Abstract

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 wideband 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 a MU-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 linear 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 most, if not all, linear 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. Additional rate constraints based on the quality of service (QoS) requirements are included in the WSRM problem are solved via MSE reformulation in [7].

The problem of precoder design for the MIMO BC system are solved either by using a centralized controller or by using decentralized algorithms where each BS handles their own problems and exchange limited information via backhaul. The distributed approaches are based on the primal or dual decomposition [8], [9]. In 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 [10]. The dual approach controls the distributed subproblems by

fixing the *'interference price'* for each BS as detailed in [11].

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 data rate, the corresponding problem guarantees fairness among 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 [12], [13]. In the backpreassure algorithm, the differential queues between the source and the destination nodes are used as the weights scaling the transmission rate [14].

Earlier studies on the queue minimization problem was summarized in the survey paper [15]. In particular, the problem of power allocation to minimize the number of backlogged packets was considered in [16] by geometric programming formulations. Since the problem considered in [16] 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 backpreassure algorithm mentioned above was extended in [17] 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 alternating directions method of multipliers (ADMM) scheme to identify the precoders independently across the BSs by exchanging limited information via backhaul. 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 $\mathcal{C} = \{1, 2, \ldots, N\}$ 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

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

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. Assuming independent detection of data streams, we can write the signal-to-interference-plus-noise ratio (SINR) as

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

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

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

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 $\gamma_{l,k,n}$.[1] 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 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 $\sum_{k\in\mathcal{U}} |v_k|^q$. 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 $\tilde{v}_k \triangleq a_k^{1/q} v_k$, a_k is the weighting factor which is incorporated to control user priority based on their respective QoS, $\gamma_{l,k,n}$ is defined in (2), $\mathbf{M}_{k,n} \triangleq [\,\mathbf{m}_{1,k,n}\,\mathbf{m}_{2,k,n}\ldots\mathbf{m}_{L,k,n}\,]$ 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. It can be easily extended for user specific streams L_k instead of using the common L streams for all users. In (6b), we consider the sum power constraint for each BS across all sub-channels. Before discussing the proposed solutions, we consider the existing algorithm to

[1]This can be achieved by Gaussian signaling

solve the issue of minimizing the number of backlogged packets with additional constraints required by problem.

For practical and tractability reasons, we 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 (5) 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 [15], [16] 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 (13) 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 [13].
- 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.

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.[2] One such algorithm is the *backpreassure algorithm*, discussed in detail in [12]–[14]. 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 *backpreassure 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 with the objective of minimizing the number of queued packets waiting at the BSs. The weights used in the precoder design algorithm should be a function of the number of backlogged packets corresponding to each user. In order to find the weights, we adopt the Lyapunov drift minimization as discussed in [13], where the Lyapunov function is given by the squared sum of the number backlogged packets as

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

where $\mathbf{Q}(i)$ denotes the stacked user queues at the i^{th} slot. The Lyapunov function provides a measure of congestion in the system, as discussed in [13, Ch. 3]. Now the Lyapunov function drift is given by

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

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

[2]routers or user terminals

where the inequality is due to the upper bound

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

In order to minimize the number of backlogged packets at each instant, minimization of the Lyapunov drift (9) is carried over all possible control decisions 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

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

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

where the second term on the r.h.s is due to i.i.d assumption on the arrivals. We can bound the first term on the r.h.s by a constant B for all possible control actions taken for a given channel condition [13].

Now the expression in (11) looks similar to the WSRM problem if the weights are replaced by the number of backlogged packets corresponding to the users. The above discussed approach is extended for the wireless networks in [17], where the queue weighted sum rate maximization is considered as the objective function to determine the transmit precoders. 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). With this, the Q-WSRM problem for a wireless system is given by

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

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

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n})\leq Q_k,\ \forall\ k\in\mathcal{U} \tag{12c}$$

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

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

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

The problem defined in (13) provides a greedy queue minimizing approach as compared to (12), since the resource allocation is driven by the channel conditions in comparison with the number of queued packets as in (12). Note that in both formulations, the resources allocated to the users are limited by the backlogged packets with an explicit rate constraint defined by (12c).

