Traffic aware Resource Allocation schemes for Multi-Cell MIMO-OFDM systems

Ganesh Venkatraman, Antti Tölli, Le-Nam Tran and Markku Juntti
Centre for Wireless Communications (CWC), Department of Communications Engineering (DCE),
University of Oulu, Oulu, FI-90014
Email: {gvenkatr, antti.tolli, le.nam.tran, markku.juntti}@ee.oulu.fi

***Abstract*—We consider a multi-cell multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) system with multiple-users (MU) contending for the space-frequency resources in the downlink direction. The problem is to determine the transmit precoders by the base stations (BSs) in a coordinated manner to minimize the total number of backlogged packets in the BSs, which are destined for the users in the system. Since the problem is similar to the precoder design for a MIMO interference broadcast channel (IBC) system, traditionally it is solved by the weighted sum rate maximization (WSRM) objective with the number of backlogged packets as the corresponding weights, *i.e*, longer the queue size, higher the priority. In contrast, we address the queue minimizing downlink precoder design as a joint nonconvex optimization problem over space-frequency resources. We employ successive convex approximation (SCA) technique to solve the problem by a sequence of convex subproblems using inner approximations. Initially, we discuss the centralized joint space-frequency resource allocation (JSFRA) solutions based on SCA and by mean squared error (MSE) reformulation. Then we address the distributed precoder design for the centralized schemes using primal and alternating directions method of multipliers (ADMM) method. Finally, we discuss the distributed precoder design problem by solving the Karush-Kuhn-Tucker (KKT) expressions to obtain the closed form solutions for the transmit and the receive precoders. Numerical results are shown to compare them.**

***Index Terms*—MIMO-OFDM, IBC, Precoder design, WSRM**

I. Introduction

In a network with multiple base stations (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. We consider the problem of transmit design over the space-frequency resources provided by the MIMO orthogonal frequency division multiplexing (OFDM) framework in the downlink interference broadcast channel (IBC) to minimize the number of queued packets. Since the space-frequency resources are shared by multiple users associated with different BSs, it can be viewed as a resource allocation problem.

In general, the resource allocation problems are solved by assigning a binary variable to each user indicating the presence or the absence on a particular resource [1]. 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.

The queue minimizing precoder designs are closely related to the weighted sum rate maximization (WSRM) problem with additional rate constraints determined by the number of backlogged packets for each user in the system. The topics on multiple-input multiple-output (MIMO) IBC precoder design have been studied extensively with different performance criteria in the literature. Due to the nonconvex nature of the MIMO IBC precoder design problems, the successive convex approximation (SCA) method has become a powerful tool to deal with these problems [28]. For example, in [2], 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 by using arithmetic-geometric inequality is proposed in [3].

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 [4], [5] can also be used to solve the WSRM problem. In [6], [7], the WSRM problem is reformulated via MSE, casting the problem as a convex one for fixed linearization coefficients. In this way, the original problem is expressed in terms of the MSE weight, precoders, and decoders. Then the problem is solved using an alternating optimization method, i.e., finding a set of variables while the remaining others are fixed. The MSE reformulation for the WSRM problem is also studied in [8] by using the SCA to solve the problem in an iterative manner. Additional rate constraints based on the quality of service (QoS) requirements are included in the WSRM problem and solved via MSE reformulation in [9], [10].

The problem of precoder design for the MIMO IBC system are solved either by using a centralized controller or by using decentralized algorithms where all BSs handles their own subproblems 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 [11], [12]. In the primal decomposition, the so-called coupling interference variables are fixed for the subproblem at each BS to find the

optimal precoders. The fixed interference are then updated by using the subgradient method as discussed in [13]. The dual and ADMM approach controls the distributed subproblems by fixing the *'interference price'* for each BS as detailed in [14].

By adjusting the weights in the WSRM objective properly, we can find arbitrary rate-tuple in the rate region that maximizes the suitable objective measures. For example, if the weight of each user is set to be inversely proportional to 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 [15], [16]. In backpressure algorithm, the differential queues between the source and the destination nodes are used as the weights scaling the transmission rate [17].

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

In this paper, we consider the problem of precoder design across the space-frequency resources to minimize the total number of queued packets waiting in 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 organization 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 IBC scenario in an OFDM framework with N sub-channels and N_B BSs each equipped with N_T transmit antennas, serving in total K users each with N_R receive antennas. The set of users associated with BS b is denoted by $\mathcal{U}_b$ and the set $\mathcal{U}$ represents all users in the system, i.e., $\mathcal{U} = \underset{b\in\mathcal{B}}{\cup}\mathcal{U}_b$, where $\mathcal{B}$ is the set of 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

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

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

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

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 $\lambda_k(i) \sim \text{Pois}(A_k)$ 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

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

where $[x]^+ \equiv \max\{x, 0\}$ and t_k denotes the transmission in bits for user k. For a MIMO-OFDM system,

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

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

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

(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 $\gamma_{l,k,n}$[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

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

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

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

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

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 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

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

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

Before proceeding further, we 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 [18], [19] alters the priorities for the queue deviation function as

- $\ell_{q=1}$ 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 when the queue size is large enough for all users.
- $\ell_{q=2}$ prioritizes users with higher number of queued packets before considering the users with a smaller number of backlogged packets. For example, it could be more ideal for the delay limited scenario when the packet arrival rates of the users are similar, since the number of backlogged packets is proportional to the delay in the transmission following the Little's law [16].
- $\ell_{q=\infty}$ 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 [15]–[17]. 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

