}\;\mathbf{m}_{l,k,n}. \quad (22b)$$

Since the scaling present in the optimal receiver (21b) has no impact on the received SINRs, the MMSE receiver in (22b) can also be used without compromising the performance or the convergence behavior.

The proposed algorithm is referred to as queue minimizing JSFRA scheme with a per BS power constraint, and it is outlined in Algorithm 1. The iterative procedure repeats until the improvement on the objective is less than a predetermined tolerance parameter or the maximum number of iterations is reached. Instead of initializing $\mathbf{u}_{l,k,n}$ arbitrarily to a feasible point, transmit precoders can also be initialized with any feasible point $\tilde{\mathbf{m}}_{l,k,n}$, which is then used to find $\mathbf{u}_{l,k,n}$ as briefed in Algorithm 1. For a fixed receive beamformer $\mathbf{w}_{l,k,n}$, the SCA iteration is carried out until convergence or for the predefined iterations, say, $J_{\max}$ for the optimal transmit precoders $\mathbf{m}_{l,k,n}$. Next, the receive beamformers are updated based on either (21b) or (22b) using the fixed transmit precoders $\mathbf{m}_{l,k,n}$. This procedure is carried out until convergence of the queue deviation or for fixed number of iterations by $I_{\max}$ as outlined in Algorithm 1.

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: $i=0$ and transmit precoders $\tilde{\mathbf{M}}_{k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{W}_{k,n}$ and $\mathbf{u}_{l,k,n}$ using (22b) and (18) using $\tilde{\mathbf{M}}_{k,n}$
repeat
  initialize $j=0$
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (19)
    update the constraint set (18) with $\mathbf{u}_{l,k,n}$ and $\mathbf{m}_{l,k,n}$ using (17)
    $j=j+1$
  until *SCA convergence or* $j\geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (20) or (22b) with the updated precoders $\mathbf{m}_{l,k,n}$
  $i=i+1$
until *Queue convergence or* $i\geq I_{\max}$

C. JSFRA Scheme via MSE Reformulation

In this section, we solve the JSFRA problem by exploiting the equivalence between the MSE and the achievable sum rate for the receivers designed based on the MMSE criterion [5]. The MSE $\epsilon_{l,k,n}$, for the data symbol is given by

$$\epsilon_{l,k,n} = \mathbb{E}\big[(d_{l,k,n}-\hat{d}_{l,k,n})^2\big] = \big|1-\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\big|^2 + \sum_{(j,i)\neq(l,k)}\big|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\big|^2 + \tilde{N}_0, \quad (23)$$

where $\hat{d}_{l,k,n}$ is the estimate of the transmitted symbol. Using the MMSE receive beamformer (22b) in the MSE expression (23) and in the SINR expression (2), we can arrive at the following relation between the MSE and the SINR as

$$\epsilon_{l,k,n} = \frac{1}{1+\gamma_{l,k,n}}. \quad (24)$$

The above equivalence is valid only if the receivers are based on the MMSE criterion. Using the equivalence in (24), the WSRM objective can be reformulated as the weighted minimum mean squared error (WMMSE) equivalent to obtain the precoders for the MU-MIMO scenario as discussed in [6]–[8]. Note that the receiver is invariably based on the MMSE criterion irrespective of the ℓ_q norm used in the objective function to obtain the optimal transmit precoders $\mathbf{m}_{l,k,n}$.

Let $v'_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ denote the queue deviation corresponding to user k and $\tilde{v}'_k \triangleq a_k^{1/q} v'_k$ represents the weighted equivalent. By using the relaxed MSE expression in (23), the problem in (6) can be expressed as

$$\underset{\substack{t_{l,k,n},\mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \quad (25a)$$

$$\text{subject to } t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n}) \quad (25b)$$

$$\sum_{(j,i)\neq(l,k)}\big|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\big|^2 + \tilde{N}_0 + \big|1-\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\big|^2 \leq \epsilon_{l,k,n} \quad (25c)$$

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}^{\mathrm{H}}_{l,k,n}\right) \leq P_{\max}, \forall b, \quad (25d)$$

An alternative MSE formulation given by (25) is non-convex even for the fixed $\mathbf{w}_{l,k,n}$ due to the constraint (25b), which is in fact a DC constraint. We resort to the SCA approach [23] by relaxing the constraint by a sequence of convex subsets using first order Taylor series approximation around a fixed MSE point $\tilde{\epsilon}_{l,k,n}$ as

$$-\log_2(\tilde{\epsilon}_{l,k,n}) - \frac{(\epsilon_{l,k,n}-\tilde{\epsilon}_{l,k,n})}{\log(2)\,\tilde{\epsilon}_{l,k,n}} \geq t_{l,k,n}, \quad (26)$$

Using the above approximation for the rate constraint, the problem defined in (25) is solved for optimal transmit precoders $\mathbf{m}_{l,k,n}$, MSEs $\epsilon_{l,k,n}$, and the user rates over each sub-channel $t_{l,n,k}$ given the fixed receive beamformers. Once the optimal precoders are obtained, the local MSE variable $\tilde{\epsilon}_{l,k,n}$ is updated with the current update $\epsilon_{l,k,n}$. The optimization problem for a fixed receive beamformers $\mathbf{w}_{l,k,n}$ is given as

$$\underset{t_{l,k,n},\mathbf{m}_{l,k,n},\epsilon_{l,k,n}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \tag{27a}$$

$$\text{subject to} \quad (25c), (25d), \text{ and } (26). \tag{27b}$$

D. Reduced Complexity Resource Allocation

The complexity involved in the JSFRA scheme scales significantly with the increase in the number of sub-channels considered in the formulation. In addition to the increased complexity, the rate of convergence to the optimal precoders also degrades due to its dependency on the problem size. In order to mitigate this, we can adopt the decomposition techniques discussed in [11], [12] to distribute the precoder design across each sub-channels independently with some coupling parameters that are exchanged across the sub-channel wise subproblems.

As an alternative sub-optimal solution, we propose queue minimizing spatial resource allocation (SRA), which aims at designing the precoders independently across each sub-channel in a sequential manner by taking the current backlogged packets into account. In this approach, the transmit power is fixed across each sub-channel n as $P_{\max,n}$ at each BS. The power split can either follow equal sharing or any predetermined pattern as in partial frequency reuse. For the fixed sub-channel power, the JSFRA formulation presented in Section III-B or III-C is performed over each sub-channel in a sequential manner. The coupling variable across the sub-channel is the total number of backlogged packets associated with each user. Let $Q_{k,n}$ is the total number of queued bits used in the optimization problem carried out for the sub-channel n. Now, for the first sub-channel, $n = 1$, the queues are given by $Q_{k,n=1} = Q_k$, which corresponds to the actual number of backlogged packets present in the current instant i. Using the updated number of queued packets, the JSFRA algorithm is performed over the sub-channel $n = 1$ to obtain the rates associated with each user. Now, for $n = 2$, the number of queued packets for each user is updated as $Q_{k,n=2} = Q_k - \sum_{l=0}^{L} t_{l,k,n=1}$. With this updated number of queued packets, the JSFRA algorithm is performed for $n = 1$ sub-channel to obtain the optimal precoders. In general, the number of queued packets to be used for each user in a given sub-channel n is given by

$$Q_{k,n} = \max\left\{Q_k - \sum_{r=1}^{n-1}\sum_{l=1}^{L} t_{l,k,r}, 0\right\}, \; \forall\, k \in \mathcal{U}. \tag{28}$$

In (28), Q_k denotes the total number of queued bits waiting to be transmitted for the user k during the current slot and $t_{l,k,r}$ is the rate or guaranteed bits allocated over the sub-channel r. However, the proposed scheme is sensitive to the order in which the sub-channels are selected for the optimization problem.

The proposed queue minimizing (QM) SRA scheme depend on the choice of the sub-channel selection, which is evident from the sequential approach used in the sub-optimal problem formulation. However, the proposed approach provides faster convergence in contrast to the JSFRA formulation over all sub-channel mainly due to the reduction in the size of the sub-channel wise sub problem. The choice of the sub-channel selection is random and it is difficult to determine the order, since the optimization variables $\mathbf{m}_{l,k,n}$ are vectors.