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. 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 (12c) 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 achive an efficient allocation by identifying the optimal space-frequency resource for each user in the system. The queue deviation objective provides an alternate and a more efficient way to perform the resource allocation as compared to the Q-WSRM approach. In this approach, we present an iterative algorithm to solve (6) locally based on the idea of alternating optimization and successive convex approximation. Using (2), we can reformulate

the problem defined in (6) as

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

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

$$\beta_{l,k,n} \geq N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{15c}$$

$$\text{(6b)} \tag{15d}$$

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. In this formulation, we relaxed the equality constraint in (2) by the inequalities in (15b) and (15c). However, this step is 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 [2]. 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 $\beta_{l,k,n}$. The problem in (15) is known to be NP-hard even for the single antenna case [6], [7]. 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 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) [18].

By fixing the receivers, the problem now is to find optimal transmit beamformers for a given set of linear receivers which is 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 $\beta_{l,k,n}$) [19]. 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

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

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

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 [20], 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

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

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

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

$$\text{subject to} \quad \beta_{l,k,n} \geq N_0\|\mathbf{w}_{l,k,n}\|^2 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{18b}$$

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

$$\text{and (17)} \tag{18d}$$

Now, the optimal linear receiver for the fixed transmit precoders $\mathbf{M}_{i,n}\,\forall i \in \mathcal{U},\, \forall n \in \mathcal{C}$ is obtained by minimizing (6) w.r.t $\mathbf{w}_{l,k,n}$, which is given by

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

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

When the number of queued packets for each user is significantly large, the queue minimizing objective is in fact the sum rate maximization problem. Since the optimal receivers for the sum rate maximization problem with the fixed transmit precoders is the MMSE receivers, we can use the MMSE receiver as the receive beamformers without affecting the optimal solution. When the number of queued packets are

smaller than the achievable transmission rate of the users, the receiver defined by (19) provides better solution compared to the MMSE based receivers given by

$$\mathbf{w}_{l,k,n} = \mathbf{R}_{l,k,n}^{-1}\, \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n} \tag{20a}$$

$$\mathbf{R}_{l,k,n} = \sum_{i\in\mathcal{U}} \sum_{j=1}^{L} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n} \mathbf{m}_{j,i,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\, \mathbf{I}_{N_R} \tag{20b}$$

For the single receive antenna scenario, the receivers in (19) and (20) achieves the same optimal point, since the SINR expression is independent of the receive beamformer.

The proposed algorithm is referred as queue minimizing (QM) 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.

Algorithm 1: Algorithm of JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{C}$
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 (20) and $\tilde{\mathbf{u}}_{l,k,n}$ with (17) using $\tilde{\mathbf{m}}_{l,k,n}$.
repeat
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (18)
    update the constraint set (17) with $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ using (16) 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 (19) or (20) with the recent precoders $\mathbf{m}_{l,k,n}$.
  $j = j + 1$.
until *Queue convergence or* $j \geq J_{\max}$

Convergence: In the proposed solution, we replaced (15b) by a convex constraint using the first order approximation, which basically means that we do not solve the problem exactly. According to the

traditional block coordinate descent method (BCDM), we need to solve a sub problem when fixing a set of variables to the global optimum to ensure the convergence to a stationary point. If we just approximate the objective, then the convergence is guaranteed [21]. In our case, we solve the sub problem inexactly, so the convergence proof of BCDM does not apply to our problem. In this problem, we used alternating optimization with SCA method, which provides monotonic convergence since the objective is improved at each step *i.e* $f^{(i)} \geq f^{(i+1)}$, assuming $f^{(i)}$ is the objective function at the i^{th} SCA iteration. For a fixed receive beamformers, the problem is guaranteed to converge to a KKT point at each SCA points as discussed in [22]. Now, by alternating the transmit and the receive precoders, we perform BCDM, which provides monotonic convergence of the objective. Since the SCA method is also considered in our problem, the convergence is not straight forward. We can only speculate the convergence based on the monotonicity of the objective values. Since receive beamformers are based on the MMSE objective, it improves the objective value for each update, providing a monotonic increase in the objective value.

