***Abstract*—In this paper, we consider the linear precoder design for the coordinated multi-cell multi-user (MU) multiple-input multiple-output (MIMO) technique with the orthogonal frequency division multiplexing (OFDM) transmission for the downlink broadcast channel. Each base station (BS) serves the users associated with it, which causes interference to the neighboring BS users. The design objective is to minimize the number of queued packets with all the coordinating BSs in the network, since the transmissions are guided by the backlogged packets. Traditionally, the problem is solved by weighing the transmission rates of the users by the corresponding length of the queued packets in the weighted sum rate maximization (WSRM) method, *i.e*, longer the queue, higher the priority. In this work, we approach the problem by formulating it as a nonconvex optimization problem and by applying successive convex approximation (SCA) method, we find the transmit and the receive beamformers in a coordinated manner. In the first part of the paper, we discuss the centralized solutions for the proposed problem in which the precoders are designed across the space-frequency resources jointly to minimize the number of backlogged packets, which is referred as the joint space-frequency resource allocation (JSFRA) scheme. In the second part, we discuss the precoder design across the coordinating BSs in a distributed manner using primal and the alternating directions method of multipliers (ADMM) based approaches. In addition to that, we also propose an iterative algorithm by solving the Karush-Kuhn-Tucker (KKT) conditions for the JSFRA scheme based on the mean squared error (MSE) relaxation, which can be solved in a distributed way across the coordinating BSs. The proposed solutions are compared to the traditional queue weighted sum rate maximization (Q-WSRM) approaches mentioned above in terms of the rate of convergence and the backlogged packets remaining after a scheduling instant.**

I. Introduction

The last mile wireless connectivity poses significant bottleneck in the overall data traffic for the interconnected networks. The main challenges in the wireless networks are due to the scarcity of the available resources either in terms of power or spectrum usage and the complexity of the receiver algorithms which are proportional to the mobile battery drain. In order to overcome the receiver complexity, orthogonal frequency division multiplexing (OFDM) based transmissions are introduced for the wide band transmissions. To improve data rate, multiple antennas are installed at base stations (BSs) and/or at user terminals to avail additional freedom in the form of spatial dimension. The inclusion of multiple-input multiple-output (MIMO) technique in wireless networks provides higher data rate or lower outage for the same transmission power and bandwidth.

In a network with multiple BSs serving multiple users (MU), 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. In this work, we consider the problem of transmit design over the space-frequency resources provided by the MIMO-OFDM framework in the downlink broadcast transmission to minimize the number of queued packets. Since the space-frequency resources are shared by multiple users associated with different BSs, the problem of interest can be viewed as a resource allocation to minimize the total number of backlogged packets of all BSs.

In general, the resource allocation problems are solved by assigning a binary variable to each user indicating the presence or the absence on a particular resource. In contrast to that, we use the transmit beamformers, which are the complex vectors, as a decision variable in determining the presence or the absence of a user on a particular resource. The purpose of using the transmit beamformers for the scheduling is two fold. Firstly, it determines the transmission rate on a certain resource and secondly, by making the transmit beamformer to be a zero vector, the corresponding user will not be scheduled on a certain resource.

In order to reduce the complexity involved in the precoder design problem, linear precoding are assumed at the BSs and detection is used at users. The queue minimizing network optimization objective is used to design the beamformers across the coordinating BSs, since the transmissions are guided by the available backlogged packets. To achieve the best performance, we propose a joint resource allocation scheme over the space and frequency dimensions among the coordinating BSs to minimize the time that the packets stay in the queues prior to the transmission, and, hence, to avoid packet drops as an indirect objective.

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 (see Section III for further discussions). The topics on MIMO broadcast channel (BC) precoder design have been studied extensively with different performance criteria in the literature. Due to the nonconvex nature of the MIMO BC precoder design problems, the successive convex approximation (SCA) method has gradually become a powerful tool to deal with these problems. For example, in [1], the nonconvex part of the objective is linearized around an operating point in order to solve the WSRM problem in an iterative manner. Similar approach of solving the WSRM problem via arithmetic-geometric inequality is proposed in [2].

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 [3], [4] is also used to solve the WSRM problem. In [5], [6], the WSRM problem is reformulated via MSE, casting the problem as a convex one for fixed MMSE receivers. 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 set of variables while the remaining others are fixed. The MSE reformulation for the WSRM problem is also studied in [7] by using the SCA to solve the problem in an iterative manner. Additional rate constraints based on the quality of service (QoS) requirements are included in the WSRM problem and solved via MSE reformulation in [8].

The problem of precoder design for the MIMO BC system are solved either by using a centralized controller or by using decentralized algorithms where all BSs often handles its own problem and exchange limited information via backhaul. The distributed approaches are based on primal, dual or alternating directions method of multipliers (ADMM) decomposition, which are discussed in a detailed manner in [9], [10]. In the primal decomposition, the so-called coupling interference variables are fixed for the subproblem at each BS to find the optimal precoders. The fixed interference are then updated by using the subgradient method as discussed in [11]. The dual and ADMM approach controls the distributed subproblems by fixing the *'interference price'* for each BS as detailed in [12].

By adjusting the weights properly, we can find arbitrary rate-tuple in the rate region of the system that maximizes other performance measures by solving the resulting WSRM problem. For example, if the weight of each user is set to be inversely proportional to his/her average data rate, the corresponding problem guarantees fairness on an average among the users. As an approximate method, we may assign weights based on the current queue size of 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 a function of it, which is the outcome of minimizing the Lyapunov drift between the current and the future queue states [13], [14]. In the backpressure algorithm, the differential queues between the source and the destination nodes are used as the weights scaling the transmission rate [15].

Earlier studies on the queue minimization problem was summarized in the survey paper [16]. In particular, the problem of power allocation to minimize the number of backlogged packets was considered in [17] by geometric programming formulations. Since the problem considered in [17] assumed single antenna transmitters and receivers, the queue minimizing problem reduces to the one of optimal power allocation. In the context of wireless networks, the backpressure algorithm mentioned above was extended in [18] by formulating the corresponding user queues as the weights in the WSRM problem.

In this paper, we consider the problem of precoder design across the space-frequency resources to minimize the total number of queued packets of all BSs. For this highly nonconvex problem, we first propose two centralized methods. In the first method, we relax the nonconvex constraint by the first order Taylor approximation around an operating point, which is updated in an iterative manner until convergence or to a certain accuracy. In the second method, we reformulate the joint space-frequency resource allocation (JSFRA) problem using the MSE equivalence with the rate expression to solve for the optimal precoders. For distributed implementation, we further proposed decentralized approaches based on primal and ADMM scheme to identify the precoders independently across the BSs by exchanging limited information via back-haul. We also proposed an iterative algorithm by solving the Karush-Kuhn-Tucker (KKT) equations, which can be implemented easily in a distributed manner.

The remainder of this paper is as follows. In Section II, we introduce the system model and the problem formulation for the queue minimizing precoder design. Existing and the proposed precoder designs for the JSFRA problem 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 BC scenario in an OFDM framework with N sub-channels and N_B BSs each equipped with N_T transmit antennas, serving 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 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.

In this paper we adopt linear 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 the 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 receiver, a receive beamforming vector $\mathbf{w}_{l,k,n}$ is employed at each user. Consequently, the received data symbol corresponding to the l^{th} spatial stream over sub-channel n at user k is given by

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

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), $L = rank(\mathbf{H}_{b,k,n}) = min(N_T, N_R)$ 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

$$\gamma_{l,k,n} = \frac{\left| \mathbf{w}_{l,k,n}^{\mathrm{H}}\, \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n} \right|^2}{N_0 \|\mathbf{w}_{l,k,n}\|^2 + \sum_{(j,i) \neq (l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2} \tag{2}$$

Let Q_k be the number of backlogged packets which are 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 packet arrivals of $A_k = \mathbf{E}_i\{\lambda_k\}$ packets/bits, where $\lambda_k(i) \sim \mathrm{Pois}(A_k)$ represents the instantaneous number of packets or bits arriving for the user k at the i^{th} instant. The total number of queued packets at the $(i+1)^{\text{th}}$ instant for the user k, represented by $Q_k(i+1)$, 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 transmission in bits for user k. For a MIMO-OFDM system,

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

where $t_{l,k,n}$ denotes the transmitted 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 $t_{l,k,n} \leq \log_2(1 + \gamma_{l,k,n})$ for the signal-to-interference-plus-noise ratio (SINR) of

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

$\gamma_{l,k,n}$.[2] 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 weighted ℓ_q-norm of all the queue deviation given by

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

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

$$\underset{\mathbf{M}_{k,n}, \mathbf{W}_{k,n}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \quad (6a)$$

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

where $\tilde{v}_k \triangleq a_k^{1/q} v_k$, and a_k is the weighting factor which is incorporated to control user priority based on their respective QoS, $\mathbf{M}_{k,n} \triangleq [\, \mathbf{m}_{1,k,n}\, \mathbf{m}_{2,k,n} \ldots \mathbf{m}_{L,k,n} \,]$ comprises the beamformers associated with the user k for n^{th} sub-channel transmission, and $\mathbf{W}_{k,n} \triangleq [\, \mathbf{w}_{1,k,n}\, \mathbf{w}_{2,k,n} \ldots \mathbf{w}_{L,k,n} \,]$ stacks the receive beamformers respectively[3]. In (6b), we consider a BS specific sum power constraint for each BS across all sub-channels.