[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

moves towards the undesirable state. Following the approach in [16], the scalar measure for the queue stability is given by

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

where $\mathbf{Q}(i)$ 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 [16, Ch. 3]. The Lyapunov function drift, expressed as $\Delta(\mathbf{Q}(i)) = \mathrm{L}\left[\mathbf{Q}(i+1)\right] - \mathrm{L}\left[\mathbf{Q}(i)\right]$, is given by

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

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

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

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

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

$$- \mathbb{E}_{\boldsymbol{\lambda},\mathbf{t}}\Big\{\sum_{k\in\mathcal{U}} Q_k(i)t_k(i)|\mathbf{Q}(i)\Big\} \tag{11b}$$

where the first term in (11b) follows from the Poisson arrival process.

Assuming the second order moments of the transmissions and the arrival rates are bounded, it can be replaced by a upper bound B as in (11b), in order to eliminate from the optimization problem [16]. The expectation is performed across all possible arrivals and the transmission rates for the given number of queued packets $\mathbf{Q}(i)$ and the channel state information at the i^{th} slot.

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 [20], 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 optimizing each term inside the expectation operator $\mathbf{E}_{\boldsymbol{\lambda},\mathbf{t}}$ in (11b), which maximizes the

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

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

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 rate constrained version of the Q-WSRM, denoted by Q-WSRM extended (Q-WSRME) problem for a cellular system, is given by with the additional constraint

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

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 in (12) with $Q_k = 1, \forall k \in \mathcal{U}$. This approach provides a greedy queue minimizing allocation as compared to Q-WSRME, since the resource allocation is driven by the channel conditions in comparison 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 SCA approach

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

$$\underset{t_k}{\text{minimize}}\ \sum_k v_k^2 = \underset{t_k}{\text{minimize}}\ \sum_k Q_k^2 - 2\,Q_k t_k + t_k^2 \tag{14}$$

It is evident that (14) 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 proposed queue deviation formulation, the explicit rate constraint is not needed, since it is handled by the objective function itself. This makes the problem simpler and allows us to employ efficient algorithms to distribute the precoder design problem across each BSs independently by exchanging minimal information exchange [12]. In contrast to the WSRM formulation, the JSFRA and the Q-WSRM problems include the sub-channels jointly to achieve an efficient allocation by identifying the optimal space-frequency resource for each user 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) to obtain the transmit precoders in a centralized manner by using the idea of alternating optimization and successive convex approximation. Using (2), we can reformulate the problem defined in (6) as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\\ \beta_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{15a}$$

$$\text{subject to} \quad \gamma_{l,k,n} \leq \frac{|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}|^2}{\beta_{l,k,n}} \triangleq f(\tilde{\mathbf{u}}_{l,k,n}) \tag{15b}$$

$$\beta_{l,k,n} \geq \tilde{N}_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2 \quad (15c)$$

and (6b) (15d)

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 and $\tilde{N}_0 = N_0 \|\mathbf{w}_{l,k,n}\|^2$ be the effective noise variance. In this formulation, we relaxed the equality constraint in (2) by the inequalities in (15b) and (15c). However, this step without loss of optimality leads to the same solution, since the inequalities in (15b) and (15c) are active for an optimal solution, following the same arguments as those in [3]. Intuitively, (15b) denotes the SINR constraint for $\gamma_{l,k,n}$, and (15c) 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 [3], the problem can be shown to be NP-hard even for the single antenna case. The reformulation in (15) 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 (15b) and (15c). Towards this end, we resort to the traditional coordinate descent technique by fixing the linear receivers, and finding 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) [22].

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 $\mathbf{w}_{l,k,n}$, (15c) can be written as a second-order cone (SOC) constraint. Thus, the difficulty is due to the non-convexity in (15b). To arrive at a tractable formulation, we adopt the SCA method to handle (15b) 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 (15b) 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 $\beta_{l,k,n}$) [23]. According to the SCA method, we relax (15b) 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 (15b) 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

$$p_{l,k,n} \triangleq \Re\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right\} \quad (16a)$$

$$q_{l,k,n} \triangleq \Im\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right\} \quad (16b)$$

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 [24], 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 (15b) by the following linear inequality constraint as

$$2\frac{\tilde{p}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(p_{l,k,n} - \tilde{p}_{l,k,n}) + 2\frac{\tilde{q}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(q_{l,k,n} - \tilde{q}_{l,k,n}) + \frac{\tilde{p}_{l,k,n}^2 + \tilde{q}_{l,k,n}^2}{\tilde{\beta}_{l,k,n}}\left(1 - \frac{\beta_{l,k,n} - \tilde{\beta}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\right) \geq \gamma_{l,k,n} \quad (17)$$

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

$$\underset{\substack{\mathbf{m}_{l,k,n},\\ \gamma_{l,k,n},,\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \quad (18a)$$

$$\text{subject to} \quad \beta_{l,k,n} \geq \tilde{N}_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2 \quad (18b)$$

$$\sum_{n=1}^{N} \sum_{k\in\mathcal{U}_b} \mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}\right) \leq P_{\max},\ \forall\, b \quad (18c)$$

and (17) (18d)