IV. Distributed Solutions

This section addresses the distributed precoder designs for the proposed JSFRA scheme. The formulation in (19) or (27) requires a centralized controller to perform the precoder design for all users belonging to the coordinating BSs. In order to design the precoders independently at each BS with the minimal information exchange via backhaul, iterative decentralization methods are considered. In particular, the primal decomposition and the ADMM based dual decomposition approaches are addressed.

To begin with, let $\bar{\mathcal{B}}_b$ denote the set $\mathcal{B}\backslash\{b\}$ and $\bar{\mathcal{U}}_b$ represents the set $\mathcal{U}\backslash\mathcal{U}_b$. In order to study the decomposition based solutions, we consider the solution proposed in (19), which is based on the Taylor series approximation for the nonconvex constraint. The following discussions are equally valid for the MSE based solution outlined in (27) as well. Since the objective of (19) can be decoupled across each BS, the centralized problem can be equivalently written as

$$\underset{\gamma_{l,k,n},\mathbf{M}_{k,n},\mathbf{W}_{k,n},\beta_{l,k,n}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{29a}$$

$$\text{subject to} \quad (19\text{b}) - (19\text{d}), \tag{29b}$$

where $\tilde{\mathbf{v}}_b$ denotes the vector of of weighted queue deviation corresponding to users $k \in \mathcal{U}_b$.

Following similar approach as in [13], [14], the coupling constraint (19b) or (25c) can be expressed by grouping the interference contribution from each BS in $\mathcal{B}$ as

$$N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{j=1,j\neq l}^{L} |\mathbf{w}_{l,k,n}^{\rm H}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b}$$
$$+ \sum_{i\in\mathcal{U}_{b_k}\backslash\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\rm H}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 \leq \beta_{l,k,n}, \tag{30}$$

where $\zeta_{l,k,n,b}$ is the total interference caused by BS b to the $l^{\rm th}$ stream of user $k \in \mathcal{U}_{b_k}$ on the $n^{\rm th}$ sub-channel, is upper bounded by

$$\zeta_{l,k,n,b} \geq \sum_{i\in\mathcal{U}_b}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\rm H}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2, \forall b \in \bar{\mathcal{B}}_{b_k}. \tag{31}$$

The coupling variable $\beta_{l,k,n}$ can be decoupled using the variable $\zeta_{l,k,n,b}$, which limits the total interference caused by the transmission from BS b to the $l^{\rm th}$ spatial stream of user k over the sub-channel n. In order the solve for the global optimal precoders, we need to find the coupling variables $\zeta_{l,k,n,b}$ by either the primal or dual decomposition method. In both approaches, the coupling constraint (19b) for the SCA and (25c) for the MSE relaxation schemes are decoupled to perform the distributed precoder design problem.

A. Decomposition based Approaches

1) Primal Decomposition Approach: The primal decomposition approach decomposes the problem by fixing the interference variables $\zeta_{l,k,n,b} \forall k, b$ in order to perform the precoder design independently across each BS. Once the optimal precoders are designed at each BS with the fixed interference constraints (30), the dual variables corresponding to the interference constraints are exchanged between the cooperating BSs in $\mathcal{B}$ to update the interference variables $\zeta_{l,k,n,b}$ for the next iteration until convergence. The primal approach is discussed extensively for the min-power problem in [13] and much of the current work follows similar approach.

Convergence: The convergence of the primal decomposition is similar to that of the centralized problem if the interference variables $\zeta_{l,k,n,b}$ are allowed to converge to a stationary point. In practice, we can limit the number of exchanges to $J_{\max}$ after which the SCA update is performed until convergence or for $I_{\max}$ times. The update of $\tilde{p}_{l,k,n}, \tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ can be made in conjunction with the receiver update $\mathbf{W}_{k,n}$. The receiver update can be made by using the precoded pilot transmission from each user as in [27].

2) ADMM approach: The ADMM decomposition method is based on the dual decomposition, however it shows better convergence properties. In contrast to the primal decomposition problem, the ADMM method relaxes the interference constraints by including it in the objective function of each subproblem with a penalty pricing [11], [12]. In order to decouple the problem (29), the coupling variables $\zeta_{l,k,n,b}$ in (30) are replaced by the respective local copies $\boldsymbol{\zeta}^{\{b\}}, \forall b \in \mathcal{B}$, which are then solved for an optimal solution. Now the sub problems are coupled by the global consensus vector $\boldsymbol{\zeta}$ maintaining the complete stacked interference profile of all users in the system as

$$\begin{aligned}\boldsymbol{\zeta} &= \Big[\zeta_{1,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(|\bar{\mathcal{U}}_1|),1,1}, \\ &\quad \ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),1,N_B}, \ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),N,N_B}\Big] \quad (32a)\\ n_{b_k} &= |\boldsymbol{\zeta}^{\{b_k\}}| = NL \sum_{b \in \mathcal{B}} \left|\bar{\mathcal{U}}_b\right|. \quad (32b)\end{aligned}$$

Let $\boldsymbol{\zeta}(b_k)$ denote the consensus entries corresponding to BS b_k. Let $\boldsymbol{\nu}^{\{b_k\}}$ represent the stacked dual variables corresponding to the equality condition $\boldsymbol{\zeta}^{\{b_k\}} = \boldsymbol{\zeta}(b_k)$ used in the subproblems. In order to limit the local interference assumptions $\zeta^{\{b_k\}}_{l,k,n,b}$ in BS b_k, the ADMM method augments a scaled quadratic penalty of the interference deviation between the local and consensus value for the interference from the BS b as $\zeta_{l,k,n,b}$ in the objective function. At optimality, the locally assumed and the consensus interference values will be equal, providing no contribution to the objective function. The optimal step size used to update the dual variables is the scaling factor ρ used to scale the penalty term in the objective function [2], [12]. The equality constraint for the local and the consensus interference vector $\boldsymbol{\zeta}^{\{b_k\}} = \boldsymbol{\zeta}(b_k)$ present in each subproblem is relaxed by the taking the partial Lagrangian. Now, the subproblem at BS b for the i^{th} iteration is given by

$$\underset{\substack{\gamma_{l,k,n}, \mathbf{W}_{k,n}, \mathbf{M}_{k,n}\\ \beta_{l,k,n}, \boldsymbol{\zeta}^{\{b\}(i)}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}^{\{b\}(i-1)T}\left(\boldsymbol{\zeta}^{\{b\}(i)} - \boldsymbol{\zeta}^{(i-1)}(b)\right)$$

$$+\frac{\rho}{2}\Big\| \underbrace{\boldsymbol{\zeta}^{\{b\}(i)}}_{\text{local}} - \underbrace{\boldsymbol{\zeta}^{(i-1)}(b)}_{\text{consensus}} \Big\|_2^2 \qquad (33a)$$

subject to

$$\beta_{l,k,n} \geq \sum_{j=1, j\neq l}^{L} |\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{\hat{b}\in\bar{\mathcal{B}}_b} \zeta^{\{b\}(i-1)}_{l,k,n,\hat{b}}$$

$$+\sum_{i\in\mathcal{U}_b\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b,k,n}\mathbf{m}_{j,i,n}|^2 + \tilde{N}_0 \qquad (33b)$$

$$\zeta^{\{b\}(i)}_{l',k',n,b} \geq \sum_{k\in\mathcal{U}_{\hat{b}}} \sum_{l=1}^{L} |\mathbf{w}^{\mathrm{H}}_{l',k',n}\mathbf{H}_{b,k',n}\mathbf{m}_{l,k,n}|^2, \forall k' \in \bar{\mathcal{U}}_b \quad (33c)$$

$$\text{(18) and (6b),} \qquad (33d)$$

where the superscript i represents the current iteration or the information exchange index and $\boldsymbol{\zeta}^{(i-1)}$ denotes the updated global interference level from the $(i-1)^{\text{th}}$ information exchange of the local interference vector $\boldsymbol{\zeta}^{\{b\}(i-1)}, \forall b \in \mathcal{B}$.