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

$$\epsilon_{l,k,n} = \mathbb{E}\big[(d_{l,k,n} - \hat{d}_{l,k,n})^2\big] \tag{21a}$$

$$= \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 + N_0 \left\|\mathbf{w}_{l,k,n}\right\|^2 \tag{21b}$$

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 (21) with the MMSE receiver shown in (20), we obtain the following relation between the MSE and the SINR as

$$\epsilon_{l,k,n} = \frac{1}{1+\gamma_{l,k,n}} \tag{22}$$

where $\gamma_{l,k,n}$ is the received SINR as in (2). Using the equivalence in (22), the WSRM objective can be reformulated as the weighted minimum mean squared error (WMMSE) objective to obtain the precoders for the MU-MIMO scenario as discussed in [5], [6].

Let $v'_k = Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}$ denotes 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, (15) is written as

$$\underset{\substack{t_{l,k,n},\mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \tag{23a}$$

$$\text{subject to} \quad t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n}) \tag{23b}$$

$$\epsilon_{l,k,n} \geq \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 + N_0 \left\|\mathbf{w}_{l,k,n}\right\|^2 \tag{23c}$$

$$\text{and (6b)} \tag{23d}$$

where (23c) bounds the MSE by $\epsilon_{l,k,n}$ and (23b) relaxes the transmitted rate $t_{l,k,n}$ using the the MSE relation.

The queue minimizing JSFRA problem using alternative MSE formulation given by (23) is non-convex due to the constraint (23b). 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 (23b). The first order Taylor approximation around a fixed MSE value $\tilde{\epsilon}_{l,k,n}$ for (23b) is given by

$$-\log_2(\tilde{\epsilon}_{l,k,n}) - \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\, \tilde{\epsilon}_{l,k,n}} \geq t_{l,k,n} \tag{24}$$

Now, using the above approximation for the rate constraint, the optimization problem is solved for the optimal precoders $\mathbf{m}_{l,k,n}$, MSEs $\epsilon_{l,k,n}$ and the users rates over each sub-channel $t_{l,n,k}$. Once the optimal values are available, the local MSE value $\tilde{\epsilon}_{l,k,n}$ is now updated with the new value $\epsilon_{l,k,n}$. The optimization problem for a fixed receive beamformers $\mathbf{w}_{l,k,n}$ is given as

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

$$\text{subject to} \quad \text{(6b), (24)} \tag{25b}$$

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

Convergence: The convergence of the MSE reformulation approach follows the same arguments detailed in [5], [6]. 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 (25). The MSE relaxation method converges to the same stationary KKT point as of the original WSRM problem by making the graidents 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 reamining number of queued bits in the formulation. It provides a sub-optimal solution to the optimal way of distributing the sub-channel wise precoder design via primal/dual decomposition methods discussed in [8], [9].

The proposed queue minimizing (QM) spatial resource allocation (SRA) scheme decouples the problem by fixing the power across 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 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

$$Q_{k,n} = \max\left\{Q_k - \sum_{j=1}^{n-1}\sum_{l=1}^{L} t_{l,k,j}, 0\right\},\ \forall\ k \in \mathcal{U} \tag{26}$$

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 (26), 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 very sensitive to the order in which the sub-channels are selected for the optimization problem. The algorithmic representation of the queue minimizing (QM) SRA scheme is shown in Algorithm 2

IV. Distributed Solutions

This section addresses the distributed precoder designs for the proposed JSFRA scheme. The formulation in (18) or (25) requires a centralized controller to perform the precoder design for all users belonging

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 $\mathcal{C} \rightarrow \tilde{\mathcal{C}}$
for $n \leftarrow 1$ **to** N **do**
  update $Q_{k,n}$ using (26) and let $\hat{n} = \tilde{\mathcal{C}}(n)$.
  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 (20) and $\tilde{\mathbf{u}}_{l,k,\hat{n}}$ with (17) using $\tilde{\mathbf{m}}_{l,k,\hat{n}}$.
  repeat
    repeat
      solve for the transmit precoders $\mathbf{m}_{l,k,\hat{n}}$ using (18)
      update the constraint set (17) with $\tilde{p}_{l,k,\hat{n}}$, $\tilde{q}_{l,k,\hat{n}}$ and $\tilde{\beta}_{l,k,\hat{n}}$ using (16) 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 (20) with the recent precoders $\mathbf{m}_{l,k,n}$.
    $j = j + 1$.
  until *Queue convergence or* $j \geq J_{\max}$
end