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

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

$$\text{subject to} \quad (18b), (18c), (18d), \text{ and } (17) \quad (19b)$$

Solving (19) using KKT conditions, we obtain the following iterative expression for the receive beamformer $\mathbf{w}_{l,k,n}$ as

$$\widetilde{\mathbf{R}}_{l,k,n} = \sum_{(j,i)\neq(l,k)} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n} \mathbf{m}_{j,i,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\, \mathbf{I}_{N_R} \quad (20a)$$

$$\mathbf{w}_{l,k,n}^{(i)} = \left(\frac{\tilde{\beta}_{l,k,n} \mathbf{m}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n}^{\mathrm{H}} \mathbf{w}_{l,k,n}^{(i-1)}}{\|\mathbf{w}_{l,k,n}^{(i-1)} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\|^2}\right) \widetilde{\mathbf{R}}_{l,k,n}^{-1} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n} \quad (20b)$$

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 (19). Note that (20) is obtained by iterating over the fixed $\mathbf{w}_{l,k,n}^{(i-1)}$ at each SCA iteration until convergence or for fixed number of iterations. It can be seen that the optimal receiver expression in (20) is in fact a scaled version of the MMSE receiver, which is given by

$$\mathbf{R}_{l,k,n} = \sum_{i\in\mathcal{U}} \sum_{j=1}^{L} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n} \mathbf{m}_{j,i,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\, \mathbf{I}_{N_R} \quad (21a)$$

$$\mathbf{w}_{l,k,n} = \mathbf{R}_{l,k,n}^{-1}\, \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n} \quad (21b)$$

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. For a fixed receive beamformer $\mathbf{w}_{l,k,n}$, the SCA iteration is carried out until convergence or for the predefined iterations, say $J_{\max}$ for the optimal transmit precoders $\mathbf{m}_{l,k,n}$. Next, the receive beamformers are updated based on either (20) or (21) using the fixed transmit precoders $\mathbf{m}_{l,k,n}$. This procedure is carried out until convergence of the queue deviation or for fixed number of iterations defined by $I_{\max}$ as outlined in Algorithm 1.

Algorithm 1: Algorithm of JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n}\, \forall\, l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0$ and transmit precoders $\tilde{\mathbf{M}}_{k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{W}_{k,n}$ and $\tilde{\mathbf{u}}_{l,k,n}$ using (21) and (17) using $\tilde{\mathbf{M}}_{k,n}$
repeat
  initialize $j = 0$
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (18)
    update the constraint set (17) with $\tilde{\mathbf{u}}_{l,k,n}$ and $\mathbf{m}_{l,k,n}$ using (16)
    $j = j + 1$
  until *SCA convergence or* $j \geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (19) or (21) with the updated precoders $\mathbf{m}_{l,k,n}$
  $i = i + 1$
until *Queue convergence or* $i \geq I_{\max}$

Convergence: In order to prove the convergence of the proposed iterative algorithm, we require the following conditions to be satisfied at each step [25]

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

In the proposed solution, we replaced (15b) by a convex constraint using the first order approximation, which is majorized by the quadratic-over-linear function in (15b) from below around a fixed point $\tilde{\mathbf{u}}_{l,k,n}^{(i)}$. Since the SCA method is adopted in the proposed algorithm, the constraint approximation satisfies the following conditions as in [26]

$$f(\tilde{\mathbf{u}}_{l,k,n}) \leq \bar{f}(\tilde{\mathbf{u}}_{l,k,n}, \tilde{\mathbf{u}}_{l,k,n}^{(i)}) \tag{22a}$$

$$f(\tilde{\mathbf{u}}_{l,k,n}^{(i)}) = \bar{f}(\tilde{\mathbf{u}}_{l,k,n}^{(i)}, \tilde{\mathbf{u}}_{l,k,n}^{(i)}) \tag{22b}$$

$$\nabla f(\tilde{\mathbf{u}}_{l,k,n}^{(i)}) = \nabla \bar{f}(\tilde{\mathbf{u}}_{l,k,n}^{(i)}, \tilde{\mathbf{u}}_{l,k,n}^{(i)}) \tag{22c}$$

where $\bar{f}(\mathbf{x}, \mathbf{x}^{(i)})$ is the approximate function of $f(\mathbf{x})$ around the point $\mathbf{x}^{*(i)}$. The stationary point of the relaxed convex problem satisfies the KKT conditions of the original nonconvex problem, which can be obtained by using conditions in (22). It can be seen that the SCA relaxed formulation converges to a local stationary point at each iteration.

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 locally optimal transmit precoders to minimize the objective function. Since the SCA subproblem is relaxed around the $i-1^{\text{th}}$ optimal point,*i.e* $\mathbf{x}^{*(i-1)}$, for the i^{th} iteration, the domain of the problem in the i^{th} step includes optimal point from the $i-1^{\text{th}}$ iteration as well. Therefore, at each SCA steps, the objective function can either be equal to or smaller than the previous value, thereby leading to the monotonic convergence of the objective function.

Once the problem is converged to a stationary transmit precoders, the receive beamformers are updated based on the receivers mentioned in (20) or (21). The monotonic nature of the objective function is preserved by the receive beamformer update, since the receiver minimizes the objective value for the fixed transmit precoders, and hence the proposed JSFRA scheme guarantees to converge to a stationary point of the original nonconvex problem.