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

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

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

Before proceeding further, we note that the constraint in (7) is handled implicitly by the definition of the ℓ_q in the objective of (6). As a proof, suppose that $t_k > Q_k$ for a certain k at 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 an optimal solution has been obtained. The choice of the norm ℓ_q used in the objective function [16], [17] alters the priorities for the queue deviation function as

- With $\ell_{q=1}$, the objective results in greedy allocation *i.e*, emptying the queue of users with good channel condition before considering the users with worse channel conditions. As a special case, it is easy to see that (6) reduces to the WSRM problem (14) when the queue size is large enough for all users.
- With $\ell_{q=2}$, the objective 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 backlogged packets is proportional to the delay in the transmission following the Little's law [14].
- With $\ell_{q=\infty}$, the objective 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 combinatorial and the nonconvex nature of the problem. In addition to that, 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, the solutions provided in the paper performs both precoder design and the scheduling of users in a joint manner. Before discussing the proposed solutions, we consider the existing algorithm to solve the issue of minimizing the number of backlogged packets with additional constraints required by 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.[4] One such algorithm is the *backpressure algorithm*, discussed in detail in [13]–[15]. The algorithm 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 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 MIMO-OFDM framework, in which the multiple BSs acts as the source nodes and the user terminals as the receiver nodes. The control policy in the form of transmit precoders are designed to minimize the number of queued packets waiting at the BSs. In order to find the optimal algorithm, we use the Lyapunov function which is predominantly used in the control theory for the system stability. Since at each time slot, the system can be described by the channel conditions and the number of backlogged packets of each user, Lyapunov function is used to provide a scalar measure, which grows large when the system moves towards the undesirable state. Following the approach

[2] This can be achieved by Gaussian signaling

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

[4] routers or user terminals

in [14], the scalar measure for the queue stability is given by

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

where $\mathbf{Q}(i)$ denotes the stacked user queues at the i^{th} slot and $\frac{1}{2}$ is used for the convenience. The Lyapunov function provides a measure of congestion in the system, as discussed in [14, Ch. 3]. Now the Lyapunov function drift is given by

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

$$\leq \sum_{k\in\mathcal{U}} \frac{\lambda_k^2(i)+t_k^2(i)}{2} + \sum_{k\in\mathcal{U}} Q_k(i)\left\{\lambda_k(i) - t_k(i)\right\} \tag{9b}$$

where the inequality is due to the upper bound

$$\left[\max(Q-t,0)+\lambda)\right]^2 \leq Q^2 + t^2 + \lambda^2 + 2Q(\lambda - t) \tag{10}$$

In order to minimize the squared sum at each instant, minimization of the Lyapunov drift (9) is carried over all possible rate allocations in the form of transmission rates t_k to users in the system. The Lyapunov drift conditioned on the current backlogged packets $\boldsymbol{Q}(i)$ is given by

$$\underset{\mathbf{t}}{\text{minimize}} \quad \Delta(\mathbf{Q}(i)) \triangleq \mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\left\{\mathrm{L}\left[\mathbf{Q}(i+1)\right] - \mathrm{L}\left[\mathbf{Q}(i)\right]|\mathbf{Q}(i)\right\} \tag{11a}$$

$$\leq \mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\left\{\sum_{k\in\mathcal{U}}\frac{\lambda_k^2(i)+t_k^2(i)}{2}|\mathbf{Q}(i)\right\} + \sum_{k\in\mathcal{U}} Q_k(i)A_k(i)\mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\left\{\sum_{k\in\mathcal{U}} Q_k(i)t_k(i)|\mathbf{Q}(i)\right\} \tag{11b}$$

where the second term on the r.h.s follows from the Poisson arrival process. Assuming the second order moments of the transmission rates and the arrival rates are bounded, it can be replaced by a suitable bound B in order to eliminate from the optimization problem [14].

Now the expression in (11) 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 [18], where the queue weighted sum rate maximization is considered as the objective function to determine the transmit precoders. The Q-WSRM formulation is given by

$$\underset{\mathbf{M}_{k,n},\mathbf{W}_{k,n}}{\text{maximize}} \quad \sum_{k\in\mathcal{U}} Q_k\left(\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n})\right) \tag{12a}$$

$$\text{subject to.} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}\right) \leq P_{\max},\ \forall\, b \tag{12b}$$

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

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n}) \leq Q_k,\ \forall\, k\in\mathcal{U} \tag{13}$$

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

As a special case of the problem defined in (12), we can formulate the sum rate maximization problem by setting the weights in (12a) as unity, leading to the problem as

$$\underset{\mathbf{M}_{k,n},\mathbf{W}_{k,n}}{\text{maximize}} \quad \sum_{k\in\mathcal{U}}\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n}) \tag{14a}$$

$$\text{subject to.} \quad \text{(12b) and (13)} \tag{14b}$$

The problem defined in (14) provides a greedy queue minimizing approach as compared to Q-WSRME, since the resource allocation is driven by the channel conditions in comparison with the number of queued packets as in Q-WSRME. Note that in both formulations, the resources allocated to the users are limited by the backlogged packets with an explicit maximum rate constraint defined by (13).

B. JSFRA scheme via Successive Convex Approximation (SCA)

The problem defined in (12) ignores the second order term arising from the Lyapunov drift minimization objective by the limiting it to a constant value. In fact, Eq. (5) provides similar expression when the exponent is set to be $\ell_{q=2}$ as

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

It is evident that (15) is equivalent to (11) if the second order terms are ignored. Limiting t_k^2 by a constant value, the Q-WSRM formulation requires the explicit rate constraint (13) to avoid the resource wastage in the form of over allocation. In the current queue deviation formulation, the explicit rate constraint is not needed, since it is handled by the objective function itself. 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 in the system. The queue deviation objective provides an alternative approach to perform the resource allocation without the additional rate constraints as in Q-WSRME approach. In this approach, we present an algorithm to solve (6) for the transmit precoders in a centralized way by using the idea of alternating optimization and successive convex approximation. Using (2), we can reformulate the problem defined in (6) as

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

$$\text{subject to} \quad \gamma_{l,k,n} \leq \frac{\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}\right|^2}{\beta_{l,k,n}} \triangleq f(\tilde{\mathbf{u}}_{l,k,n}) \tag{16b}$$

$$\beta_{l,k,n} \geq N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{(j,i)\neq(l,k)}|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{16c}$$

$$\text{(6b)} \tag{16d}$$

where $\tilde{\mathbf{u}}_{l,k,n} \triangleq \{\mathbf{w}_{l,k,n}^{\mathrm{H}},\mathbf{H}_{b_k,k,n},\mathbf{m}_{l,k,n},\beta_{l,k,n}\}$ be the vector which needs to be identified for the optimal allocation. In this formulation, we relaxed the equality constraint in (2) by the inequalities in (16b) and (16c). However, this step is without loss of optimality leads to the same solution, since the inequalities in (16b) and (16c) are active for an optimal solution, following the same arguments as those in [2]. Intuitively, (16b) denotes the SINR constraint for $\gamma_{l,k,n}$, and (16c) gives an upper bound for the total interference seen by the user $k \in \mathcal{U}_b$, denoted by the variable $\beta_{l,k,n}$. Similar to the WSRM problem in [2], the problem can be shown to be NP-hard even for the single antenna case. The reformulation in (16) allows a tractable solution as presented below. First, we note that the constraints (6b) 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 find the optimal transmit beamformers. Recall the original coordinate descent method assumes that the optimization variables belong to disjoint sets (Cartesian product of sets, to be precise) [19].