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 (PD) and the ADMM based dual decomposition (DD) 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 $\mathcal{U}\backslash\mathcal{U}_b$. The centralized problem[3] can be equivalently written as

$$\underset{\substack{\gamma_{l,k,n},p_{l,k,n},q_{l,k,n}\\ \mathbf{M}_{k,n},\mathbf{W}_{k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{27a}$$

$$\text{subject to} \quad \text{(18b)} - \text{(18d)}, \tag{27b}$$

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 [10], [11], the coupling constraint[4] given by (18b) can be expressed

[3](18) for the SCA approach and (25) for MSE reformulation approach

[4]coupling constraint in MSE scheme is given by (25c)

by grouping the interference contribution from each BSs in the system as

$$\beta_{l,k,n} \geq \sum_{\substack{j=1 \\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2 + \sum_{i \in \mathcal{U}_{b_k} \setminus \{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \sum_{b \in \bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} \; + \; N_0 \|\mathbf{w}_{l,k,n}\|^2 \tag{28}$$

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

$$\zeta_{l,k,n,b} \geq \sum_{i \in \mathcal{U}_b} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2, \forall b \in \bar{\mathcal{B}}_{b_k} \tag{29}$$

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 optimal coupling variables $\zeta_{l,k,n,b}$ by either using PD or by using DD method. In both approaches, the coupling constraint (18b) for the SCA scheme and (25c) for the MSE relaxation scheme 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 (28), 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 WSRM problem in [10] 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 (DD), but shows better convergence properties. In contrast to the primal decomposition problem, the alternating directions method of multipliers (ADMM) method relaxes the interference constraints by including it in the objective function of each subproblem with a penalty pricing [8], [9]. In order to decouple the problem (27), the coupling variables $\zeta_{l,k,n,b}$ in (28) 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

$$\boldsymbol{\zeta} = \Big[\zeta_{1,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(1),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_1(|\bar{\mathcal{U}}_1|),1,1}, \ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),1,N_B}, \ldots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),N,N_B}\Big] \tag{30a}$$

$$n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| = NL \sum_{b \in \mathcal{B}} \big|\bar{\mathcal{U}}_b\big| \tag{30b}$$

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 [9], [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

$$\underset{\substack{\gamma_{l,k,n},\mathbf{W}_{k,n},\mathbf{M}_{k,n}\\ \beta_{l,k,n},\boldsymbol{\zeta}^{\{b\}(i)}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}^{\{b\}(i-1)T}\left(\boldsymbol{\zeta}^{\{b\}(i)} - \boldsymbol{\zeta}^{(i-1)}(b)\right) + \frac{\rho}{2}\Big\| \underbrace{\boldsymbol{\zeta}^{(i-1)}(b)}_{\text{consensus}} - \underbrace{\boldsymbol{\zeta}^{\{b\}(i)}}_{\text{locals}} \Big\|_2^2 \tag{31a}$$

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

$$+ \sum_{i\in\mathcal{U}_b\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{j,i,n}|^2 + N_0\|\mathbf{w}_{l,k,n}\|^2 \tag{31b}$$

$$\zeta_{l',k',n,b}^{\{b\}(i)} \geq \sum_{k\in\mathcal{U}_{\hat{b}}} \sum_{l=1}^{L} |\mathbf{w}_{l',k',n}^{\mathrm{H}}\mathbf{H}_{b,k',n}\mathbf{m}_{l,k,n}|^2, \ \forall k' \in \bar{\mathcal{U}}_b, \ \forall n \in \mathcal{C} \tag{31c}$$

$$(17) \text{ and } (38b) \tag{31d}$$

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 (31) 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

$$\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)} = \frac{1}{2}\left(\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i)} + \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b_k\}(i)}\right) \tag{32}$$

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

$$\nu_{l,k,n,b}^{\{b_k\}(i)} = \nu_{l,k,n,b}^{\{b_k\}(i-1)} + \rho\left(\zeta_{l,k,n,b}^{\{b_k\}(i)} - \zeta_{l,k,n,b}^{(i)}\right), \forall b, b_k \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b \tag{33}$$