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 [4], [5]. The MSE $\epsilon_{l,k,n}$, for the data symbol is given by

$$\mathbb{E}\big[(d_{l,k,n} - \hat{d}_{l,k,n})^2\big] = \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 + N_0 \left\|\mathbf{w}_{l,k,n}\right\|^2 = \epsilon_{l,k,n} \tag{23}$$

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 (23) with the MMSE receiver shown in (21), we obtain the following relation between the MSE and the SINR as

$$\epsilon_{l,k,n} = \frac{1}{1 + \gamma_{l,k,n}} \tag{24}$$

where $\gamma_{l,k,n}$ is the received SINR as in (2). Using the equivalence in (24), the WSRM objective can be reformulated as the weighted minimum mean squared error (WMMSE) equivalent to obtain the precoders for the MU-MIMO scenario as discussed in [6]–[8].

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, by relaxing the MSE in (23) and the rate MSE equivalence, (15) is written as

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

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

$$\left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + N_0 \left\|\mathbf{w}_{l,k,n}\right\|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 \leq \epsilon_{l,k,n} \tag{25c}$$

$$\text{and (6b)} \tag{25d}$$

The alternative MSE formulation given by (25) is nonconvex even for the fixed $\mathbf{w}_{l,k,n}$ due to the constraint (25b). We adopt the SCA method as in Section II-B to relax the constraint by a sequence of convex subsets using first order approximations around $\tilde{\epsilon}_{l,k,n}$ as

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

Now, using the above approximation for the rate constraint, the problem defined by (25) 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

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

$$\text{subject to} \quad \text{(6b), (25c), and (26)} \tag{27b}$$

Convergence: Following the similar approach in Section III-B, at each iteration, the SCA subproblems converge to a stationary point of the original nonconvex problem. The uniqueness of the precoders are justified if there is no phase ambiguity in the stationary solution. By reorganizing (25c)

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

we can see that the ambiguity in the phase rotations for the transmit and the receive beamformers are avoided by the real component in the MSE expression.

At each SCA step, the transmit precoders are obtained uniquely by minimizing (25) due to the convex nature of the relaxed problem. Using the transmit precoders, the receive beamformers are obtained using the MMSE receivers. The MMSE receiver minimizes the objective value for the fixed transmit precoders, leading to the monotonic convergence of the objective function. At each SCA step, the optimal value of the previous iteration is also included in the domain of the problem, the objective value can either decrease or stays the same after each iteration.

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 optimal approach is to decompose the problem over each sub-channel with a fixed transmit power constraint for each sub-channel. The power allocated for each sub-channel is controlled by a master problem based on different algorithms as discussed in [11], [12].

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

$$Q_{k,n} = \max\Big\{Q_k - \sum_{r=1}^{n-1}\sum_{l=1}^{L} t_{l,k,r}, 0\Big\},\ \forall\, k \in \mathcal{U} \tag{29}$$

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

IV. Distributed Solutions

This section addresses the distributed precoder designs for the proposed JSFRA scheme. The formulation in (18) or (27) requires a centralized controller to perform the precoder design for all users belonging to the coordinating BSs. In order to design the precoders independently at each BS with the minimal information exchange via backhaul, iterative decentralization methods are considered. In particular, the primal decomposition and the ADMM based dual decomposition approaches are addressed.

To begin with, let $\bar{\mathcal{B}}_b$ denote the set $\mathcal{B}\backslash\{b\}$ and $\bar{\mathcal{U}}_b$ represents the set $\mathcal{U}\backslash\mathcal{U}_b$. In order to study the decomposition based solutions, we consider the solution proposed in the (18), which is based on the Taylor approximation for the nonconvex constraint. The following discussions are equally valid for the MSE based solution outlined in (27) as well. Since the objective of (18) can be decoupled across each BS, the centralized problem can be equivalently written as

$$\underset{\gamma_{l,k,n},\mathbf{M}_{k,n},\mathbf{W}_{k,n},\beta_{l,k,n}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{30a}$$

$$\text{subject to} \quad \text{(18b)} - \text{(18d)}, \tag{30b}$$

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 [13], [14], the coupling constraint in (18b) or (25c) can be expressed by grouping the interference contribution from each BSs in the coordinating set B as

$$N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{j=1,j\neq l}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} + \sum_{i\in\mathcal{U}_{b_k}\backslash\{k\}}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 \leq \beta_{l,k,n} \tag{31}$$

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

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

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 (18b) for the SCA and (25c) for the MSE relaxation schemes are decoupled to perform the distributed precoder design problem.

A. Decomposition based Approaches

1) Primal Decomposition Approach: The primal decomposition approach decomposes the problem by fixing the interference variables $\zeta_{l,k,n,b}\forall k, b$ in order to perform the precoder design independently across each BS. Once the optimal precoders are designed at each BS with the fixed interference constraints (31), the dual variables corresponding to the interference constraints are exchanged between the cooperating BSs in $\mathcal{B}$ to update the interference variables $\zeta_{l,k,n,b}$ for the next iteration until convergence. The primal approach is discussed extensively for the min-power problem in [13] and much of the current work follows the same.

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 a stationary point. 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 [27].