By fixing the receivers, the problem now is to find optimal transmit beamformers for a given set of linear receivers 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 the SCA method to handle (16b) by replacing the original non-convex constraint by the series of convex constraints. Note that the function $f(\tilde{\mathbf{u}}_{l,k,n})$ in (16b) 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 $\beta_{l,k,n}$) [20]. According to the SCA method, we relax (16b) to a convex constraint in each iteration of the iterative procedure. Since $f(\tilde{\mathbf{u}}_{l,k,n})$ is convex, a concave approximation of (16b) can be easily found by considering the first order approximation of $f(\tilde{\mathbf{u}}_{l,k,n})$ around the current operation point. 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

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

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

and hence $f(\tilde{\mathbf{u}}_{l,k,n}) = (p_{l,k,n}^2 + q_{l,k,n}^2)/\beta_{l,k,n}$. Note that $p_{l,k,n}$ and $q_{l,k,n}$ are just symbolic notation and not the newly introduced optimization variables. In CVX [21], for example, we declare $p_{l,k,n}$ and $q_{l,k,n}$ with the '*expression*' qualifier. 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

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

In summary, for the fixed linear receivers, the JSFRA problem to find transmit beamformers is shown by

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

$$\text{subject to} \quad \beta_{l,k,n} \geq N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \quad (19b)$$

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

$$\text{and (18)} \quad (19d)$$

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) w.r.t $\mathbf{w}_{l,k,n}$ as

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

$$\text{subject to} \quad \text{(19b), (19c), (19d), and (18)} \quad (20b)$$

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

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

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

where $\mathbf{w}_{l,k,n}^{(i-1)}$ is the receive beamformer from the earlier iteration, upon which the linear relaxation is performed for the nonconvex constraint in (20). It can be seen that the optimal receiver expression in (21) is the scaled version of the MMSE receiver, which is given by

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

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

The proposed algorithm is referred as queue minimizing JSFRA scheme with a per BS power constraint which 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 $\tilde{\mathbf{u}}_{l,k,n}$ 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 $\tilde{\mathbf{u}}_{l,k,n}$ in an efficient manner as briefed in Algorithm 1. In Algorithm 1, the SCA iterations are carried until convergence or for maximum of $I_{\max}$ iterations for the optimal $\mathbf{w}_{l,k,n}$ receive beamformers and the outer iterations are for the convergence of the number of queued bits, which is limited by the maximum of $J_{\max}$ iterations.

Convergence: In order to prove the convergence of the proposed algorithm, we require the following conditions to be satisfied

- convergence of the SCA subproblem
- uniqueness of the transmit and the receive beamformers obtained
- monotonic convergence of the objective function

Algorithm 1: Algorithm of JSFRA scheme

Input: $a_k,\ Q_k,\ \mathbf{H}_{b,k,n},\ \ \forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n}\ \forall\, l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0$, $j = 0$ and the transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{w}_{l,k,n}$ with (22) and $\tilde{\mathbf{u}}_{l,k,n}$ with (18) using $\tilde{\mathbf{m}}_{l,k,n}$.
repeat
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (19)
    update the constraint set (18) with $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ using (17) with the precoders $\mathbf{m}_{l,k,n}$ obtained from the previous step.
    $i = i + 1$.
  until *SCA convergence or* $i \geq I_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (20) or (22) with the recent precoders $\mathbf{m}_{l,k,n}$.
  $j = j + 1$.
until *Queue convergence or* $j \geq J_{\max}$

In the proposed solution, we replaced (16b) by a convex constraint using the first order approximation, the linear approximation is majorized by the quadratic-over-linear function in (16b) from below around a fixed point $\tilde{\mathbf{u}}^{(i)}_{l,k,n}$. Since the SCA method is adopted in the proposed algorithm, the constraint approximation satisfies the following conditions as in [22]

$$\begin{aligned} f(\tilde{\mathbf{u}}_{l,k,n}) &\leq \bar{f}(\tilde{\mathbf{u}}_{l,k,n}, \tilde{\mathbf{u}}^{(i)}_{l,k,n}) && (23a)\\ f(\tilde{\mathbf{u}}^{(i)}_{l,k,n}) &= \bar{f}(\tilde{\mathbf{u}}^{(i)}_{l,k,n}, \tilde{\mathbf{u}}^{(i)}_{l,k,n}) && (23b)\\ \partial f(\tilde{\mathbf{u}}^{(i)}_{l,k,n}) &= \partial \bar{f}(\tilde{\mathbf{u}}^{(i)}_{l,k,n}, \tilde{\mathbf{u}}^{(i)}_{l,k,n}) && (23c) \end{aligned}$$

where $\bar{f}(x, x^{(i)})$ is the approximate function $\bar{f}(x)$ around the point $x^{(i-1)}$. At each iteration $(i-1)$, the optimal value x^* obtained by solving the approximate problem is a feasible solution to next approximation at (i) iteration by following (23). It can be easily verified that the optimal point of each SCA problem satisfies the KKT condition of the original problem, which in fact is the stationary point of the SCA subproblem.

The uniqueness of the transmit and the receive beamformers can be justified by forcing one antenna to be real valued to exclude the phase ambiguity arising from the complex precoders. The monotonic convergence of the objective function can be justified by the following arguments. At each SCA iteration, the relaxed subproblem is solved for the optimal transmit precoders to minimize the objective function. Since the SCA subproblem is relaxed around the $(i-1)^{\text{th}}$ optimal point for the $(i)^{\text{th}}$ iteration, the solution of the $(i)^{\text{th}}$ iteration includes optimal point from the $(i-1)^{\text{th}}$ iteration as well. Therefore, at each SCA iteration, the objective function can either be equal to or smaller than the previous solution, thereby leading to the monotonic convergence of the objective function.

Once the problem converged to an optimal point, the receive beamformers are updated based on the receivers mentioned in (21) or (22). Since the receiver minimizes the objective value for the fixed transmit precoders, the proposed JSFRA scheme guarantees to converge to an optimal point.

C. JSFRA scheme via MSE reformulation

In this section, we solve the JSFRA problem by exploiting the equivalence between the MSE and the achievable capacity for the receivers designed based on the MMSE criterion [3], [4]. The MSE for the data symbol, represented as $\epsilon_{l,k,n}$, is given by

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

where $\mathbf{m}_{l,k,n}, \mathbf{w}_{l,k,n}$ denotes the transmit and the receive beamformer and $\hat{d}_{l,k,n}$ is the received symbol as in (1). Now, replacing the receive beamformer in (24) with the MMSE receiver shown in (22), we obtain the following relation between the MSE and the SINR as

$$\epsilon_{l,k,n} = \frac{1}{1 + \gamma_{l,k,n}} \tag{25}$$

where $\gamma_{l,k,n}$ is the received SINR as in (2). Using the equivalence in (25), 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 [5]–[7].

Let $v'_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ denote the queue deviation corresponding to the user k and $\tilde{v}'_k \triangleq a_k^{1/q} v'_k$ represents the weighted equivalent. Now, using the MSE relation, (16) is written as

$$\begin{aligned} \underset{\substack{t_{l,k,n}, \mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n}, \mathbf{w}_{l,k,n}}}{\text{minimize}} \quad & \|\tilde{\mathbf{v}}'\|_q\\ \text{subject to} \quad & t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n})\\ & \epsilon_{l,k,n} \geq \big|1 - \mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\big|^2 + \sum_{(j,i)\neq(l,k)} \big|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,i,n}\mathbf{m}_{j} \\ & \text{and (6b)} \end{aligned}$$

where (26c) bounds the MSE by $\epsilon_{l,k,n}$ and (26b) relaxes the transmitted rate $t_{l,k,n}$ using the MSE relation.

The queue minimizing JSFRA problem using alternative MSE formulation given by (26) is non-convex even for the fixed $\mathbf{w}_{l,k,n}$ due to the constraint (26b). In order to solve this efficiently, we use the SCA method as discussed earlier in Section II-B by using the linear under estimator for the convex function on the r.h.s of (26b). The first order Taylor approximation around a fixed MSE value $\tilde{\epsilon}_{l,k,n}$ for (26b) is given by

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

Now, using the above approximation for the rate constraint, the problem defined by (26) can be solved optimal transmit precoders $\mathbf{m}_{l,k,n}$, MSEs $\epsilon_{l,k,n}$ and the users rates over each sub-channel $t_{l,n,k}$ for 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

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

$$\text{subject to} \quad \text{(6b), (27)} \tag{28b}$$

$$\epsilon_{l,k,n} \geq \left|1 - \mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,i,n}\mathbf{m}_{j,i,n}\right|^2 + N_0 \|\mathbf{w}_{l,k,n}\|^2 \tag{28c}$$

Convergence: Following the similar argument from Section III-B, we can see that the SCA subproblems converges to a stationary point. The transmit precoders obtained by the MSE relaxations are unique due to the MSE expression in (28c), since it removes the phase ambiguity due to the real valued expression for the transmit precoders as

$$\epsilon_{l,k,n} \geq 1-2\Re\{\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\}+\sum_{\forall(j,i)} \left|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,i,n}\mathbf{m}_{j,i,n}\right|^2 + N_0 \|\mathbf{w}_{l,k,n}\|^2 \tag{29}$$

where $\Im\left\{\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\}$ is zero due to the real valued MSE. Since the transmit precoders are obtained by minimizing the weighted MSE criterion, it converges to the stationary KKT point at each iteration. The receiver updates are made based on the MMSE criterion, it provides the optimal receive beamformers for the given transmit precoders. Alternating between these two precoder designs guarantees the monotonicity of the objective function used in the problem (28). The MSE relaxation method converges to the same stationary KKT point as of the original WSRM problem by making the gradients equal.