Now, the local problem (33) at each BS b is solved either by the SCA approach discussed in Section III-B or by using the MSE reformulation approach outlined in Section III-C. Once the local problems are solved at each BS, the new update for the global interference vector $\boldsymbol{\zeta}^{(i)}$ and the dual variables $\boldsymbol{\nu}^{\{b\}(i)}$ are performed at each BS independently by exchanging the corresponding local copies of the interference vector $\boldsymbol{\zeta}^{\{b\}(i)}, \forall b \in \mathcal{B}$. Since the entries in $\boldsymbol{\zeta}^{(i)}$ relate exactly to two BSs only, each entry in $\boldsymbol{\zeta}^{(i)}$ can be updated by exchanging the local copies from the corresponding two BSs. For instance, the entry $\zeta^{(i)}_{l,\mathcal{U}_{b_k}(1),n,b}$ depends on the local interference value $\zeta^{\{b_k\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b}$ assumed by the BS b_k and the actual interference caused by BS b as in $\zeta^{\{b\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b}$ as

$$\zeta^{(i)}_{l,\mathcal{U}_{b_k}(1),n,b} = \frac{1}{2}\left(\zeta^{\{b\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b} + \zeta^{\{b_k\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b}\right). \qquad (34)$$

The dual variable vector $\boldsymbol{\nu}^{\{b_k\}}$, which includes the stacked dual variables of the interference equality constraint at BS b_k, are updated using the subgradient as

$$\nu^{\{b_k\}(i)}_{l,k,n,b} = \nu^{\{b_k\}(i-1)}_{l,k,n,b} + \rho\left(\zeta^{\{b_k\}(i)}_{l,k,n,b} - \zeta^{(i)}_{l,k,n,b}\right). \qquad (35)$$

The distributed precoder design using ADMM approach is shown in Algorithm 2.

Convergence: The convergence of the ADMM method follows the same argument as the centralized algorithm if each distributed algorithm is allowed to converge to a stationary value for the fixed SCA point. Since the subproblem solved at each BS is convex, the ADMM method converges to a stationary point [12] for the fixed SCA value. The receive beamformers are updated along with the SCA update of $\tilde{\mathbf{u}}_{l,k,n}$. Combining the receiver update with the SCA update improves the convergence speed due to the fact that the MMSE receivers are optimal for the fixed transmit beamformers, providing monotonic increase in the objective function.

B. 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

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 $\mathbf{w}_{l,k,n} \forall l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0$ and the transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying total power constraint (6b)
update $\mathbf{w}_{l,k,n}$ with (22b) and $\tilde{\mathbf{u}}_{l,k,n}$ with (18)
initialize the global interference vectors $\boldsymbol{\zeta}^{(0)} = \mathbf{0}^{\mathrm{T}}$
initialize the interference threshold $\boldsymbol{\nu}^{\{b\}(0)} \forall b \in \mathcal{B} = 0$
foreach *BS* $b \in \mathcal{B}$ **do**
  repeat
    initialize $j = 0$
    repeat
      solve for $\mathbf{M}_{k,n}$ and the local interference $\boldsymbol{\zeta}^{\{b\}}$ using (33)
      exchange $\boldsymbol{\zeta}^{\{b\}(j)}$ among BSs in $\mathcal{B}$
      update dual variables in $\boldsymbol{\nu}^{\{b\}(j+1)}$ using (35)
      update the consensus vector $\boldsymbol{\zeta}^{(j+1)}$ using (34)
      $j = j + 1$
    until *convergence or* $j \geq J_{\max}$
    downlink precoded pilot transmission with $\mathbf{M}_{k,n}$
    update $\mathbf{W}_{k,n}$ and notify to all BSs in $\mathcal{B}$ using uplink precoded pilots as in [27]
    update $\tilde{\boldsymbol{u}}_{l,k,n}$ using (16c) and (17) for SCA or $\tilde{\epsilon}_{l,k,n}$ using (25c) for MSE approach
    $i = i + 1$
  until *convergence or* $i \geq I_{\max}$
end

on the MSE reformulation method discussed in Section III-C. In contrast to Section IV-A, the problem is solved using the KKT conditions in which the transmit precoders, receive beamformers and the subgradient updates are performed at the same instant to minimize the global queue deviation objective with few number of iterations. The proposed methods in this section provide algorithms that can be of practical importance in the TDD system due to the limited signaling requirements. Similar work has been considered for the WSRM problem with minimum rate constraints in [9], [10]. Since the formulation in [9], [10] are similar to the Q-WSRME scheme with an additional maximum rate constraint (14), it requires explicit dual variables to handle the maximum rate constraint, thereby making the problem difficult to solve in an iterative manner.

In the proposed JSFRA formulation, the maximum rate constraints are implicitly handled by the objective function without the need of explicit constraints. However, the KKT conditions cannot be formulated due to the non-differential objective function. The non-differentiability is due to the absolute value operator present in the norm function. 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 of 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, which means also ignoring the queues in the first place as well.

With the assumption of either one of the above conditions to be true, the problem in (27) can be written as

$$\underset{\substack{t_{l,k,n}, \mathbf{M}_{k,n}, \\ \epsilon_{l,k,n}, \mathbf{W}_{k,n}}}{\text{minimize}} \quad \sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{U}_b} a_k \left(Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n} \right)^q \tag{36a}$$

subject to

$$\alpha_{l,k,n} : \left| 1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n} \right|^2 + \widetilde{N}_0 + \sum_{(x,y) \neq (l,k)} \left| \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_y,k,n} \mathbf{m}_{x,y,n} \right|^2 \leq \epsilon_{l,k,n} \tag{36b}$$

$$\sigma_{l,k,n} : \log_2(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2) \tilde{\epsilon}_{l,k,n}} \leq -t_{l,k,n} \tag{36c}$$

$$\delta_b : \sum_{n=1}^{N} \sum_{k \in \mathcal{U}_b} \sum_{l=1}^{L} \operatorname{tr}\left(\mathbf{m}_{l,k,n} \mathbf{m}_{l,k,n}^{\mathrm{H}} \right) \leq P_{\max}, \forall b, \tag{36d}$$

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (36b), (36c) and (36d).

The problem in (36) is solved using the KKT expressions, which are obtained by the derivative of the Lagrangian function w.r.t the primal and the dual variables, complementary slackness conditions, and the primal, dual feasibility requirements as shown in Appendix B. Upon solving, we obtain the iterative solution as

$$\mathbf{m}_{l,k,n}^{(i)} = \Bigg(\sum_{x \in \mathcal{U}} \sum_{y=1}^{L} \alpha_{y,x,n}^{(i-1)} \mathbf{H}_{b_k,x,n}^{\mathrm{H}} \mathbf{w}_{y,x,n}^{(i-1)} \mathbf{w}_{y,x,n}^{\mathrm{H}\,(i-1)} \mathbf{H}_{b_k,x,n} + \delta_b \mathbf{I}_{N_T} \Bigg)^{-1} \alpha_{l,k,n}^{(i-1)} \mathbf{H}_{b_k,k,n}^{\mathrm{H}} \mathbf{w}_{l,k,n}^{(i-1)} \tag{37a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \Bigg(\sum_{x \in \mathcal{U}} \sum_{y=1}^{L} \mathbf{H}_{b_x,k,n} \mathbf{m}_{y,x,n}^{(i)} \mathbf{m}_{y,x,n}^{\mathrm{H}\,(i)} \mathbf{H}_{b_x,k,n}^{\mathrm{H}} + \mathbf{I}_{N_R} \Bigg)^{-1} \mathbf{H}_{b_k,k,n} \, \mathbf{m}_{l,k,n}^{(i)}. \tag{37b}$$

$$\epsilon_{l,k,n}^{(i)} = \left| 1 - \mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}^{(i)} \right|^2 + N_0 \, \|\mathbf{w}_{l,k,n}^{(i)}\|^2 + \sum_{(x,y) \neq (l,k)} \left| \mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)} \mathbf{H}_{b_y,k,n} \mathbf{m}_{x,y,n}^{(i)} \right|^2 \tag{37c}$$

$$t_{l,k,n}^{(i)} = -\log_2(\epsilon_{l,k,n}^{(i-1)}) - \frac{\left(\epsilon_{l,k,n}^{(i)} - \epsilon_{l,k,n}^{(i-1)} \right)}{\log(2) \, \epsilon_{l,k,n}^{(i-1)}} \tag{37d}$$

$$\sigma_{l,k,n}^{(i)} = \left[\tfrac{a_k \, q}{\log(2)} \left(Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}^{(i)} \right)^{(q-1)} \right]^{+} \tag{37e}$$

$$\alpha_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i-1)} + \rho \left(\frac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i)}} - \alpha_{l,k,n}^{(i-1)} \right) \tag{37f}$$