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 [11], [12]. In order to decouple the problem (30), the coupling variables $\zeta_{l,k,n,b}$ in (31) are replaced by the respective local copies $\boldsymbol{\zeta}^{\{b\}}, \forall b \in \mathcal{B}$, which are then solved for an optimal solution. Now the sub problems are coupled by the global consensus vector $\boldsymbol{\zeta}$ maintaining the complete stacked interference profile of all users in the system as

$$\begin{aligned}\boldsymbol{\zeta} \;&= \big[\zeta_{1,\bar{\mathcal{U}}_1(1),1,1},\ldots,\zeta_{L,\bar{\mathcal{U}}_1(1),1,1},\ldots,\zeta_{L,\bar{\mathcal{U}}_1(|\bar{\mathcal{U}}_1|),1,1},\\ &\quad \ldots,\zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),1,N_B},\ldots,\zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),N,N_B}\big] \quad (33a)\\ n_{b_k} &= |\boldsymbol{\zeta}^{\{b_k\}}| = NL\sum_{b\in\mathcal{B}}\left|\bar{\mathcal{U}}_b\right| \quad (33b)\end{aligned}$$

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 [12], [28]. The equality constraint for the local and the consensus interference vector $\boldsymbol{\zeta}^{\{b_k\}} = \boldsymbol{\zeta}(b_k)$ present in each subproblem is relaxed by the taking the partial Lagrangian. Now, the subproblem at BS b for the i^{th} iteration is given by

$$\begin{aligned}\underset{\substack{\gamma_{l,k,n},\mathbf{W}_{k,n},\mathbf{M}_{k,n}\\ \beta_{l,k,n},\boldsymbol{\zeta}^{\{b\}(i)}}}{\text{minimize}} \quad & \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}^{\{b\}(i-1)T}\left(\boldsymbol{\zeta}^{\{b\}(i)} - \boldsymbol{\zeta}^{(i-1)}(b)\right)\\ &+\frac{\rho}{2}\Big\|\underbrace{\boldsymbol{\zeta}^{\{b\}(i)}}_{\text{local}} - \underbrace{\boldsymbol{\zeta}^{(i-1)}(b)}_{\text{consensus}}\Big\|_2^2 \quad (34a)\\ \text{subject to} \quad & \\ \beta_{l,k,n} \geq \sum_{j=1,j\neq l}^{L} & |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{\hat{b}\in\bar{\mathcal{B}}_b}\zeta_{l,k,n,\hat{b}}^{\{b\}(i-1)}\\ + \sum_{i\in\mathcal{U}_b\setminus\{k\}}\sum_{j=1}^{L} & |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{j,i,n}|^2 + N_0\|\mathbf{w}_{l,k,n}\|^2 \quad (34b)\\ \zeta_{l',k',n,b}^{\{b\}(i)} \geq \sum_{k\in\mathcal{U}_{\hat{b}}} & \sum_{l=1}^{L}|\mathbf{w}_{l',k',n}^{\mathrm{H}}\mathbf{H}_{b,k',n}\mathbf{m}_{l,k,n}|^2, \forall k' \in \bar{\mathcal{U}}_b \quad (34c)\\ & (17) \text{ and } (6b) \quad (34d)\end{aligned}$$

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 $(i-1)^{\mathrm{th}}$ information exchange of the local interference vector $\boldsymbol{\zeta}^{\{b\}(i-1)}, \forall b \in \mathcal{B}$.

Now, the local problem (34) 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 $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)}$ depends on the local interference value $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b_k\}(i)}$ assumed by the BS b_k and the actual interference caused by the BS b as in $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i)}$ as

$$\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)} = \frac{1}{2}\left(\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i)} + \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b_k\}(i)}\right) \quad (35)$$

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

$$\nu_{l,k,n,b}^{\{b_k\}(i)} = \nu_{l,k,n,b}^{\{b_k\}(i-1)} + \rho\left(\zeta_{l,k,n,b}^{\{b_k\}(i)} - \zeta_{l,k,n,b}^{(i)}\right) \quad (36)$$

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

Algorithm 2: Distributed JSFRA scheme using ADMM

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n}\, \forall\, l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0$ and the transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying total power constraint (6b)
update $\mathbf{w}_{l,k,n}$ with (21) and $\tilde{\mathbf{u}}_{l,k,n}$ with (17)
initialize the global interference vectors $\boldsymbol{\zeta}^{(0)} = \mathbf{0}^{\mathrm{T}}$
initialize the interference threshold $\boldsymbol{\nu}^{\{b\}(0)} \forall b \in \mathcal{B} = 0$
foreach *BS* $b \in \mathcal{B}$ **do**
  repeat
    initialize $j = 0$
    repeat
      solve for $\mathbf{M}_{k,n}$ and the local interference $\boldsymbol{\zeta}^{\{b\}}$ using (34)
      exchange $\boldsymbol{\zeta}^{\{b\}(j)}$ among BSs in $\mathcal{B}$
      update dual variables in $\boldsymbol{\nu}^{\{b\}(j+1)}$ using (36)
      update the consensus vector $\boldsymbol{\zeta}^{(j+1)}$ using (35)
      $j = j + 1$
    until *convergence or* $j \geq J_{\max}$
    downlink precoded pilot transmission with $\mathbf{M}_{k,n}$
    update $\mathbf{W}_{k,n}$ and notify to all BSs in $\mathcal{B}$ using uplink precoded pilots as in [27]
    update $\tilde{\boldsymbol{u}}_{l,k,n}$ using (15c) and (16) for SCA or $\tilde{\epsilon}_{l,k,n}$ using (25c) for MSE approach
    $i = i + 1$
  until *convergence or* $i \geq I_{\max}$
end

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

B. Decomposition 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 [6], [7] by directly solving the KKT conditions. Similar work has been considered for the WSRM problem with the minimum rate constraints in [9], [10]. 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 (27) can be written as

$$\underset{\substack{t_{l,k,n},\mathbf{M}_{k,n},\\ \epsilon_{l,k,n},\mathbf{W}_{k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{U}_b} a_k \left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^q \tag{37a}$$

subject to

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

$$\sigma_{l,k,n} : \log_2(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\tilde{\epsilon}_{l,k,n}} \leq -t_{l,k,n} \tag{37c}$$

$$\delta_b : \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max},\ \forall\, b \tag{37d}$$

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (37b), (37c) and (37d).