D. Reduced Complexity Resource Allocation (Per Sub-Channel 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 provide an alternative sub-optimal solution, in which the precoders are designed over each sub-channel independently in a sequential manner by taking the remaining number of queued bits in the formulation. The centralized precoder design can also be performed across each sub-channel by using the decomposition methods discussed in [9], [10].

The proposed queue minimizing spatial resource allocation (SRA) formulation enables us to solve for the transmit precoders of all the users associated with the coordinating BS in the set $\mathcal{B}$ over each sub-channel independently by fixing the transmit power on each sub-channel to a constant value $P_{\max,n}$ as compared to the global power constraint defined by (6b). In contrast to the decomposition based approach for the sub-channel wise resource allocation, where the primal/dual variables are exchanged, this method requires the update on the number of queued bits before each sub-channel wise optimization. The total number of queued bits for each user are updated by the difference between the total number of queued bits present during the current slot to the total number of bits that are guaranteed by the earlier sub-channel allocations for the same slot as

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

where $Q_{k,n}$ is the total number of queued bits used in the optimization problem carried out for the sub-channel n. In the expression (30), Q_k denotes the total number of queued bits waiting to be transmitted for the user k during the current slot and $t_{l,k,j}$ is the rate or guaranteed bits allocated over the sub-channel j. The current scheme is sensitive to the order in which the sub-channels are selected for the optimization problem. The algorithmic representation of the SRA scheme is shown in Algorithm 2

Algorithm 2: Algorithm of SRA scheme

Input: $a_k,\ Q_k,\ \mathbf{H}_{b,k,n},\ \ \forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U}$
Input: permute $\mathcal{N} \rightarrow \tilde{\mathcal{N}}$
for $n \leftarrow 1$ **to** N **do**
  update $Q_{k,n}$ using (30) and let $\hat{n} = \tilde{\mathcal{N}}(n)$.
  Output: $\mathbf{m}_{l,k,\hat{n}}$ and $\mathbf{w}_{l,k,\hat{n}}\ \forall\, l \in \{1,2,\ldots,L\}$
  Initialize: $i = 0,\ j = 0$ and the transmit precoders $\tilde{\mathbf{m}}_{l,k,\hat{n}}$ randomly satisfying per sub-channel power constraint $P_{\max,\hat{n}}$
  update $\mathbf{w}_{l,k,\hat{n}}$ with (22) and $\tilde{\mathbf{u}}_{l,k,\hat{n}}$ with (18) using $\tilde{\mathbf{m}}_{l,k,\hat{n}}$.
  repeat
    repeat
      solve for the transmit precoders $\mathbf{m}_{l,k,\hat{n}}$ using (19)
      update the constraint set (18) with $\tilde{p}_{l,k,\hat{n}}$, $\tilde{q}_{l,k,\hat{n}}$ and $\tilde{\beta}_{l,k,\hat{n}}$ using (17) with the precoders $\mathbf{m}_{l,k,\hat{n}}$ obtained from the previous step.
      $i = i + 1$.
    until *SCA convergence or* $i \geq I_{\max}$
    update the receive beamformers $\mathbf{w}_{l,k,n}$ using (22) with the recent precoders $\mathbf{m}_{l,k,n}$.
    $j = j + 1$.
  until *Queue convergence or* $j \geq J_{\max}$
end

IV. Distributed Solutions

This section addresses the distributed precoder designs for the proposed JSFRA scheme. The formulation in (19) or (28) 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 the

(19), which is based on the Taylor approximation for the nonconvex constraint. The following discussions are equally valid for the MSE based solution outlined in (28) as well. Since the objective of (19) can be decoupled across each BS, the centralized problem can be equivalently written as

$$\begin{aligned} \underset{\substack{\gamma_{l,k,n},p_{l,k,n},q_{l,k,n}\\ \mathbf{M}_{k,n},\mathbf{W}_{k,n},\beta_{l,k,n}}}{\text{minimize}} \quad & \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q && \text{(31a)}\\ \text{subject to} \quad & \text{(19b)} - \text{(19d)}, && \text{(31b)} \end{aligned}$$

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

Following the similar approach as in [11], [12], the coupling constraint in (19b) or (28c) can be expressed by grouping the interference contribution from each BSs in the coordinating set B as

$$\beta_{l,k,n} \geq \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2 + \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} + N_0\|\mathbf{w}_{l,k,n}\|^2 \quad (32)$$

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

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

The coupling variable $\beta_{l,k,n}$ can be decoupled using the variable $\zeta_{l,k,n,b}$, which limits the interference caused by the transmission from BS b to the user k^{th} corresponding data stream. In order the solve for the global optimal precoders, we need to find the coupling variables $\zeta_{l,k,n,b}$ by either primal decomposition or by dual decomposition method. In both approaches, the coupling constraint (19b) for the SCA and (28c) 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 (32), the dual variables corresponding to the interference constraints are exchanged between the cooperating BSs $\mathcal{B}$ to update the interference variables $\zeta_{l,k,n,b}$ for the next iteration and is carried out until convergence. The primal approach is discussed extensively for the min-power problem in [11] and much of the current work follows the same. Much of the details are provided in the Appendix A.

Convergence: The convergence of the primal is similar to that of the centralized problem if the interference variables $\zeta_{l,k,n,b}$ are allowed to converge to the optimal values. But in practice, we can limit the number of exchanges to $J_{\max}$ after which 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 [23].

2) ADMM approach: In this section, we discuss the ADMM decomposition method, which is basically based on the dual decomposition, but 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 [9], [10]. In order to decouple the problem (31), the coupling variables $\zeta_{l,k,n,b}$ in (32) 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

$$\begin{aligned} \boldsymbol{\zeta} = \big[&\zeta_{1,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(|\bar{\mathcal{U}}_1|),1,1},\\ &\ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),1,N_B}, \ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),N,N_B}\big] && \text{(34a)}\\ n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| &= NL \sum_{b\in\mathcal{B}} |\bar{\mathcal{U}}_b| && \text{(34b)} \end{aligned}$$

Let $\boldsymbol{\zeta}(b_k)$ denotes the consensus entries corresponding to the BS b_k. Let $\boldsymbol{\nu}^{\{b_k\}}$ represents 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 $\zeta_{l,k,n,b}^{\{b_k\}}$ at the BSs 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 [10], [24]. 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 the BS b for the i^{th} iteration is given by

$$\begin{aligned} \underset{\substack{\gamma_{l,k,n},\mathbf{W}_{k,n},\mathbf{M}_{k,n}\\ \beta_{l,k,n},\boldsymbol{\zeta}^{\{b\}(i)}}}{\text{minimize}} \quad & \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}^{\{b\}(i-1)T} \left(\boldsymbol{\zeta}^{\{b\}(i)} - \boldsymbol{\zeta}^{(i-1)}(b)\right) + \frac{\rho}{2} \Big\| \underbrace{\boldsymbol{\zeta}^{(i-1)}(}_{\text{consens}} \\ \text{subject to} \quad & \beta_{l,k,n} \geq \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b,k,n} \mathbf{m}_{j,k,n}|^2 + \sum_{\hat{b}\in\bar{\mathcal{B}}_b} \zeta_{l,k,n,\hat{b}}^{\{b\}(i-1)} \\ & \qquad + \sum_{i\in\mathcal{U}_b\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b,k,n} \mathbf{m}_{j,i,n}|^2 + N_0\|\mathbf{v} \\ & \zeta_{l',k',n,b}^{\{b\}(i)} \geq \sum_{k\in\mathcal{U}_{\hat{b}}} \sum_{l=1}^{L} |\mathbf{w}_{l',k',n}^{\mathrm{H}} \mathbf{H}_{b,k',n} \mathbf{m}_{l,k,n}|^2, \ \forall k' \in \bar{\mathcal{U}}_b, \\ & \text{(18) and (40b)} \end{aligned}$$

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 (35) 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 $\boldsymbol{\zeta}^{(i)}$ relates exactly two BSs only, each entry in the $\boldsymbol{\zeta}^{(i)}$ can be updated by exchanging the local copies from the corresponding two BSs only. 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 the BS b as in $\zeta^{\{b\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b}$ as

$$\zeta^{(i)}_{l,\mathcal{U}_{b_k}(1),n,b} = \frac{1}{2}\left(\zeta^{\{b\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b} + \zeta^{\{b_k\}(i)}_{l,\mathcal{U}_{b_k}(1),n,b}\right) \tag{36}$$

The dual variable entries in the vector $\boldsymbol{\nu}^{\{b_k\}}$, which is the stacked dual variables corresponding to the interference equality constraint at the BS b_k, are updated using the subgradient as

$$\nu^{\{b_k\}(i)}_{l,k,n,b} = \nu^{\{b_k\}(i-1)}_{l,k,n,b} + \rho\left(\zeta^{\{b_k\}(i)}_{l,k,n,b} - \zeta^{(i)}_{l,k,n,b}\right), \forall b, b_k \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b \tag{37}$$

Convergence: The convergence of the ADMM method follows the same argument as the centralized algorithm if each distributed algorithm is allowed to converge to the optimal value for a fixed SCA point. Since subproblem solved at each BS is convex, the ADMM method converges to the optimal value [10] for a given SCA point. The receive beamformers are updated at each SCA update provides monotonic increase in the objective function, since the MMSE receive beamformers are optimal for the fixed transmit precoders obtained by solving the subproblems until convergence. The algorithmic representation of the ADMM based approach for decentralization is given in Algorithm 3.