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 [9] 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, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{C}$
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 (20) and $\tilde{\mathbf{u}}_{l,k,n}$ with (17) 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 (31)
    exchange $\boldsymbol{\zeta}^{\{b\}(j)}$ across the coordinating BSs in $\mathcal{B}$
    update the dual variables in $\boldsymbol{\nu}^{\{b\}(j+1)}$ using (33)
    update the global interference vector $\boldsymbol{\zeta}^{(j+1)}$ using (32)
    $j = j + 1$
  until *convergence or* $j \geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (20) with 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 (16) and (15c) for the SCA approach (or)
  update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (25c) 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 rate constraints in [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 in an iterative approach as in [25].

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 $Q_k \gg \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ to ignore the absolute operator.

With the assumption of either one of the above conditions to be true, the problem in (25) can be written as

$$\underset{\substack{t_{l,k,n},\mathbf{M}_{k,n},\\ \sigma_{l,k,n},\mathbf{W}_{k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{U}_b} a_k \left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^q \tag{34a}$$

subject to

$$\alpha_{l,k,n}: \quad \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 + N_0\left\|\mathbf{w}_{l,k,n}\right\|^2 \tag{35}$$

$$+\sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}\right|^2 \leq \epsilon_{l,k,n} \tag{36a}$$

$$\sigma_{l,k,n}: \quad -\log(\tilde{\epsilon}_{l,k,n}) - \frac{(\epsilon_{l,k,n}-\tilde{\epsilon}_{l,k,n})}{\tilde{\epsilon}_{l,k,n}} \geq t_{l,k,n}\,\log(2) \tag{36b}$$

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

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (36a), (36b) and (36c). The equality of (36a) is due to the equivalence of the MSE expression with the transmit and the receive precoders.

By forming the Lagrangian of (34) with the corresponding dual variables as shown in (34), we can obtain the KKT expressions by differentiating with respect to the variables present in the problem as de-

tailed 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

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

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

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

$$\sigma_{l,k,n}^{(i)} = \frac{a_k\, q}{\log(2)} \Big(Q_k - \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}^{(i)} \Big)^{(q-1)} \tag{37d}$$

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

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

The dual variable $\alpha^{(i)}$ is updated with memory as in (37e) to avoid abrupt oscillations due to $\sigma < 0$ from (37d) when $Q_k < \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}$. The parameter $\rho \in [0,1]$ provides a linear combination of the previous to the current value of the dual variable, dictating system reaction for the excess allocation.

The KKT solutions provided in (37) 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 (37) are evaluated sequentially until convergence or to a certain accuracy. In (37), all expressions are in closed form except the transmit precoders (37a), which depends on the BS specific dual variable δ_b. It can be solved efficiently by the bisection method satisfying the power constraint (36c). After each iteration instant, the transmit and the receive precoders are exchanged across the coordinating BSs in $\mathcal{B}$ to obtain the next operating point.

The receive beamformers from the users can be informed to the coordinating BSs by using the precoded uplink pilot signaling, where the precoders used for the uplink pilots are the receive beamformers $\sqrt{\alpha_{l,k,n}^{(i-1)}}\mathbf{w}_{l,k,n}^{*(i-1)}$ evaluated at the receivers. Upon receiving the uplink precoded pilots by the BS, the effective channel $\sqrt{\alpha_{l,k,n}^{(i-1)}}\mathbf{w}_{l,k,n}^{*\,(i-1)}\mathbf{H}_{b,k,n}^{\mathrm{T}}$ can be used in the expression (37a) to update the transmit

precoders at the respective BSs [23], $\mathbf{x}^*$ represents the conjugate of $\mathbf{x}$. The algorithmic representation of the KKT based scheme is shown in Algorithm. 4.