The problem in (37) is solved using theKKT expressions, which is obtained by taking the derivative of the Lagrangian function w.r.t the primal and the dual variables as shown in the Appendix A. Upon solving, we obtain the iterative solution as

$$\mathbf{m}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\alpha_{y,x,n}^{(i-1)}\mathbf{H}_{b_k,x,n}^{\mathrm{H}}\mathbf{w}_{y,x,n}^{(i-1)}\mathbf{w}_{y,x,n}^{\mathrm{H}\,(i-1)}\mathbf{H}_{b_k,x,n} + \delta_b\mathbf{I}_{N_T}\Big)^{-1}\alpha_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)} \tag{38a}$$

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

$$t_{l,k,n}^{(i)} = -\log_2(\epsilon_{l,k,n}^{(i-1)}) - \frac{\left(\epsilon_{l,k,n}^{(i)} - \epsilon_{l,k,n}^{(i-1)}\right)}{\log(2)\,\epsilon_{l,k,n}^{(i-1)}} \tag{38c}$$

$$\sigma_{l,k,n}^{(i)} = \left[\frac{a_k\, q}{\log(2)}\left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}^{(i)}\right)^{(q-1)}\right]^+ \tag{38d}$$

$$\alpha_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i-1)} + \rho\left(\frac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i)}} - \alpha_{l,k,n}^{(i-1)}\right) \tag{38e}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L}\mathbf{H}_{b_x,k,n}\mathbf{m}_{y,x,n}^{(i)}\mathbf{m}_{y,x,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_x,k,n}^{\mathrm{H}} + \mathbf{I}_{N_R}\Big)^{-1}\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}^{(i)} \tag{38f}$$

Since the dual variables $\alpha^{(i)}$ and $\sigma^{(i)}$ are interdependent

in (39a), 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 (38e) to obtain the other variables in (38). At each iteration, the dual variable $\alpha^{(i)}$ is updated linearly from the earlier value $\alpha^{(i-1)}$ by a step size of $\rho \in [0, 1]$. 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 (38a). 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 (38) 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 (38) are evaluated sequentially until convergence or to a certain accuracy. In (38), all expressions are in closed form except the transmit precoders (38a), which depends on the BS specific dual variable δ_b. It can be solved efficiently by the bisection method satisfying the power constraint (37d). 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 [27], 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 (38a) 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. 3.

Convergence: The iterative method presented in Algorithm 3 converges to the stationary point if the dual variables $\alpha_{l,k,n}$ are allowed to converge or for fixed number of iterations $J_{\max}$. 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 $\epsilon_{l,k,n}$ [12]. 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 (38a).

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.

Algorithm 3: KKT approach for the JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B}$, $\forall\, k \in \mathcal{U}$, $\forall\, n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n}\, \forall\, l \in \{1, 2, \ldots, L\}$
Initialize: $i = 1$, $\mathbf{w}_{l,k,n}^{(0)}, \epsilon_{l,k,n}^{(0)}$ randomly, dual variables $\alpha_{l,k,n}^{(0)} = 1$, and $I_{\max}$
foreach *BS* $b \in \mathcal{B}$ **do**
  initialize $i = 0$
  repeat
    find $\mathbf{M}_{k,n}^{(i)}$ using (38a), where δ_b is obtained by bisection search satisfying (37d)
    update $\epsilon_{l,k,n}^{(i)}$, $t_{l,k,n}^{(i)}$, $\sigma_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ using (38b) and (38c), (38d) and (38e) at BS b and perform precoded downlink pilot transmission with $\mathbf{M}_{k,n}^{(i)}$
    update $\mathbf{W}_{k,n}^{(i+1)}$ using (38f) and notify to all BSs using uplink precoded pilot with $\tilde{\mathbf{w}}_{l,k,n}^{(i+1)}$ as the precoder
    $i = i + 1$
  until *until convergence or* $i \geq I_{\max}$
end

A. Centralized Solutions

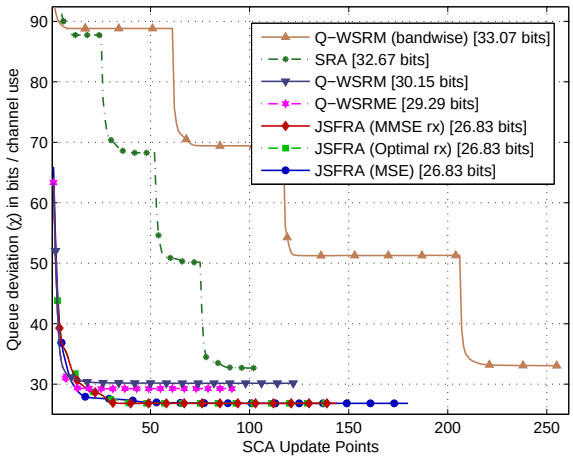
We discuss the performance of the centralized algorithms from Section III for some system configurations. To begin with, we consider a single cell single-input single-output (SISO) model operating at 10 dB signal-to-noise ratio (SNR) with $K = 3$ users sharing $N = 3$ sub-channel resources. The number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