Algorithm 3: Decentralization via ADMM for 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 the transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{w}_{l,k,n}$ with (22) and $\tilde{\mathbf{u}}_{l,k,n}$ with (18) using $\tilde{\mathbf{m}}_{l,k,n}$
initialize the global interference vectors $\boldsymbol{\zeta}^{(0)} = \mathbf{0}^{\mathrm{T}}$
initialize the interference threshold $\boldsymbol{\nu}^{\{b\}(0)} \forall b \in \mathcal{B} = 0$
for each BS $b \in \mathcal{B}$, perform the following procedure
repeat
  initialize $j = 0$
  repeat
    solve for the transmit precoders $\mathbf{M}_{k,n}$ and the local interference $\boldsymbol{\zeta}^{\{b\}}$ using (35)
    exchange $\boldsymbol{\zeta}^{\{b\}(j)}$ across the coordinating BSs in $\mathcal{B}$
    update the dual variables in $\boldsymbol{\nu}^{\{b\}(j+1)}$ using (37)
    update the global interference vector $\boldsymbol{\zeta}^{(j+1)}$ using (36)
    $j = j + 1$
  until *convergence or* $j \geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (22) by exchanging the recent precoders $\mathbf{m}_{l,k,n}$
  exchange the receive precoders $\mathbf{W}_{k,n}\ \forall k \in \mathcal{U}_b$ among the BSs in $\mathcal{B}$
  update $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ with the recent precoders using (17) and (16c) for the SCA approach (or)
  update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (28c) with equality for the MSE formulation approach
  $i = i + 1$
until *convergence or* $i \geq I_{\max}$

B. Decomposition using KKT equations in MSE formulation

The distributed solutions via primal and ADMM approaches depend on the subgradient update by using a step size parameter for the coupling variables, which affects the speed of convergence to the optimal value. In this method, we provide an alternative approach to decentralize the MSE equivalent problem considered in [5], [6] by solving the KKT conditions. Similar work has been considered for the WSRM problem with the minimum rate constraints in [8], [25]. When the queues are involved, the maximum rate constraint imposed by the number of queued packets at the BS includes a nonconvex constraint, which makes the problem difficult to solve due to the additional nonconvex maximum rate constraint (13) for the WSRM problem.

Even though the rate constraints are implicitly present in the objective function, we cannot formulate the KKT conditions readily due to the non-differentiable objective function. The non-differentiability of the objective function is due to the absolute operator present in the norm function. In order to make the objective function differentiable, we consider the following case 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 and the queues as well.

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

$$\underset{\substack{t_{l,k,n},\mathbf{M}_{k,n},\\ \epsilon_{l,k,n},\mathbf{W}_{k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{U}_b} a_k\left(Q_k-\sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^q \tag{38a}$$

subject to

$$\alpha_{l,k,n}: \quad \left|1-\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2+N_0\left\|\mathbf{w}_{l,k,n}\right\|^2 + \sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}\right|^2 \le \epsilon_{l,k,n} \tag{38b}$$

$$\sigma_{l,k,n}: \quad -\log(\tilde{\epsilon}_{l,k,n})-\frac{(\epsilon_{l,k,n}-\tilde{\epsilon}_{l,k,n})}{\tilde{\epsilon}_{l,k,n}} \ge t_{l,k,n}\log(2) \tag{38c}$$

$$\delta_b: \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\operatorname{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right)\le P_{\max},\ \forall\, b \tag{38d}$$

where $\alpha_{l,k,n},\sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (38b), (38c) and (38d).

By forming the Lagrangian of (38) with the corresponding dual variables as shown in (38), we can obtain the KKT expressions by differentiating with respect to the variables present in the problem as detailed in Appendix B. Upon solving the KKT expressions, the iterative solution $\forall k\in\mathcal{U}$, $\forall n\in\{1,\ldots,N\}$ and $\forall l\in\{1,\ldots,L\}$ is given by

$$\mathbf{m}_{l,k,n}^{(i)}=\Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\alpha_{y,x,n}^{(i-1)}\mathbf{H}_{b_k,x,n}^{\mathrm{H}}\mathbf{w}_{y,x,n}^{(i-1)}\mathbf{w}_{y,x,n}^{\mathrm{H}\,(i-1)}\mathbf{H}_{b_k,x,n}+\delta_b\mathbf{I}_{N_T}\Big)^{-1}\alpha_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)} \tag{39a}$$

$$\epsilon_{l,k,n}^{(i)}=\left|1-\mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}^{(i)}\right|^2+\sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}^{(i)}\right|^2+N_0\left\|\mathbf{w}_{l,k,n}^{(i)}\right\|^2 \tag{39b}$$

$$t_{l,k,n}^{(i)}=-\log_2(\epsilon_{l,k,n}^{(i-1)})-\frac{\left(\epsilon_{l,k,n}^{(i)}-\epsilon_{l,k,n}^{(i-1)}\right)}{\log(2)\,\epsilon_{l,k,n}^{(i-1)}} \tag{39c}$$

$$\sigma_{l,k,n}^{(i)}=\max\left(0,\frac{a_k\,q}{\log(2)}\left(Q_k-\sum_{n=1}^{N}\sum_{l=1}^{L}t_{l,k,n}^{(i)}\right)^{(q-1)}\right) \tag{39d}$$

$$\alpha_{l,k,n}^{(i)}=\alpha_{l,k,n}^{(i-1)}+\rho\left(\frac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i)}}-\alpha_{l,k,n}^{(i-1)}\right) \tag{39e}$$

$$\mathbf{w}_{l,k,n}^{(i)}=\Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\mathbf{H}_{b_x,k,n}\mathbf{m}_{y,x,n}^{(i)}\mathbf{m}_{y,x,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_x,k,n}^{\mathrm{H}}+\mathbf{I}_{N_R}\Big)^{-1}\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}^{(i)} \tag{39f}$$

Since the dual variables $\alpha^{(i)}$ and $\sigma^{(i)}$ are interdependent in the expressions in (39e), one has to be fixed to optimize for the other. In this problem, we fix the dual variable $\alpha^{(i)}$ to a fixed value as in (39e) to obtain the other variables in (39). It can be seen that, 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 thereby forcing the dual variable $\alpha_k^{(i)}<\alpha_k^{(i-1)}$, which in turn reduce the transmit precoder weights in (39a). This reduction in the precoder weights forces the allocated rate of the user k to reduce from the previous iteration as $t_k^{(i)}<t_k^{(i-1)}$.

The KKT solutions provided in (39) are solved in an iterative manner by initializing the transmit and the receive precoders $\mathbf{M}_{k,n},\mathbf{W}_{k,n}$ with the single user beamforming and the MMSE vectors. The dual variable α's corresponding to precoder weights are initialized with ones to provide equal priorities to all streams. Now, the closed form expressions in (39) are evaluated sequentially until convergence or to a certain accuracy. In (39), all expressions are in closed form except the transmit precoders (39a), which depends on the BS specific dual variable δ_b. It can be solved efficiently by the bisection method satisfying the power constraint (38d). After each iteration instant, the transmit and the receive precoders are updated across the coordinating BSs in $\mathcal{B}$ to obtain the next operating point.

In order to perform the distributed approach, we can consider the method proposed in [23], where the users evaluate the MSE, and the dual variable $\alpha_{l,k,n}$ using the transmission made by all BSs using the updated transmit precoders $\mathbf{m}_{l,k,n}^{(i-1)}$. Once the dual variables are evaluated, the users will notify the dual variables and the receive beamformers to the BSs using uplink precoded pilots, where the uplink precoder is given by $\tilde{\mathbf{w}}_{l,k,n}^{(i-1)}=\sqrt{\alpha_{l,k,n}^{(i-1)}}\mathbf{w}_{l,k,n}^{*(i-1)}$. Upon receiving the uplink precoded pilots at the BS b, the effective channel $\mathbf{H}_{b,k,n}^{\mathrm{T}}\tilde{\mathbf{w}}_{l,k,n}^{(i-1)}$ can be measured and used in the expression (39a) to update the transmit precoders, where $\mathbf{x}^*$ represents the conjugate of $\mathbf{x}$. The algorithmic representation of the distributed MSE-KKT scheme is shown in the Algorithm. 4.