Since the dual variables $\alpha^{(i)}$ and $\sigma^{(i)}$ are interdependent in (37), one has to be fixed to optimize for the other. So, the variable $\alpha^{(i)}$ is fixed as in (37f) to obtain the other variables in (37). At each iteration, the dual variables $\alpha^{(i)}$ are updated linearly from the earlier $\alpha^{(i-1)}$ by a step size of $\rho \in [0, 1]$. When the allocated rate $t_k^{(i-1)}$ is greater than the number of queued packets Q_k for a user k, the corresponding dual variable $\sigma^{(i)}$ will be negative and due to the projection operator $[x]^+$ in (37e), it will be zero, thereby forcing $\alpha_k^{(i)} < \alpha_k^{(i-1)}$ as in (37f). Once the $\alpha_k^{(i)}$ is reduced, the precoder weight in

(37a) is lowered to make the rate $t_k^{(i)} < t_k^{(i-1)}$.

The KKT expressions in (37) are solved in an iterative manner by initializing the transmit and the receive beamformers $\mathbf{M}_{k,n}, \mathbf{W}_{k,n}$ 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 (37). The transmit precoder in (37a) depends on the BS specific dual variable δ_b, which can be found by bisection search satisfying the total power constraint (36d). 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 (37).

To devise an algorithm for a practical implementation, we extend the decentralization methods discussed in [27], for our problem of minimizing the total number of backlogged packets as follows. After receiving the updated transmit precoders from all BSs in $\mathcal{B}$, each user evaluates the MMSE receiver in (37b) and notify them to the BSs via uplink precoded pilots. On receiving pilot signals, BSs update the MSE in (23) as

$$\epsilon_{l,k,n}^{(i)} = 1 - \mathbf{w}_{l,k,n}^{(i)\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}^{(i)}. \tag{38}$$

Using the current MSE value, $t_{l,k,n}^{(i)}, \sigma_{l,k,n}^{(i)}$, and $\alpha_{l,k,n}^{(i)}$ are evaluated using (37d), (37e) and (37f), and the updated 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 back-haul exchanges between BSs, as an alternative approach, users may perform all processing required and BSs will update the precoders based on the feedback information from the users. Upon receiving the transmit precoders from BSs, each user will update the receive beamformer $\mathbf{w}_{l,k,n}$, the MSE $\epsilon_{l,k,n}$, and the dual variables $\lambda_{l,k,n}$ and $\alpha_{l,k,n}$. The updated $\alpha_{l,k,n}$ and $\mathbf{w}_{l,k,n}$ are notified to the BSs using two separate precoded uplink pilot symbols with $\tilde{\mathbf{w}}_{l,k,n}^{(i)} = \sqrt{\alpha_{l,k,n}^{(i)}} \mathbf{w}_{l,k,n}^{*(i)}$ and $\bar{\mathbf{w}}_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i)} \mathbf{w}_{l,k,n}^{*(i)}$ as the precoders. On receiving the precoded uplink pilots, each BS use the effective channel $\mathbf{H}_{b,k,n}^{\mathrm{T}} \tilde{\mathbf{w}}_{l,k,n}^{(i)}$ and $\mathbf{H}_{b,k,n}^{\mathrm{T}} \bar{\mathbf{w}}_{l,k,n}^{(i)}$ in (37a) to update the transmit precoders, where $\mathbf{x}^*$ is the complex conjugate of $\mathbf{x}$. Finally, Algorithm 3 outlines the distributed precoder design using the KKT based MSE reformulated JSFRA problem.

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**
  initialize $i = 0$
  repeat
    update $\mathbf{M}_{k,n}^{(i)}$ using (37a), and perform downlink transmission
    find $\mathbf{W}_{k,n}^{(i)}$ using (37b) 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 (37c) and (37d), (37e) and (37f) at each user with the updated $\mathbf{W}_{k,n}^{(i)}$
    using precoded uplink pilots, $\mathbf{W}_{k,n}^{(i)}$ 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

V. Simulation Results

The simulations carried out in this work consider the path loss varying uniformly across all users in the system with the channels drawn from the *i.i.d.* samples. The queues are generated based on the Poisson process with the average values specified in each section presented.

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 number of packets waiting at the transmitter for each user is given by $Q_k = 4, 8$ and 4 bits, respectively.

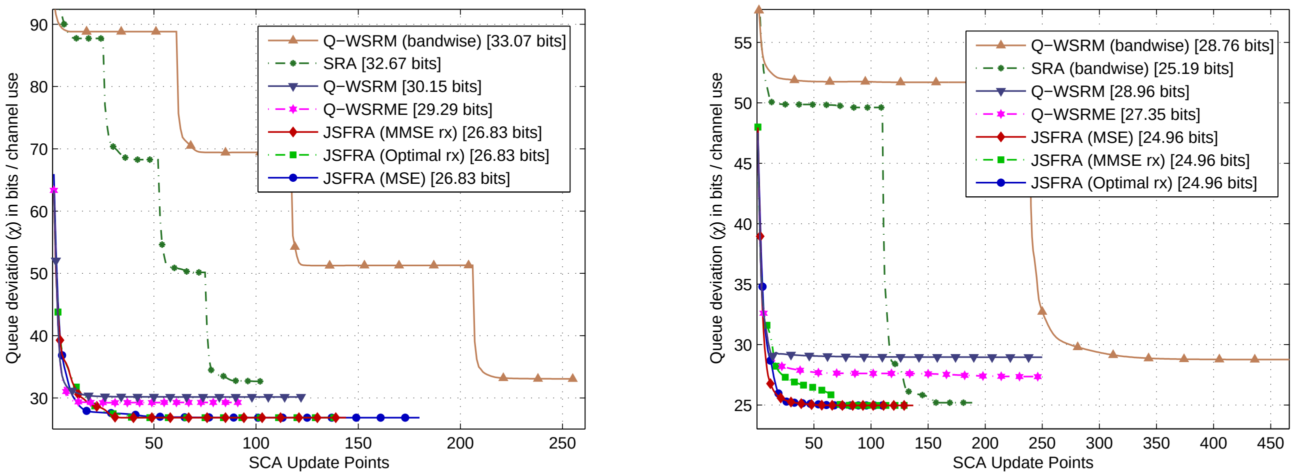
Table I tabulates the channel seen by the users over each sub-channel followed by the rates assigned by three different algorithms, Q-WSRME allocation, JSFRA approach and the band-wise Q-WSRM scheme using the WMMSE design [7]. The performance metric used for the comparison is the total number of backlogged bits left over at each slot after the allocation, which is denoted as $\chi = \sum_{k=1}^{K} \, [Q_k - t_k]^+$. Even though $\mathcal{U}(1)$ and $\mathcal{U}(3)$ has 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 waiting at the transmitter. The rate allocated for $\mathcal{U}(2)$ on the second sub-channel is higher in JSFRA scheme compared to the other schemes. It is due to the efficient allocation of the total power shared across the sub-channels.