Table. I outlines the channel seen by the users over each sub-channel followed by the rates assigned by three different algorithms, Q-WSRME allocation, JSFRA approach and the band-wise Q-WSRM scheme using WMMSE design [7]. The performance metric used for the comparison is the total number of backlogged bits left over at each slot after the allocation, which is denoted by $\chi = \sum_{k=1}^{K} [Q_k - t_k]^+$. Even though $\mathcal{U}(1)$ and $\mathcal{U}(3)$ has equal number of backlogged packets of $Q_1 = Q_3 = 4$ bits, user $\mathcal{U}(3)$ got scheduled in the first sub-channel due to the better channel condition. In contrast, the JSFRA approach assigns the first user on the first sub-channel, which reduces the total number of backlogged packets waiting at the transmitter. The rate allocated for $\mathcal{U}(2)$ on the second sub-channel is higher in JSFRA scheme compared to the other schemes. It is due to the efficient allocation of the total power shared across the sub-channels.

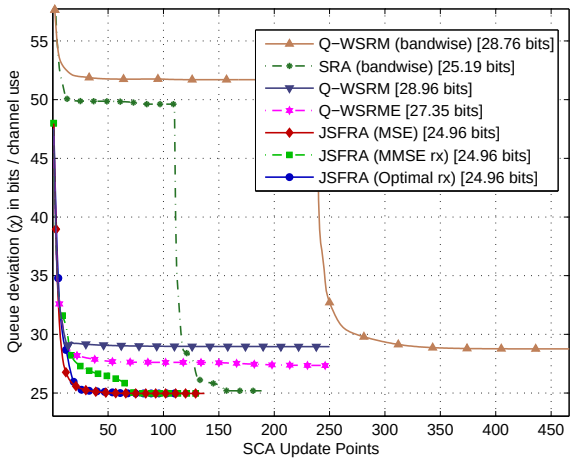
For the MIMO framework, we consider a system with $N = 3$ sub-channels and $N_B = 3$ BSs, each equipped with $N_T = 4$ transmit antennas operating at 10dB SNR, serving $|\mathcal{U}_b| = 3$ users each. The path loss between the BSs and the users are uniformly generated from $[0, -3]$ dB and the association is made by selecting the BS with lowest path loss component. Fig. 1a 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

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

TABLE I: Sub channel wise allocation for a scheduling instant



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



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

Fig. 1: Number of backlogged packets at the SCA update points

to minimize the total number of backlogged packets. The total number of queued packets for Fig. 1a is given by $Q_k = [14, 15, 14, 8, 12, 9, 12, 11, 11]$ bits and for Fig. 1b is $Q_k = [9, 12, 8, 12, 5, 4, 10, 8, 5]$ bits respectively.

The performance of the centralized algorithms are compared by the total number of residual bits remaining in the system after each SCA update in Fig. 1. The Q-WSRM algorithm is not optimal due to the problem of over allocation when the number of queued packets are small, as in Fig. 1. In contrast, the Q-WSRME algorithm provides more favorable allocation by including the explicit rate constraint to avoid the over allocation problem. It can be seen that the JSFRA algorithms converge to the optimal point for all formulations proposed in Section III-B. It can be seen that all the algorithms are Pareto-optimal and provides different performance based on the weights used to find a point in the rate region.

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

The behavior of the JSFRA algorithm for different exponents q are outlined in the Table. II for the users located

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

TABLE II: Queues for $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$

at the cell-edge of the system employing $N_T = 4$ transmit antennas. The configuration is mentioned in the caption of Table. II 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 compared using the total number of backlogged packets after each SCA update points. Fig. 2 compares the performance of the algorithms for the system configuration $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ with $N_T = 4$ transmit antennas at the BSs. Each BS serves $|\mathcal{U}_b| = 4$ users in a coordinated manner to reduce the total number of backlogged packets at each BS. The total number of queued packets assumed for both figures is $Q_k = [5, 7, 9, 11, 8, 12, 5, 4]$ bits. As pointed out in Section IV, the performance and the convergence speed of the distributed algorithms are susceptible to the step size used in

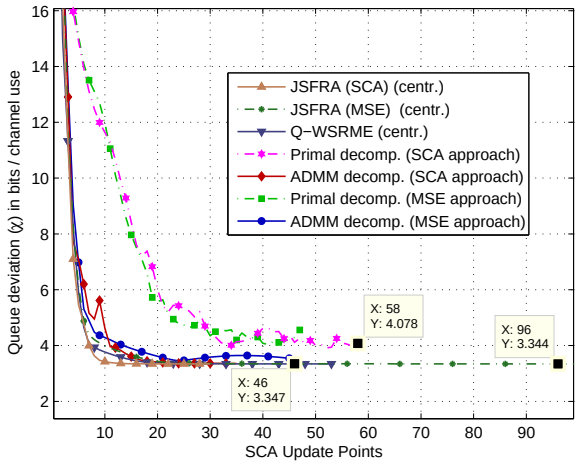


Fig. 2: χ at SCA points for $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$