Algorithm 4: 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: $i=1$ and the receive beamformers $\mathbf{w}_{l,k,n}^{(0)}$ randomly
Initialize: $\epsilon_{l,k,n}^{(0)}$ randomly and the dual variables $\alpha_{l,k,n}^{(0)}=1$
for each BS $b\in\mathcal{B}$, perform the following procedure
set the maximum iteration counter $I_{\max},J_{\max}$ to a valid number.
initialize $j=0$.
repeat
  initialize $i=0$.
  repeat
    update the transmit precoders $\mathbf{M}_{k,n}^{(i)}$ using (39a), where δ_b is identified by the bisection method satisfying (38d).
    update $\epsilon_{l,k,n}^{(i)}$, $t_{l,k,n}^{(i)}$, $\sigma_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ using (39b) and (39c), (39d) and (39e) at the receivers using the transmit precoders $\mathbf{M}_{k,n}^{(i)}$ from all BSs.
    perform precoded uplink pilot transmission from all users with the precoder $\tilde{\mathbf{w}}_{l,k,n}^{(j)}$.
    $i=i+1$
  until *until convergence or* $i\ge I_{\max}$
  evaluate the receive beamforming vector $\mathbf{W}_{k,n}^{(j+1)}$ using (39f).
  inform BSs using uplink precoded pilot transmission from all users with the precoder $\tilde{\mathbf{w}}_{l,k,n}^{(j+1)}$.
  $j=j+1$.
until *until convergence or* $j\ge J_{\max}$

Convergence: The iterative method presented in Algorithm 4 converges to the stationary point if the dual variables $\alpha_{l,k,n}$ are allowed to converge. The convergence of the dual variable is guaranteed, since the problem is convex by fixing the receive precoders $\mathbf{w}_{l,k,n}$ and the operating MSE point $\epsilon_{l,k,n}$ [10]. Once the dual variables are converged or iterated to a certain accuracy, the receivers are updated using the MMSE objective. In this algorithm, when $t_k > Q_k$, $\sigma_{l,k,n}$ will be zero (dual feasibility), thereby reducing the priority weights $\alpha_{l,k,n}$ present in the transmit precoder expression in (39a).

V. Simulation Results

The simulations carried out in this work considered 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

In this section, we discuss the performance of the centralized algorithms from Section III for a few 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 $Q_k = 5, 8$ and 6 bits respectively.

Fig. 1 plots the sub-channel wise channel seen by the users followed by the assigned rates over each sub-channel by three different algorithms, Q-WSRM allocation, JSFRA scheme and the band-wise Q-WSRM scheme using WMMSE precoder design proposed in [6]. The performance metric used for the comparison of different algorithms is the total number of backlogged bits left over at each slot after the allocation, which is denoted by $\chi = \sum_{k=1}^{K} [Q_k - t_k]^+$. It can be seen from Fig. 1, that the Q-WSRM scheme provides more priority to the third user with $Q_3 = 6$ bits compared to the first user with $Q_1 = 5$ bits while allocating the first sub-channel. In contrast to the Q-WSRM scheme, the JSFRA scheme assigns the first user on the first sub-channel thereby reducing the total number of backlogged packets waiting at the transmitter.

In order to understand the behavior in 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 users are located with the maximum interference seen from the neighboring BSs is limited to -6 dB, thereby simulating a realistic scenario. Fig. 2a shows the performance of the centralized schemes for a single receive antenna system. It compares the total number of SCA updates required by the JSFRA, SRA and the Q-WSRME schemes to perform the optimal allocations to minimize the total number of backlogged packets.

In the sub-channel wise allocations, where the total transmit power is shared equally among the sub-channels, the precoders are designed for each sub-channels independently. In this approach, the precoders are coupled via the number of queued bits which are updated using (30) before designing the precoders for the sub-channels. At each SCA points, the number

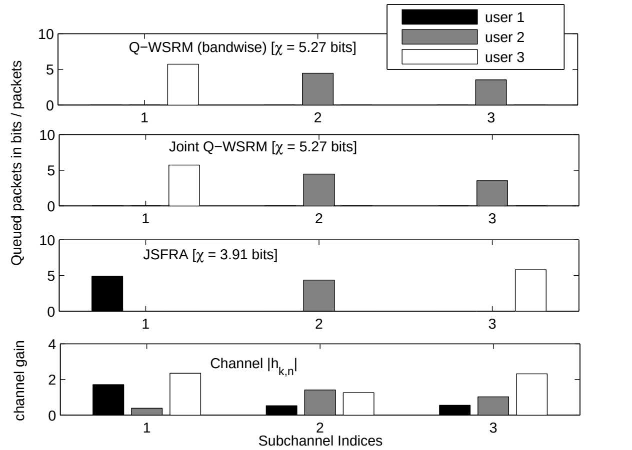


Fig. 1: Allocation for a SISO Model comparing different Algorithms

q	user indices								χ
1	15.00	3.95	5.26	8.95	7.03	11.90	12.00	9.73	25.15
2	11.23	3.93	10.76	10.65	10.27	9.68	8.77	5.90	27.77
∞	11.41	4.41	10.41	10.41	10.41	8.41	8.41	6.41	28.68
Q_k	15.0	8.0	14.0	14.0	14.0	12.0	12.0	10.0	

TABLE I: Queue information for the system $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$

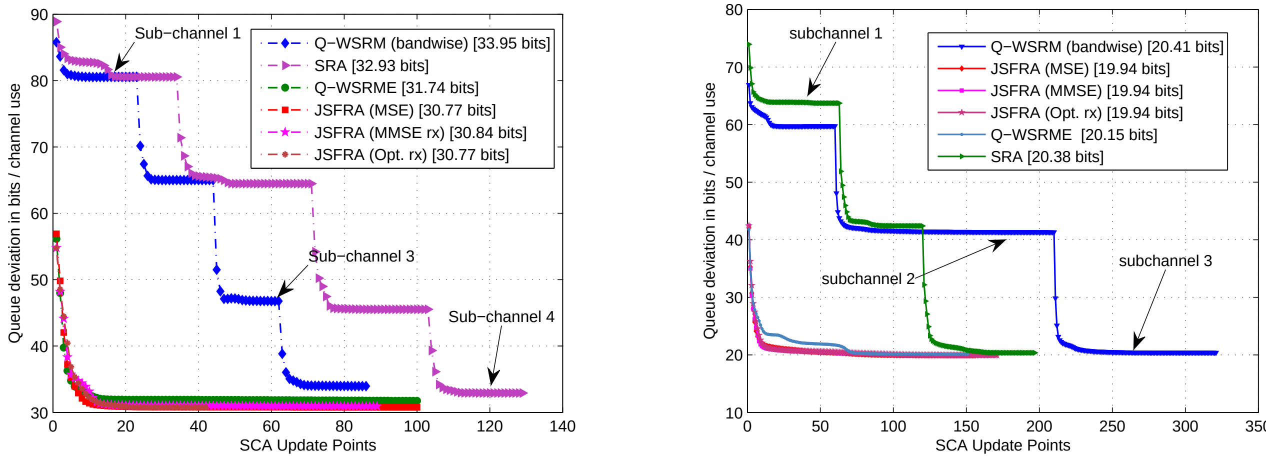
of queued bits are reduced significantly with the introduction of a new sub-channel since the algorithm starts with a random initial point before it converges to an optimal[5] precoders. The total number of backlogged bits at each SCA update instant are plotted in Fig. 2a for the discussed centralized schemes and the convergence point are marked with the data tips. Fig. 2b compares the performance of the centralized algorithms for the $N_R = 2$ receive antennas. In all the schemes, the receive beamformers are updated at the SCA update points instead of updating after the SCA convergence as in Algorithm. 1. 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 are outlined in the Table. I for the users located at the cell-edge of the system employing $N_T = 4$ transmit antennas. The configuration is mentioned in the caption of Table. I along with the number of queued bits for each user. It is evident that the algorithm minimizes the 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. The ℓ_∞ norm provides the fair allocation by making the queued deviation equal for all the users after the current allocation irrespective of their channel gains.

B. Distributed Solutions

The performance of the distributed algorithms are studied in this section by comparing with the centralized algorithms.

[5]due to the nonconvex nature of the original problem



(a) System $\{N, N_B, K, N_T, N_R\} = \{4, 3, 9, 4, 1\}$

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

Fig. 2: Number of backlogged packets at the SCA update points

The performance is compared using the total number of backlogged packets after each SCA update points. Fig. 3a 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. As pointed out in Section IV, the performance and the convergence speed of the distributed algorithms are susceptible to the choice of the step size used in the subgradient update. Since the interference values are fixed for local subproblem in the primal approach, it may lead to infeasible solutions if the initial or any intermediate update using (44) is not valid.