For a MIMO framework, we consider a system with $N = 3$ sub-channels and $N_B = 3$ BSs, each equipped with $N_T = 4$ transmit antennas operating at 10dB SNR, serving $|\mathcal{U}_b| = 3$ users each. The path loss between the BSs and the users are uniformly generated from $[0, -3]$ dB and the association is made by selecting the BS with the lowest path loss component. Fig. 1(a) shows the performance of the centralized schemes for a single receive antenna system. The total number of queued packets for Fig. 1(a) is given by $Q_k = [14, 15, 14, 8, 12, 9, 12, 11, 11]$ bits and for Fig. 1(b) is $Q_k = [9, 12, 8, 12, 5, 4, 10, 8, 5]$ bits respectively.

The performance of the centralized algorithms are compared in terms of the total number of residual bits remaining in the system after each SCA update in Fig. 1. The Q-WSRM algorithm is not optimal due to the problem of over-allocation when the number of queued packets are few in number. In contrast, the Q-WSRME algorithm provides more favorable allocation by including the explicit rate constraint to avoid

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 $\{N, N_B, K, N_T, N_R\} = \{4, 3, 9, 4, 1\}$ (b). System Model $\{N, N_B, K, N_T, N_R\} = \{2, 3, 9, 4, 2\}$

Fig. 1. Total number of backlogged packets χ present in the system after each SCA updates

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	

the over-allocation. It can be seen that the JSFRA algorithms converge to the optimal point for all formulations proposed in Section III-B.

For both scenarios in Fig. 1, the Q-WSRME performs marginally inferior to the JSFRA algorithms due to the weights used in the algorithm. The performance loss is attributed to the fact that the Q-WSRME algorithm favors the users with the large number of backlogged packets as compared to the users with better channel conditions. Fig. 1(b) compares the algorithms for $N_R = 2$ receive antenna case. In all figures, the receivers are updated along with the SCA update instants *i.e*, $J_{\max} = 1$ in Algorithm 1. It is also noted that the performance degradation by performing the group update is very minimal. Since the receiver minimizes the objective for the fixed transmit precoders, the convergence is monotonic as can be seen from the figures.

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 performance of the distributed algorithms are compared using the total number of backlogged packets after each SCA update points. Fig. 2 compares the performance of the algorithms for the system configuration $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ with $N_T = 4$ transmit antennas at the BSs. Each BS serves $|\mathcal{U}_b| = 4$ users in a coordinated manner to reduce the total number of backlogged packets at each BS. The total number of queued packets assumed for both figures is $Q_k = [5, 7, 9, 11, 8, 12, 5, 4]$ bits. As pointed out in Section IV, the performance and the convergence speed of the distributed algorithms are susceptible to the step size used in the subgradient update. 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 plots the performance of the primal and the ADMM solutions for the JSFRA scheme using the SCA and by MSE relaxation at each SCA point. In between the SCA updates, the primal or the ADMM scheme is performed for $J_{\max} = 20$ iterations to exchange the respective coupling variables. In Fig. 2, the total number of backlogged packets at each SCA points

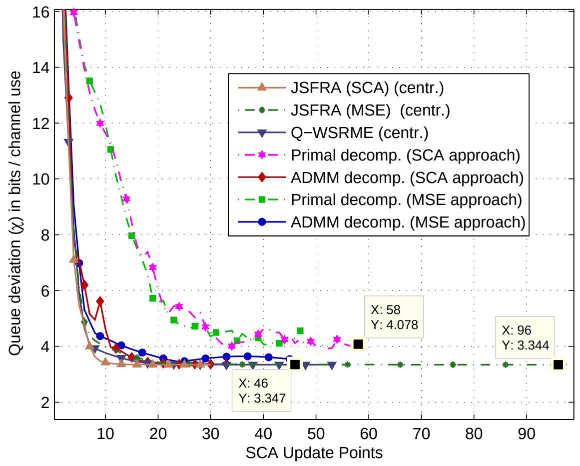


Fig. 2. Convergence behaviour of the centralized and the distributed algorithms for a system $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$

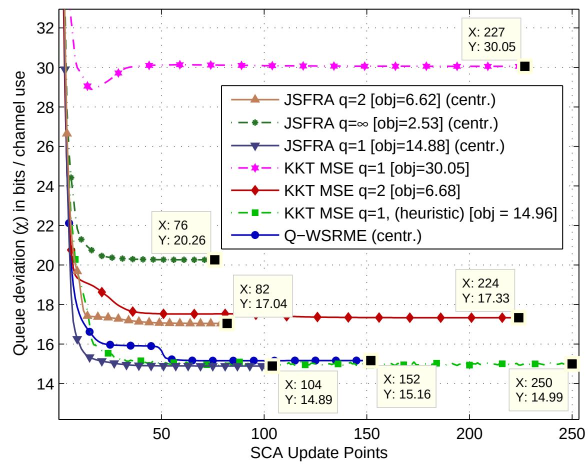


Fig. 3. Impact of varying q in the total number of backlogged packets after each SCA update for a system $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$

are plotted without the inner loop iterations of $J_{\max}$ times for the primal or the dual variables convergence. It can be seen from Fig. 2 that the distributed algorithms approach the centralized performance by exchanging minimal information between the coordinating BSs.

Fig. 3 compares the performance of the centralized and the KKT algorithm in Section IV-B for different exponents by plotting the total number of backlogged packets at each SCA update point. The ℓ_1 norm JSFRA scheme provides better performance over other schemes due to the greedy objective. The KKT approach for ℓ_1 norm is not defined due to the non-differentiability of the objective as discussed in the Section IV-B. If used for ℓ_1 norm, the problem of 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 (37e), which will always be equal to one. A heuristic method based on subdifferential calculus in [2] is proposed in Fig. 3 by assigning zero for $\sigma_{l,k,n}$ when the queue deviation is negative, *i.e*, $Q_k - t_k < 0$. It is required to address the problem of over-allocation in the ℓ_1 norm for dropping the absolute value operator from the objective function. It can be seen that the heuristic method oscillates near the optimal point with the deviation determined by the factor ρ used in (37f).

The objective values are mentioned in the legend for all the schemes and the objective of the ℓ_2 norm is not the same as that of the ℓ_1 norm used for plotting. For simulations, we update all variables in (37) at once at each iteration, *i.e*, $J_{\max} = 1$, which is well justified for the practical implementations due to the signaling overheads. The ℓ_2 norm for the JSFRA and the KKT approach achieves nearly the same value of 6.62 with different χ, due to the limited number of iterations for the dual variable convergence between each SCA update. Fig. 3 also shows the effect of dropping the squared rate variable from the objective in the Q-WSRME scheme compared to the ℓ_2 norm which includes it. By dropping it, the Q-WSRME scheme minimizes the number of queued packets in a prioritized manner based on the respective queues. On contrary, the ℓ_2 norm allocate rates to the users with the higher

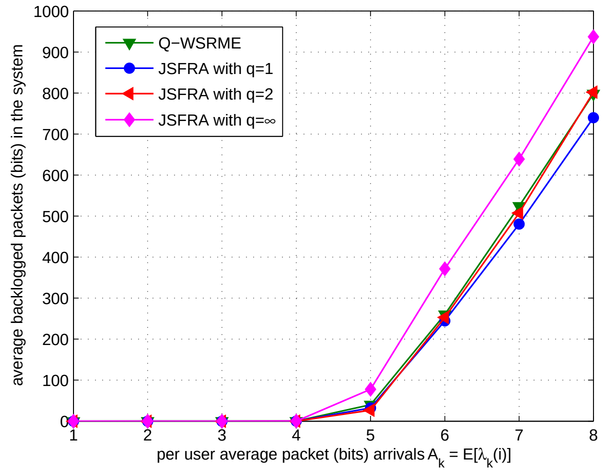


Fig. 4. System $\{N, N_B, K, N_R\} = \{4, 2, 12, 1\}$

number of queued packets before addressing the users with the smaller number of queued packets.

C. Average Backlogged Packets Over Slots

This section is dedicated to understand the performance of the proposed JSFRA scheme with different values of ℓ_q over multiple time slots. The performance is compared with the existing Q-WSRME scheme for different arrival rates A_k. Fig. V-C compares the performance of the centralized algorithms for different ℓ_q values using the average number of backlogged packets present in the system after each time instant. The horizontal axis indicate the average arrivals in bits per user, which is constant for all user. The instantaneous arrival follows Poisson distribution. The model considered for simulation is a 4×1 MIMO system with $N = 4$ sub-channels with $N_B = 2$ BSs. The path loss is modeled as a uniform random variable $[0, -3]$ dB with the maximum SINR seen by any user is 6 dB. The average is performed over 100 time slots.