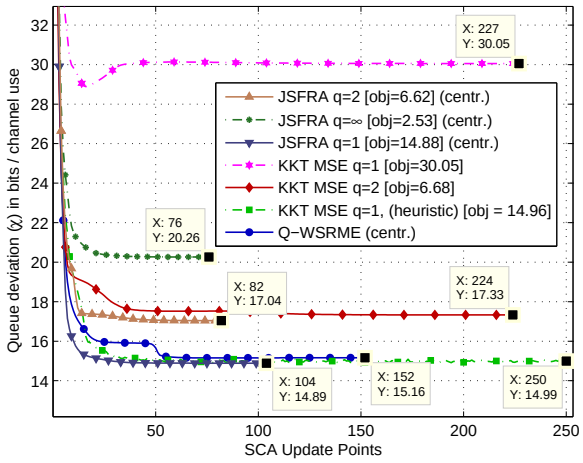


Fig. 3: χ at SCA points for $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$

the subgradient update. Due to the fixed interference levels in the primal approach, it may lead to infeasible solutions if the initial or any intermediate update is not feasible.

The Fig. 2 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 scheme is performed for $J_{\max} = 20$ iterations to exchange the respective coupling variables. In Fig. 2, the total number of backlogged packets at each SCA points are plotted without the inner loop iterations of $J_{\max}$ times for the primal or the dual variables convergence. It can be seen from Fig. 2 that the distributed algorithms approaches the centralized performance by exchanging minimal information between the coordinating BSs.

Fig. 3 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. 3, the plots are compared by the surplus number of backlogged packets at the end of each iteration or the SCA update point. Fig. 3 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 (38d) to 0 when the queue deviation is negative $Q_k - t_k < 0$, if not, then it will be the same as in (38d). 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 (38e).

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 once at each iteration, *i.e*, $J_{\max} = 1$, which is well justified for the practical implementations due to the signaling overheads. Fig. 3 also compares the effect of dropping the squared rate variable from the objective in the Q-WSRME scheme compared to the ℓ_2 norm which includes it. By dropping 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 paper, we addressed the allocation of space-frequency resources to the users in a multi-cell multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) system. The resource allocation is considered as a joint space-frequency precoder design problem since the allocation of a resource to a user can be achieved by a non-zero precoding vector. We proposed the joint space-frequency resource allocation (JSFRA) scheme by adopting the successive convex approximation (SCA) technique to model the non-convex constraint as a sequence of convex subsets to design the precoders for minimizing the number of queued packets. Additionally, an alternative approach using mean squared error (MSE) relaxation is also proposed with the same objective by fixing the receivers based on MSE minimization. We also proposed the distributed solutions for the centralized JSFRA problem using alternating directions method of multipliers (ADMM) and primal decomposition methods. Finally, we proposed an iterative algorithm to determine the precoders in a decentralized manner based on the Karush-Kuhn-Tucker (KKT) conditions for the MSE reformulated JSFRA scheme. Numerical results shows that the proposed algorithms perform better than the existing approaches.

APPENDIX A
KKT CONDITIONS FOR MSE APPROACH

In order to solve for an iterative precoder design algorithm, the KKT expressions for the problem in (37) are obtained by differentiating the Lagrangian by assuming the equality constraint for (37b) and (37c). At the stationary point, the following conditions are to be satisfied.

$$\nabla_{\epsilon_{l,k,n}} : -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \tag{39a}$$

$$\nabla_{t_{l,k,n}} : -q\, a_k \Big(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\Big)^{(q-1)} + \frac{\sigma_{l,k,n}}{\log_2(e)} = 0 \tag{39b}$$

$$\nabla_{\mathbf{m}_{l,k,n}} : \sum_{y\in\mathcal{U}}\sum_{x=1}^{L} \alpha_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_k,y,n}\mathbf{w}_{x,y,n}\mathbf{w}^{\mathrm{H}}_{x,y,n}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n} + \delta_b\mathbf{m}_{l,k,n} = \alpha_{l,k,n}\mathbf{H}^{\mathrm{H}}_{b_k,k,n}\mathbf{w}_{l,k,n}, \tag{39c}$$

$$\nabla_{\mathbf{w}_{l,k,n}} : \sum_{(x,y)\neq(l,k)} \mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}^{\mathrm{H}}_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_y,k,n}\mathbf{w}_{l,k,n} + \mathbf{I}_{N_R}\mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n} \tag{39d}$$

In addition to the primal constraints given in (37b), (37c) and (37d), the complementary slackness criterion must also be satisfied at the stationary point. Upon solving the above expression in (39) with the complementary slackness conditions, we obtain the iterative algorithm to determine the transmit and the receive beamformers as shown in (38).

REFERENCES

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

[2] 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.

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

[4] 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.

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

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

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

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

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

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

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

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

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

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

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

[16] M. Neely, *Stochastic network optimization with application to communication and queueing systems*, ser. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2010, vol. 3, no. 1.

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

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

[19] 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.

[20] 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.

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

[22] 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.

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

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

[25] G. Scutari, F. Facchinei, P. Song, D. Palomar, and J.-S. Pang, "Decomposition by Partial Linearization: Parallel Optimization of Multi-Agent Systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 641–656, Feb 2014.

[26] 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.

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

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