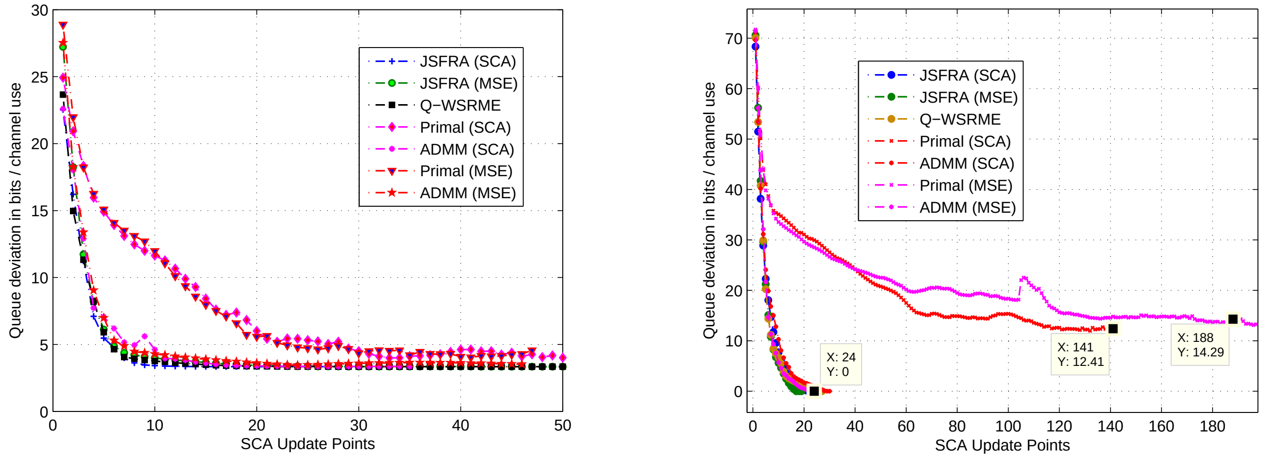
The Fig. 3a plots the performance of the primal and the ADMM solutions for the JSFRA scheme using SCA and by MSE relaxation at each SCA points. In between the SCA updates, the primal or the ADMM schemes are performed for $J_{\max} = 20$ iterations by exchanging either the fixed interference variables in the primal approach or the dual variables in the ADMM scheme. It can be seen from Fig. 3a that the distributed algorithms approaches the centralized performance by exchanging minimal information between the coordinating BSs.

In Fig. 3b, the performance of the distributed algorithms are studied for $K = 12$ users utilizing $N = 6$ sub-channels. The system considers $N_B = 3$ BSs, each having $N_T = 4$ transmit antennas serving $|\mathcal{U}_b| = 4$ users equipped with $N_R = 2$ antennas respectively. The users are assumed to be scattered over the cell boundary experiencing the path loss following the uniform distribution between $[0, -6]$ dB. The performance of the algorithms are similar to the $N_B = 2$ BS scenario discussed earlier.

Fig. 3b plots the performance of the centralized and the distributed algorithms at each SCA update. In case of the distributed algorithms, in between each SCA update, the primal or the ADMM exchanges are performed for $J_{\max} = 20$ iterations. In practice, $J_{\max} = 1$ can be set to perform the SCA update, ADMM or primal update, and the receive beamformers $\mathbf{W}_{k,n}$ update at the same instant. The data tips are used to highlight the convergent points of various algorithms. The performance of the JSFRA schemes using the primal decomposition are notably inferior compared to the ADMM approach for the same schemes. It is mainly attributed to the difficulty in selecting the step size for the system employing $N_B \geq 3$ BSs.

Fig. 4 compares the performances of the centralized algorithm based on the MSE reformulation with the iterative approach proposed in Section IV-B based on solving the KKT conditions. In Fig. 4, the plots are compared by the surplus number of backlogged packets at the end of each iteration or the SCA update point. Fig. 4 shows that the ℓ_1 norm for JSFRA scheme provides better performance over rest of the schemes due to its 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, thereby performs the worst of all other approaches. The heuristic method used in the figure is obtained by forcing the dual variable $\sigma_{l,k,n}$ in (39d) to 0 when the queue deviation is negative $Q_k - t_k < 0$, if not, then it will be the same as in (39d). The additional condition can be justified due to the dropping of absolute value operator from the objective. It can be seen that the heuristic method oscillates near the optimal point with the deviation determined by the factor ρ used in (39e).

The objective values are mentioned in the legend for all the schemes, since the objective of ℓ_2 norm is not the same as ℓ_1 norm, which is used for the plot. The ℓ_2 norm for the JSFRA and the KKT based approach achieves nearly the same objective value of 6.62 but different χ, since the dual variables $\alpha_{l,k,n}$ and $\sigma_{l,k,n}$ are not iterated until convergence between each SCA steps in the KKT approach. In the simulation, we update all the variables at each iteration due to the practical issues in the signaling between the BSs. Fig. 4 also compares the effect of dropping the squared rate variable from the objective in the Q-WSRME scheme compared to the ℓ_2 norm



(a) System $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$

(b) System $\{N, N_B, K, N_R\} = \{6, 3, 12, 2\}$

Fig. 3: Number of backlogged packets at each SCA points

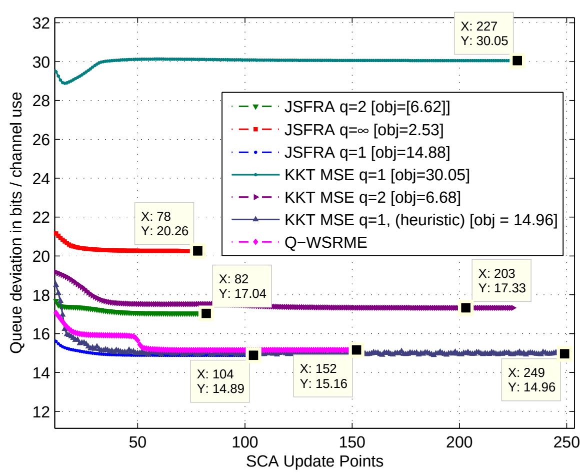


Fig. 4: Convergence plot for $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$ model

which includes it. By dropping the squared rate variable, 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 number of queued packets before addressing the users with the smaller number of queued packets.

Acknowledgment

This work has been supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia Solutions Networks, Xilinx Ireland, Academy of Finland.

VI. Conclusions

In this work, we presented the queue minimizing objective to minimize the number of backlogged packets by designing the transmit precoders across the space-frequency dimension in a multi-cell multi-user (MU) multiple-input multiple-output (MIMO) system. We proposed the joint space-frequency resource allocation (JSFRA) scheme, which adopts the successive convex approximation (SCA) technique to model the nonconvex constraint as a convex constraint in an iterative manner to design the precoders for the queue minimizing objective. We also proposed an alternative approach using the mean squared error (MSE) relaxation for the same problem when the receive beamformers are based on the minimum mean squared error (MMSE) receivers. The exponent used in the proposed formulation can be modified to achieve either greedy allocation or equally fair allocation. We also proposed the distributed solutions for the proposed centralized JSFRA problem using the alternating directions method of multipliers (ADMM) method, which has better convergence behavior compared to the primal decomposition. In addition, we also proposed an iterative algorithm based on the Karush-Kuhn-Tucker (KKT) conditions for the MSE reformulated JSFRA scheme to design the precoders in a distributed manner using the closed form solutions.

Appendix A
Primal Decomposition Approach

By fixing the interference values ζ_{l',k',n,b_k} corresponding to the interference from the base station (BS) b_k, the constraint involving the coupling variables (19b) can be relaxed using the equivalent formulation in (32). Now, the subproblem for the BS $b_k \in \mathcal{B}$ can be obtained by grouping the terms relevant to the BS b_k as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{M}_{k,n}\\ \mathbf{W}_{k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_{b_k}\|_q \tag{40a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_{b_k}} \mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}\right) \le P_{\max}, \tag{40b}$$

$$\beta_{l,k,n} \ge \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}}\sum_{j=1}^{L}|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} + N_0\|\mathbf{w}_{l,k,n}\|^2 \tag{40c}$$

$$\zeta_{l',k',n,b_k} \ge \sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}|\mathbf{w}_{l',k',n}^{\mathrm{H}}\mathbf{H}_{b_k,k',n}\mathbf{m}_{l,k,n}|^2, \ \forall k' \in \bar{\mathcal{U}}_{b_k}, \ \forall n \in \mathcal{C} \tag{40d}$$

$$\text{and (18),} \tag{40e}$$

Let $\boldsymbol{\zeta}^{\{b_k\}}$ be the vector representing the fixed interference levels relevant to the BS b_k in a fully connected network[6], which is given by

$$\boldsymbol{\zeta}_{k,n,b} = [\ \zeta_{1,k,n,b}, \dots, \zeta_{L,k,n,b}\ ] \tag{41a}$$

$$\boldsymbol{\zeta}_n^{\{b_k\}} = \Big[\ \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(1)}, \dots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \dots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(|\mathcal{U}_{b_k}|),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \dots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(1),n,b_k}, \dots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(|\bar{\mathcal{U}}_{b_k}|),n,b_k}\ \Big] \tag{41b}$$

$$\boldsymbol{\zeta}^{\{b_k\}} = \Big[\ \boldsymbol{\zeta}_1^{(b_k)}, \dots, \boldsymbol{\zeta}_N^{(b_k)}\ \Big], \tag{41c}$$

where the length of the vector $\boldsymbol{\zeta}^{\{b_k\}}$ is

$$n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| = \left(|\bar{\mathcal{B}}_{b_k}||\mathcal{U}_{b_k}| + |\bar{\mathcal{U}}_{b_k}|\right) LN \tag{42}$$