It can be seen from Fig. V-C that the performance of the JSFRA scheme with ℓ_2 norm and the Q-WSRME approach are similar in the average residuals after each transmission instant. Note that the performance of the Q-WSRM scheme is similar to the Q-WSRME approach when the arrival rates are greater than the actual transmissions. It can be seen from Fig. V-C that the average number of backlogged packets are contained until the per user arrival is ≤ 4 bits. After that, the average queues become unbounded. The performance of the JSFRA scheme with ℓ_1 norm has significantly less number of average queued packets in comparison with other schemes. The performance of the JSFRA scheme with ℓ_∞ performs the worst in terms of the average number of queued packets after each transmission instant by achieving the fair service rate to all users.

VI. Conclusions

In this paper, we addressed the problem of allocating downlink space-frequency resources to the users in a multi-cell MIMO IBC system using OFDM transmission. 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 proposed the JSFRA scheme by employing the SCA technique to relax the nonconvex constraint by a sequence of convex subsets for designing the precoders to minimize the total number of user queued packets. Additionally, an alternative MSE relaxation approach is also proposed by using SCA technique to address the nonconvex constraints for a fixed MMSE receivers. We then introduced distributed 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 conditions of the MSE reformulated JSFRA method. 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 performance of the proposed algorithms with the existing solutions.

Appendix A
Tightness of SINR relaxation

In order for the constraints (16b) and (16c) to be active, there should be at least one user per BS with large enough backlogged packets that cannot be satisfied with the given power budget. To prove this, let us consider a single BS $N_T \times 1$ system with $|U|_b$ users each with Q_k backlogged packets. Let us consider the problem for a single BS b as

$$\begin{aligned} \underset{\gamma_k, \mathbf{m}_k, \beta_k}{\text{minimize}} \quad & \sum_{k \in \mathcal{U}_b} \left| Q_k - \log_2(1+\gamma_k) \right|^q && \text{(39a)} \\ \text{subject to} \quad & \gamma_k \leq \frac{|\mathbf{H}_k \mathbf{m}_k|^2}{\beta_k} && \text{(39b)} \\ & \beta_k \geq N_0 + \sum_{i \neq k} |\mathbf{H}_k \mathbf{m}_i|^2 && \text{(39c)} \\ & \sum_{k \in \mathcal{U}_b} \operatorname{tr}\left(\mathbf{m}_k \mathbf{m}_k^{\mathrm{H}}\right) \leq P_{\max}. && \text{(39d)} \end{aligned}$$

Let us assume the system has $N_T > |\mathcal{U}_b|$ and the precoders are determined for all users at the optimal point for which the objective is minimum and is nonzero. Let us consider two users, say i and j with precoders $\mathbf{m}_i$ and $\mathbf{m}_j$. In order to show the tightness of the constraints (39b) and (39c), let us assume by contradiction that (39b) and (39c) are not tight for a user i only. Let us also consider that γ_k for all users $k \in \mathcal{U}_b \backslash j$ are identified to satisfy the backlogged packets.

Now, in order to minimize the objective of the i^{th} user, γ_i should be $2^{Q_i} - 1$ to have the minimum objective. Since the constraints (39b) and (39c) are not tight for the i^{th} user, the actual SINR on the r.h.s of (39b) is greater than γ_i. It is attributed to the excess power in the transmit precoder of user i as $\mathbf{m}_i = \bar{\mathbf{m}}_i + \mathbf{V}_i \mathbf{e}$, where $\mathbf{V}_i = \mathcal{N}([\mathbf{H}_0^{\mathrm{T}}, \ldots, \mathbf{H}_{i-1}^{\mathrm{T}}, \mathbf{H}_{i+1}^{\mathrm{T}}, \ldots, \mathbf{H}_{|\mathcal{U}_b|-1}^{\mathrm{T}}]^{\mathrm{T}})$ denotes the null space corresponding to the other user channel vectors and $\mathbf{e}$ be any random vector with power ΔP. Note that the extra power has no impact on the interference constraint (39c) of other users.

Since the allocated rate of user j is less than the queued packets, we can find a precoder $\mathbf{m}_j$ with the additional power of ΔP as $\mathbf{m}'_j = \bar{\mathbf{m}}_j + \mathbf{V}_j \mathbf{e}'$. Note that the new precoder has no impact on the interference constraint of other users. The newly found precoder $\mathbf{m}'_j$ minimize the queue of user j without affecting the rates of other user, there by reduces the objective further. It is in contradiction to our original assumption that the precoders are optimal and the objective is minimum.

It can be seen that the constraints are tight as long as there is one user with unserviced backlogged packets at each BS with the given power budget. Since the objective is not to minimize the power, when the power budget is more than sufficient to service the backlogged packets of all users, the JSFRA scheme is not guaranteed to find the minimum power precoders to minimize the current backlogged packets in the system. The objective of the JSFRA problem can be modified to include the power term to find the minimum power precoders as

$$\|\tilde{\mathbf{v}}\|_q + \varphi \sum_{b \in \mathcal{B}} \sum_{k \in \mathcal{U}_b} \sum_{n=1}^{N} \sum_{l=1}^{L} \|\mathbf{m}_{l,k,n}\|_2,$$

for a small φ which doesn't affect the optimal solution. Note that the above objective is guaranteed to make the constraints (16b) and (16c) tight by relaxing the power constraint (16d).

A. Convergence Analysis

1) Centralized Problem: Let us express the JSFRA problem in (16) and (25) as

$$\begin{aligned} \underset{\mathbf{x}, \mathbf{y}, \mathbf{z}}{\text{minimize}} \quad & f(\mathbf{x}, \mathbf{y}, \mathbf{z}) && \text{(40a)} \\ \text{subject to} \quad & h(\mathbf{z}) - g_0(\mathbf{x}, \mathbf{y}) \leq 0 && \text{(40b)} \\ & g_1(\mathbf{x}, \mathbf{y}) \leq 0, && \text{(40c)} \\ & g_2(\mathbf{x}) \leq 0, && \text{(40d)} \end{aligned}$$

wher g_2, f are convex functions and h is a linear function. Let g_0, g_1 are convex functions only on $\mathbf{x}$ or $\mathbf{y}$ as the variable but not on both. Note that the (40b) correspond to the constraints in (16b) and (25b) and (40c) corresponds to the constraints in (16c) and (25c). Other convex constraints are addressed by the constraint (40d). With this, the feasible set of the problem (40) is given by

$$\mathcal{F}= \{\mathbf{x},\mathbf{y},\mathbf{z}|h(\mathbf{z})-g_0(\mathbf{x},\mathbf{y})\leq 0, g_1(\mathbf{x},\mathbf{y})\leq 0, g_2(\mathbf{x})\leq 0\}$$

In order to solve the problem, we resort to alternating optimization (AO) by fixing a block of varibles and optimize for others. In the problem (40), even after fixing the variable $\mathbf{y}$, the problem is nonconvex due to the DC constraint in (40b). In order to solve the problem after fixing the variable $\mathbf{y}$, we adopt the SCA approach presented in [28]–[30] by relaxing the nonconvex set by a sequence of convex subsets. Since the proposed method involves two level of iterations, we denote the AO iteration index by a superscript (i) and the DC constraint relaxations by a subscript k. Let $\mathcal{X}_k^{(i)}$ be a feasible set for the i^{th} AO iteration and the k^{th} SCA point for a fixed $\mathbf{y}$ and for a fixed $\mathbf{x}$, the set is represented as $\mathcal{Y}_k^{(i)}$. Since the SCA iterations are performed until convergence, let $\mathbf{x}_*^{(i)}$ denotes the converged point of $\mathbf{x}$ in the i^{th} AO iteration. For the sake of clarity, we define the optimal value of $\mathbf{z}$ obtained for the i^{th} AO iterate for the fixed $\mathbf{y}$ variable as $\mathbf{z}_{*|y}^{(i)}$.