Algorithm 4: KKT approach for the JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{C}$
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 $\alpha_{l,k,n}^{(0)} = 1$
set the maximum iteration counter $I_{\max}, J_{\max}$ to a valid number.
repeat
  foreach *BS* $b \in \mathcal{B}$ **do repeat**
    update the transmit precoders $\mathbf{M}_{k,n}^{(i)}$ using (37a), where δ_b is identified by the bisection method satisfying (36c).
    update the MSE $\epsilon_{l,k,n}^{(i)}$ and the throughput $t_{l,k,n}^{(i)}$ using (37b) and (37c).
    solve for the dual variables $\alpha_{l,k,n}^{(i)}$ and $\sigma_{l,k,n}^{(i)}$ using (37e) and (37d).
    $i = i + 1$.
  until *until convergence or* $i \geq I_{\max}$
  $j = j + 1$.
  evaluate the receive beamforming vector $\mathbf{W}_{k,n}^{(i)}$ using (37f).
  exchange the receive precoders $\mathbf{W}_{k,n}^{(i)}$ and the dual variables $\alpha_{l,k,n}^{(i)}$ across the coordinating BSs in $\mathcal{B}$.
until *until convergence or* $j \geq J_{\max}$

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 $\epsilon_{l,k,n}$ [9]. 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 (37a).

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.

Users	Queued Packets	Channel Gains			Q-WSRM approach (*backpreassure* algorithm)			JSFRA Scheme			Q-WSRM band Alloc Scheme		
		Sch-1	Sch-2	Sch-3	Sch-1	Sch-2	Sch-3	Sch-1	Sch-2	Sch-3	Sch-1	Sch-2	Sch-3
1	5	1.71	0.53	0.56	0	0	0	4.91	0	0	0	0	0
2	8	0.39	1.41	1.03	0	4.46	3.54	0	4.36	0	0	4.39	3.53
3	6	2.34	1.26	2.32	5.72	0	0	0	0	5.79	5.81	0	0
Remaining backlogged packets (χ)					5.28 bits			3.91 bits			5.28 bits		

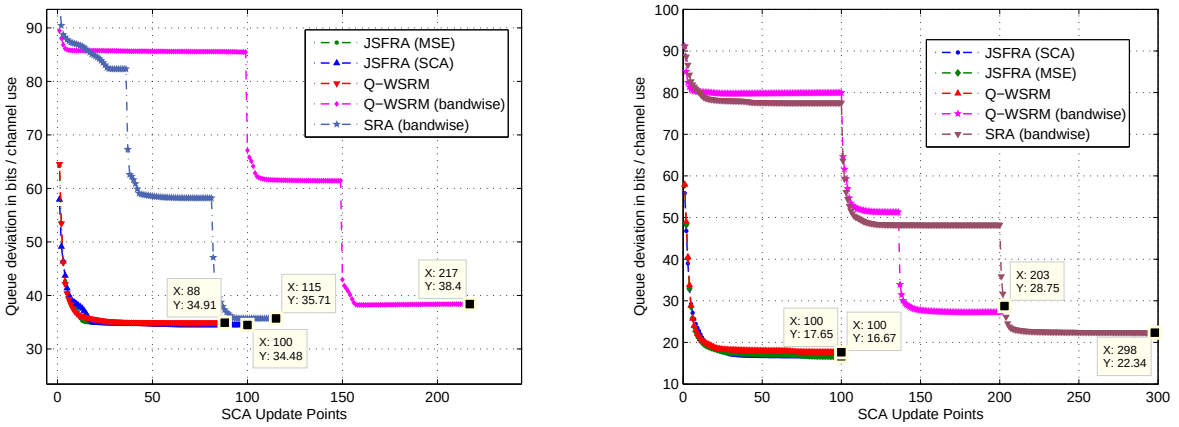
TABLE I: Sub channel wise allocation for a scheduling instant

A. Centralized Solutions

This section discusses the performance of the centralized algorithms discussed in 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. With this configuration, the resources allocated for the users on each sub-channel by various algorithms are listed in Table I.

Table I lists 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 Table I, 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. Int 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. 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-WSRM schemes to perform the optimal allocations to minimize the total number of backlogged packets.



(a) 4×1 system (b) 4×2 system

Fig. 1: Convergence plot for $\{N, N_B, K\} = \{3, 3, 9\}$ model

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

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 (26) before designing the precoders for the sub-channels. At each SCA points, the number 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. 1a for the discussed centralized schemes and the convergence point are marked with the data tips. Fig. 1b compares the performance of the centralized algorithms for the $N_R = 2$ receive antennas.