The local subproblem (40) solved at each BS are coordinated by the master problem, which updates the interference thresholds $\boldsymbol{\zeta}^{\{b\}}, \forall b \in \mathcal{B}$ for the next iteration. The master problem controlling multiple subproblems is given by

$$\underset{\boldsymbol{\zeta}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \alpha_b^\star(\boldsymbol{\zeta}^{\{b\}}) \tag{43a}$$

$$\text{subject to} \quad \boldsymbol{\zeta}^{\{b\}} \in \mathbb{R}_+^{n_b}, \forall b \in \mathcal{B}, \tag{43b}$$

where $\alpha_b^\star(\boldsymbol{\zeta}^{\{b\}})$ denotes the optimal solution for (40) with the previous value of $\boldsymbol{\zeta}^{(i-1)}$, where $\boldsymbol{\zeta}$ is the global interference vector formed by stacking the interference vector associated with each BS as $\boldsymbol{\zeta} = \left[\boldsymbol{\zeta}^{\{\mathcal{B}(0)\}}, \boldsymbol{\zeta}^{\{\mathcal{B}(1)\}}, \dots, \boldsymbol{\zeta}^{\{\mathcal{B}(|\mathcal{B}|)\}}\right]$.

The master problem to find the optimal $\boldsymbol{\zeta}^{\{b_k\}(i)}, \forall b_k \in \mathcal{B}$ is given by the following subgradient method [24] as

$$\zeta_{l,k,n,b}^{\{b_k\}(i)} = \left[\zeta_{l,k,n,b}^{\{b_k\}(i-1)} - \rho\, s_{l,k,n,b}^{\{b_k\}(i-1)}\right]^{+}, \forall b \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b, \tag{44}$$

where i is the iteration index, ρ is the positive step size, and $s_{l,k,n,b}^{\{b_k\}(i-1)}$ is the subgradient of the problem defined in (43) evaluated at $\zeta_{l,k,n,b}^{(i-1)}$. To find the subgradient $s_{l,k,n,b}^{\{b_k\}(i-1)}$, the dual variables corresponding to the interference constraints are required, which can be obtained by forming the dual problem of (41) as discussed in [11]. Now the primal and the dual variables $\mu_{l,k,n}^{\{b_k\}}$ and $\mu_{l',k',n}^{\{b_k\}}$ corresponding to the constraints (40c) and (40d) can be obtained from the solvers, which solves the dual problem as well.

To obtain the next interference iterate at each BS, the locally evaluated dual variables are exchanged among the BSs in the set $\mathcal{B}$ in order to obtain the next interference vector in a distributed manner. Once we obtain the dual variables from all the BSs, subgradients relevant to the BS b_k are evaluated by taking the difference between the two BSs associated with each interference value, *i.e*, $s_{l,k,n,b}^{\{b_k\}(i)} = \mu_{l,k,n}^{\{b_k\}} - \mu_{l,k,n}^{\{b\}}$. With the newly estimated subgradient value $s_{l,k,n,b}^{\{b_k\}(i)}$, the interference terms corresponding to the BS b_k are updated using (44). The algorithmic representation of the primal decomposition (PD) approach is detailed in Algorithm. 5.

[6] in practice it will be less due to the path loss

Algorithm 5: Decentralization via Primal Decomposition for 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 $\mathbf{w}_{l,k,n}\ \forall\, l \in \{1, 2, \dots, L\}$
Initialize: $i = 0$ and the transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{w}_{l,k,n}$ with (22) and $\tilde{\mathbf{u}}_{l,k,n}$ with (18) using $\tilde{\mathbf{m}}_{l,k,n}$
initialize the interference threshold $\zeta_{l,k,n,b}^{\{0\}}\ \forall b \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_{b_k}, \forall l, n$
for each BS $b \in \mathcal{B}$, perform the following procedure
repeat
- initialize $j = 0$
- **repeat**
  - solve for the transmit precoders $\mathbf{m}_{l,k,n}$ and dual variables $\mu_{l,k,n}^{\{b\}}$ using (41)
  - exchange $\mu_{l,k,n}^{\{b\}}$ across the coordinating BSs in $\mathcal{B}$
  - update $\zeta_{l,k,n,b}^{\{b\}(j+1)}$ using (44) locally
  - $j = j + 1$
- **until** *convergence or* $j \ge J_{\max}$
- update the receive beamformers $\mathbf{w}_{l,k,n}$ using (22) by exchanging the recent precoders $\mathbf{m}_{l,k,n}$
- exchange the receive precoders $\mathbf{W}_{k,n}\ \forall k \in \mathcal{U}_b$ among the BSs in $\mathcal{B}$
- update $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ with the recent precoders using (17) and (16c) for the SCA approach (or)
- update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (28c) with equality for the MSE formulation approach
- $i = i + 1$

until *convergence or* $i \ge I_{\max}$

APPENDIX B
KKT EXPRESSIONS FOR THE DISTRIBUTED MSE FORMULATION

In order to solve for an iterative precoder design algorithm, the KKT expressions for the problem in (38) are obtained by differentiating the Lagrangian by assuming the equality constraint for (38b) and (38c). At the stationary point, the following conditions are to be satisfied.

$$\nabla_{t_{l,k,n}}: \quad -q\left[a_k\left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^{(q-1)}\right] + \sigma_{l,k,n}\log(2) = 0 \tag{45a}$$

$$\nabla_{\epsilon_{l,k,n}}: \quad -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \tag{45b}$$

$$\nabla_{\mathbf{m}_{l,k,n}}: \sum_{y\in\mathcal{U}}\sum_{x=1}^{L}\alpha_{x,y,n}\mathbf{H}_{b_k,y,n}^{\mathrm{H}}\mathbf{w}_{x,y,n}\mathbf{w}_{x,y,n}^{\mathrm{H}}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n} + \delta_b\mathbf{m}_{l,k,n} = \alpha_{l,k,n}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n} \tag{45c}$$

$$\nabla_{\mathbf{w}_{l,k,n}}: \sum_{(x,y)\neq(l,k)}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}_{x,y,n}^{\mathrm{H}}\mathbf{H}_{b_y,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n} + \mathbf{I}_{N_R}\mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n} \tag{45d}$$

in addition to the primal constraints given in (38b), (38c) and (38d), the complementary slackness criteria are given by

$$\alpha_{l,k,n}\underbrace{\Big(\epsilon_{l,k,n} - \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 - \sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}\right|^2 - N_0\left\|\mathbf{w}_{l,k,n}\right\|^2\Big)}_{=0} \tag{46a}$$

$$\sigma_{l,k,n}\underbrace{\left(\log(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\tilde{\epsilon}_{l,k,n}} + t_{l,k,n}\log(2)\right)}_{=0} \tag{46b}$$

$$\delta_b\Big(\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}\right) - P_{\max}\Big) \tag{46c}$$

In the expressions (46a) and (46b), the value inside the braces are zero due to the equality assumptions. Now, the dual variables should satisfy $\alpha_{l,k,n} \geq 0$ and $\sigma_{l,k,n} \geq 0$. The total power constraint in (46c) need not be tight to make the dual variable δ_b to be greater than zero. In cases where the total power required to obtain the desired transmission rate is strictly less than $P_{\max}$, δ_b must be zero to satisfy the complementary slackness criterion defined in (46c). Upon solving the KKT expressions in (45) and (46), we obtain the iterative algorithm defined in the Section IV-B.

References

[1] 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.

[2] 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.

[3] P. Viswanath, V. Anantharam, and D. N. C. Tse, "Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1968–1983, 1999.

[4] 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.

[5] 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.

[6] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, sept. 2011.

[7] M. Hong, Q. Li, Y.-F. Liu, and Z.-Q. Luo, "Decomposition by successive convex approximation: A unifying approach for linear transceiver design in interfering heterogeneous networks," *arXiv preprint arXiv:1210.1507*, 2012.

[8] 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.

[9] 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.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

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

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

[22] 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.

[23] 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.

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

[25] J. Kaleva, A. Tolli, and M. Juntti, "Decentralized beamforming for weighted sum rate maximization with rate constraints," in *IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*. IEEE, 2013, pp. 220–224.