Let us consider the order for AO by optimizing variable $\mathbf{x}$ before optimizing the variable $\mathbf{y}$. Without affecting the convergence proof, let us consider the variable $\mathbf{y}$ is fixed for the AO i with the optimal value achieved from the previous iteration $i-1$ as $\mathbf{y}_*^{(i-1)}$. In order to find the optimal value of $\mathbf{x}$ for the SCA iteration k, we linearize the nonconvex function g_0 using previous SCA iterate of $\mathbf{x}$ as

$$\begin{aligned}\hat{g}_o(\mathbf{x},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)}) &= g_0(\mathbf{x}_k^{(i)},\mathbf{y}_*^{(i-1)})\\ &\quad+\nabla g_0(\mathbf{x}_k^{(i)},\mathbf{y}_*^{(i-1)})^{\mathrm{T}}(\mathbf{x}-\mathbf{x}_k^{(i)}).\end{aligned}\tag{41}$$

Using (41), the convex subproblem for i^{th} AO iteration and k^{th} SCA point for the variable $\mathbf{x}$ and $\mathbf{z}$ is given by

$$\underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad f(\mathbf{x},\mathbf{y}_*^{(i-1)},\mathbf{z}) \tag{42a}$$
$$\text{subject to} \quad h(\mathbf{z})-\hat{g}_0(\mathbf{x},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)})\leq 0 \tag{42b}$$
$$g_1(\mathbf{x},\mathbf{y}_*^{(i-1)})\leq 0, \tag{42c}$$
$$g_2(\mathbf{x})\leq 0, \tag{42d}$$

Let the set defined by the problem in (42) be represented as $\mathcal{X}_k^{(i)}\subset\mathcal{F}$. In order to prove the convergence of the convex subproblem (42) for a fixed $\mathbf{y}=\mathbf{y}_*^{(i-1)}$ operating at $\mathbf{x}_k^{(i)}$, let us assume that (42) yields $\mathbf{x}_{k+1}^{(i)}$ and $\mathbf{z}_{k+1}^{(i)}$ as the solution at the k^{th} iteration. To show $\mathbf{x}_{k+1}^{(i)}$ and $\mathbf{z}_{k+1}^{(i)}$ minimizes the objective function ans also feasible, let us assume that the point $\mathbf{x}_k^{(i)}\in\mathcal{X}_k^{(i)}$, which is feasible for (42). Since the function $g_0(\mathbf{x},\mathbf{y}_*^{(i-1)}$ is linearized at $\mathbf{x}_k^{(i)}$, it satisfies

$$h(\mathbf{z})-g_0(\mathbf{x},\mathbf{y}_*^{(i-1)})\leq h(\mathbf{z})-\hat{g}_0(\mathbf{x},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)})\leq 0, \tag{43}$$

since $g_0(\mathbf{x},\mathbf{y}_*^{(i-1)})\leq\hat{g}_0(\mathbf{x},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)}),\forall\mathbf{x}\in\mathcal{X}_k^{(i)}$. Now, $\mathbf{x}_{k+1}^{(i)}$ and $\mathbf{z}_{k+1}^{(i)}$ are the optimal and feasible point for the k^{th} subproblem (42), it satisfies

$$\begin{aligned}h(\mathbf{z}_{k+1}^{(i)})-g_0(\mathbf{x}_{k+1}^{(i)},\mathbf{y}_*^{(i-1)}) &\leq h(\mathbf{z}_{k+1}^{(i)})-\hat{g}_0(\mathbf{x}_{k+1}^{(i)},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)})\\ &\leq h(\mathbf{z}_k^{(i)})-\hat{g}_0(\mathbf{x}_k^{(i)},\mathbf{y}_*^{(i-1)};\mathbf{x}_k^{(i)})\leq 0.\end{aligned}\tag{44}$$

Using (44), we can prove that the solution $\mathbf{x}_{k+1}^{(i)}$ and $\mathbf{z}_{k+1}^{(i)}$ are feasible, since the initial point of $\mathbf{x}=\mathbf{x}_*^{(i-1)}$ was chosen to be feasible from the earlier AO iteration $i-1$. In order to prove the convergence of the objective, using (44), we can see that $\mathcal{X}_0^{(i)}\subseteq\cdots\subseteq\mathcal{X}_{k-1}^{(i)}\subseteq\mathcal{X}_k^{(i)}\subset\mathcal{F}$. Since the feasible set of the problem (42) includes the feasible sets from the earlier iteration, we arrive at

$$\begin{aligned}f(\mathbf{x}_0^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_0^{(i)}) &\geq f(\mathbf{x}_k^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_k^{(i)})\\ &\geq f(\mathbf{x}_{k+1}^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_{k+1}^{(i)})\geq f(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_{*|y}^{(i)}).\end{aligned}\tag{45}$$

Thus the sequence $f(\mathbf{x}_k^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_k^{(i)})$ is nonincreasing and converges to a critical point. Note that feasible point $(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_{*|y}^{(i)})$ need not be a stationary point of the problem (40), since it is the minimizer only in the feasible set $\mathcal{X}_*^{(i)}\subset\mathcal{F}$, which depends on $\mathbf{x}$ and $\mathbf{z}$ only.

Once the solution is found for a fixed $\mathbf{y}$, we fix $\mathbf{x}$ as $\mathbf{x}_*^{(i)}$ and optimize for $\mathbf{y}$. Even after treating $\mathbf{x}$ as a constant, the problem is still nonconvex due to the DC constraint. Following similar approach, we can find the minimizer $\mathbf{y}_k^{(i)}$ and $\mathbf{z}_k^{(i)}$ for the convex subproblem (42) at each iteration k. Note that $\mathbf{z}_k^{(i)}$ is reused since the variable $\mathbf{x}$ is fixed for the i^{th} AO iteration. The convergence and the nonincreasing behavior of the problem follows similar arguments as above. Now, the optimal solution of the converged subproblems with $\mathbf{y}$ as variable are $\mathbf{y}_*^{(i)}$ and $\mathbf{z}_*^{(i)}$. Note that the solution point $(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i)},\mathbf{z}_{*|x}^{(i)})$ is the unique minimizer in the set $\mathcal{Y}_*^{(i)}$.

Finally, to prove the global convergence of the objective, we need to show the nonincreasing behavior of the objective function between each AO update, *i.e*,

$$f(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i)},\mathbf{z}_{*|x}^{(i)})\leq f(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_{*|y}^{(i)}).$$

Let us consider an AO iteration i in which the optimal value for $\mathbf{x}$ and $\mathbf{z}$ are obtained as $\mathbf{x}_*^{(i)}$ and $\mathbf{x}_{*|y}^{(i)}$ using fixed $\mathbf{y}=\mathbf{y}_*^{(i-1)}$. In order to find $\mathbf{y}_*^{(i)}$, we fix $\mathbf{x}$ as $\mathbf{x}_*^{(i)}$ and optimize for $\mathbf{y}$. Since the problem (40) is nonconvex even after fixing $\mathbf{x}$, we have to linearize the convex function in the DC constraint (40b) around some fixed operating point of $\mathbf{y}$. Since we know that $\mathbf{y}_*^{(i-1)}$ is already a feasible point for $\mathbf{y}$ along with $\mathbf{x}_*^{(i)}$, linearization is performed around this feasible point. Using the inequality in (44), we can show that $\{\mathbf{y}_*^{(i-1)},\mathbf{x}_*^{(i)},\mathbf{z}_{*|y}^{(i)}\}\in\mathcal{Y}_0^{(i)}$. Now the optimization is performed to find the optimal $\mathbf{y}$ using the relaxed subproblem (42), the optimal solution $\mathbf{z}_0^{(i)}$ for the initial iteration after AO update follows

$$f(\mathbf{x}_*^{(i)},\mathbf{y}_0^{(i)},\mathbf{z}_0^{(i)})\leq f(\mathbf{x}_*^{(i)},\mathbf{y}_*^{(i-1)},\mathbf{z}_{*|x}^{(i)}),$$

since the feasible set includes the earlier optimal points $\{\mathbf{y}_*^{(i-1)},\mathbf{x}_*^{(i)}\}$ as the operating point for AO and SCA iteration. Note that it is not possible to say $\mathcal{X}_*^{(i)}\subseteq\mathcal{Y}_0^{(i)}$, since the set is nonconvex on $\mathbf{x},\mathbf{y}$, but $\{\mathbf{y}_*^{(i-1)},\mathbf{x}_*^{(i)},\mathbf{z}_{*|y}^{(i)}\}\in\{\mathcal{X}_*^{(i)}\cap\mathcal{Y}_0^{(i)}\}$. Now by using induction, we can show that the global problem converges to a feasible point of the nonconvex problem (40) in a nonincreasing manner.