The behavior of the JSFRA algorithm for different exponents q are outlined in the Table II 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 II along with the number of queued bits for each user. It is evident that

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

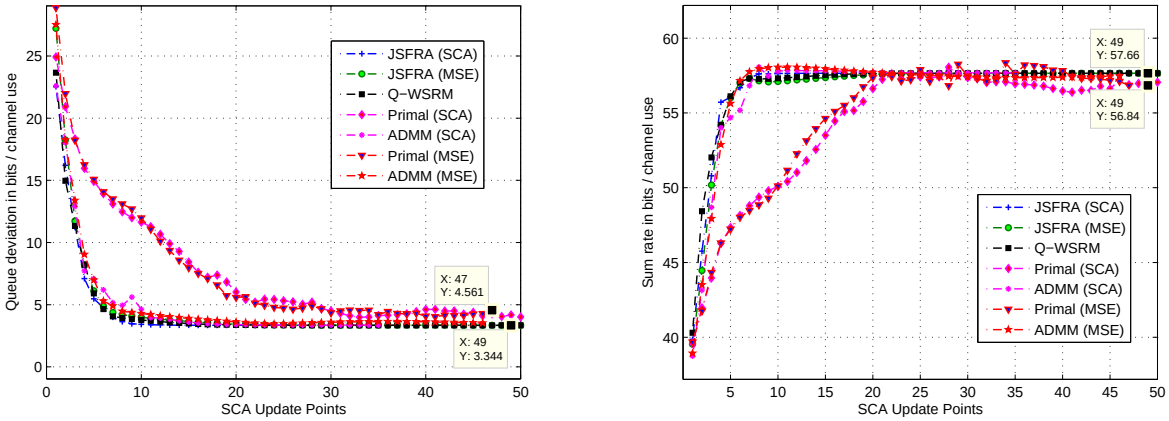
the algorithm minimizes the queued bits for the ℓ_1 norm compared to the ℓ_2 norm, which is shown in the column displaying the total number of left over packets χ in bits. The ℓ_∞ norm provides fair allocation of the resources by making the left over packets $\chi_k = 3.58$ bits. The ℓ_∞ norm provides 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. The total number of queued packets after the current allocation and the total sum rate achieved by the network are used as the performance measures. Fig. 2 shows the performance of the algorithms for the system configuration represented by $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ with $N_T = 4$ transmit antennas at each BS. Each BS serves $|\mathcal{U}_b| = 4$ users in a coordinated manner to reduce the total number of backlogged packets at each BS. Fig. 2a shows the total number of backlogged packets and Fig. 2b plots the total rate achieved by different algorithms after each SCA update. 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 (42) is not valid.

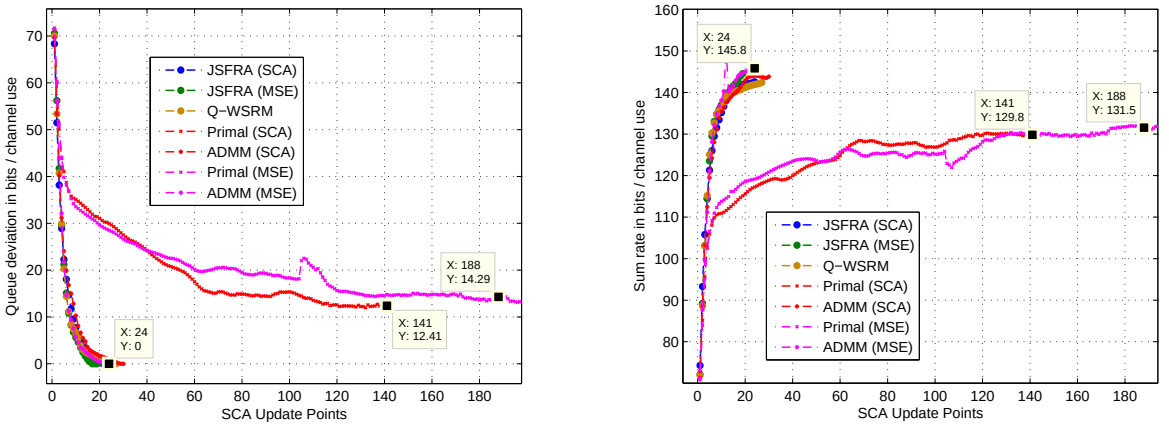
The Fig. 2 plots the performance of the primal and the ADMM solutions of the JSFRA scheme using SCA and by MSE relaxation at each SCA update points. In between the SCA updates, the primal or the ADMM schemes are performed for $J_{\max} = 20$ iterations by exchanging either the fixed interference variables in the primal approach or the dual variables in the ADMM scheme. It can be seen from Fig. 2 that the distributed algorithms approaches the centralized performance by exchanging minimal information between the coordinating BSs. The plot also includes the Q-WSRM which performs similar to the other centralized schemes discussed in this paper.