To show the converged point of the iterative algorithm is in fact the stationary point of the nonconvex problem (40), it must satisfy the KKT conditions of the nonconvex problem. Since the converged point is the minimizer for the iterative algorithm such that

$$f(\mathbf{x}_*^{(i)}, \mathbf{y}_*^{(i)}, \mathbf{z}_{*|x}^{(i)}) = f(\mathbf{x}_*^{(i+1)}, \mathbf{y}_*^{(i)}, \mathbf{z}_{*|y}^{(i+1)}) \\ = f(\mathbf{x}_*^{(i+1)}, \mathbf{y}_*^{(i+1)}, \mathbf{z}_{*|x}^{(i+1)}), \quad (46)$$

the solution is inside the feasible set $\mathcal{F}$ and $(\mathbf{x}_*^{(i+1)}, \mathbf{y}_*^{(i+1)}, \mathbf{z}_{*|x}^{(i+1)})$ is the minimizer of the objective function f_0 in the feasible set $\mathcal{X}_*^{(i+1)} \subset \mathcal{F}$. Using the discussions in [23], we can easily show that the feasible point $\mathcal{F}$ and $(\mathbf{x}_*^{(i+1)}, \mathbf{y}_*^{(i+1)}, \mathbf{z}_{*|x}^{(i+1)})$, which is the minimizer in the local neighborhood, is a stationary point of the non convex problem in (40) satisfying the constraint qualifications and the KKT expressions for the set $\mathcal{X}_*^{(i+1)} \subset \mathcal{F}$. The non differentiability of the objective function in (16) and (25) requires the subdifferential set of the objective function to include $0 \in \partial f_0(\mathbf{z}_*)$ to satisfy the KKT conditions.

The uniqueness of the convex subproblems (42) is required for the convergence of AO. For the problem (19), the uniqueness of the transmit precoders are guaranteed by the constraint (18), which can be written as

$$\gamma_{l,k,n} + \tilde{\beta}_{l,k,n}^{-2} |\tilde{\mathbf{H}}_{b_k,k,n} \tilde{\mathbf{m}}_{l,k,n}|^2 (\beta_{l,k,n} - \tilde{\beta}_{l,k,n}) \\ - \tilde{\beta}_{l,k,n}^{-1} \tilde{\mathbf{m}}_{l,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n} (\mathbf{m}_{l,k,n} - \tilde{\mathbf{m}}_{l,k,n}) \leq 0, \quad (47)$$

where $\tilde{\mathbf{H}}_{b_k,k,n} = \mathbf{w}_{l,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n}$ and for the (27) by

$$|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}|^2 \\ + \sum_{(j,i) \neq (l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2 + \widetilde{N}_0 \leq \epsilon_{l,k,n}. \quad (48)$$

Using these arguments, we can claim that the proposed JSFRA centralized solution achieves a stationary point of the original nonconvex problem with the nonincreasing objective value at each iteration.

Appendix B
KKT conditions for MSE approach

In order to solve for an iterative precoder design algorithm, the KKT expressions for the problem in (36) are obtained by differentiating the Lagrangian by assuming the equality constraint for (36b) and (36c). At the stationary points, following conditions are satisfied.

$$\nabla_{\epsilon_{l,k,n}} : -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \quad (49a)$$

$$\nabla_{t_{l,k,n}} : -q\, a_k \Big(Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n} \Big)^{(q-1)} + \frac{\sigma_{l,k,n}}{\log_2(e)} = 0 \quad (49b)$$

$$\nabla_{\mathbf{m}_{l,k,n}} : \sum_{y \in \mathcal{U}} \sum_{x=1}^{L} \alpha_{x,y,n} \mathbf{H}_{b_k,y,n}^{\mathrm{H}} \mathbf{w}_{x,y,n} \mathbf{w}_{x,y,n}^{\mathrm{H}} \mathbf{H}_{b_k,y,n} \mathbf{m}_{l,k,n} \\ + \delta_b \mathbf{m}_{l,k,n} = \alpha_{l,k,n} \mathbf{H}_{b_k,k,n}^{\mathrm{H}} \mathbf{w}_{l,k,n}, \quad (49c)$$

$$\nabla_{\mathbf{w}_{l,k,n}} : \sum_{(x,y) \neq (l,k)} \mathbf{H}_{b_y,k,n} \mathbf{m}_{x,y,n} \mathbf{m}_{x,y,n}^{\mathrm{H}} \mathbf{H}_{b_y,k,n}^{\mathrm{H}} \mathbf{w}_{l,k,n} \\ + \mathbf{I}_{N_R} \mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n}. \quad (49d)$$

In addition to the primal constraints given in (36b), (36c) and (36d), the complementary slackness criterion must also be satisfied at the stationary point. Upon solving the above expressions in (49) with the complementary slackness conditions, we obtain the iterative algorithm to determine the transmit and the receive beamformers as shown in (37).

References

[1] E. Matskani, N. Sidiropoulos, Z.-Q. Luo, and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2682–2693, July 2008.

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

[3] C. Ng and H. Huang, "Linear Precoding in Cooperative MIMO Cellular Networks with Limited Coordination Clusters," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1446–1454, December 2010.

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

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

[6] S. S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 4792–4799, 2008.

[7] 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 Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, sept. 2011.

[8] 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," *arXiv preprint arXiv:1210.1507*, 2012.

[9] J. Kaleva, A. Tölli, and M. Juntti, "Primal decomposition based decentralized weighted sum rate maximization with QoS constraints for interfering broadcast channel," in *IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2013, pp. 16–20.

[10] ——, "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.

[11] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[12] 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.

[13] H. Pennanen, A. Tölli, and M. Latva-Aho, "Decentralized coordinated downlink beamforming via primal decomposition," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 647–650, 2011.

[14] A. Tölli, H. Pennanen, and P. Komulainen, "Decentralized minimum power multi-cell beamforming with limited backhaul signaling," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 570–580, 2011.

[15] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, Dec 1992.

[16] 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.

[17] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource allocation and cross-layer control in wireless networks*. Now Publishers Inc, 2006.

[18] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, 2004.

[19] 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.

[20] K. Seong, R. Narasimhan, and J. Cioffi, “Queue proportional scheduling via geometric programming in fading broadcast channels,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1593–1602, 2006.

[21] P. C. Weeraddana, M. Codreanu, M. Latva-aho, and A. Ephremides, “Resource allocation for cross-layer utility maximization in wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 60, no. 6, pp. 2790–2809, 2011.

[22] F. Zhang and V. Lau, “Cross-Layer MIMO Transceiver Optimization for Multimedia Streaming in Interference Networks,” *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1235–1244, March 2014.

[23] 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.

[24] Z.-Q. Luo and S. Zhang, “Dynamic Spectrum Management: Complexity and Duality,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, Feb 2008.

[25] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

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

[27] P. Komulainen, A. Tölli, and M. Juntti, “Effective CSI Signaling and Decentralized Beam Coordination in TDD Multi-Cell MIMO Systems,” *IEEE Transactions on Signal Processing*, vol. 61, no. 9, pp. 2204–2218, 2013.

[28] T. Lipp and S. Boyd, “Variations and extensions of the convex-concave procedure,” 2014.

[29] 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.

[30] 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