In Fig. 3, 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 with the maximum interference power from any neighboring BS follows $\mathbb{U}(0, -6)$ dB. The performance of the algorithms are similar to the $N_B = 2$ BS scenario discussed earlier.



(a) Queue deviation (b) Sum rate performance

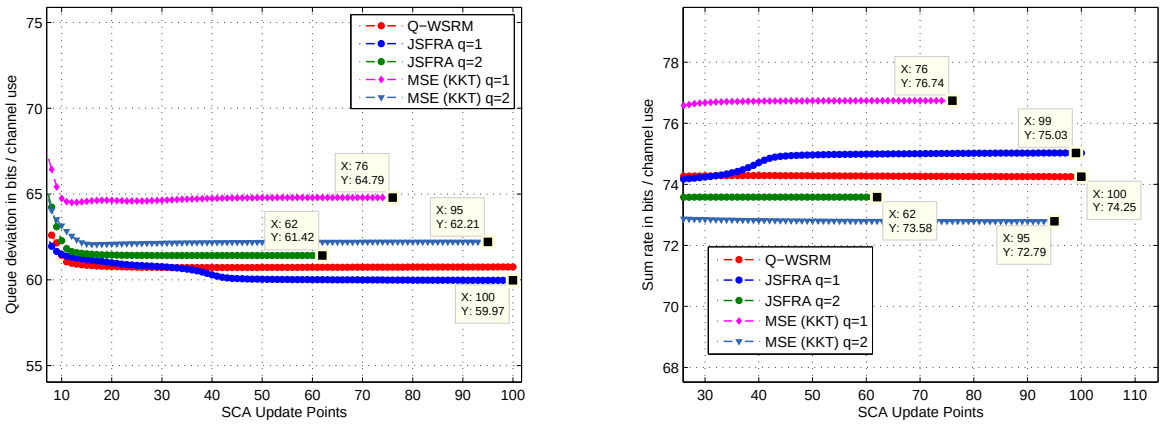
Fig. 2: Convergence plot for $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ model



(a) Queue deviation (b) Sum rate performance

Fig. 3: Convergence plot for $\{N, N_B, K, N_R\} = \{6, 3, 12, 2\}$ model

Fig. 3 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.



(a) Queue deviation

(b) Sum rate performance

Fig. 4: Convergence plot for $\{N, N_B, K, N_R\} = \{4, 2, 12, 1\}$ model

The performance of the MSE-KKT scheme discussed in Section IV-B is presented in Fig. 4. It compares the JSFRA scheme with ℓ_1 and ℓ_2 norm cases in addition to the Q-WSRM scheme. We compare the performance of the MSE-KKT scheme for the exponents $q = 1$ and $q = 2$. The performance of the ℓ_1 JSFRA scheme performs the best over other algorithms in minimizing the total number of queued bits after the current allocation due to its greedy approach. The inclusion of the additional second order rate term in (14), the ℓ_2 JSFRA scheme achieves different performance than the Q-WSRM scheme as shown in Fig. 4.

It can be seen from Fig. 4b, that the sum rate of the $q = 1$ KKT approach outperforms other schemes in spite of Fig. 4a showing inferior queue minimizing performance. This due to the lack of the rate bounding constraint (37d), which is simply given by $\sigma_{l,k,n}^{(i)} = \frac{a_k}{\log(2)}$ providing no control for the excess allocations. For $q = 2$, when the allocated rate exceeds the available queued bits, $\sigma_{l,k,n}^{(i)}$ will be negative, thereby reducing (37e) as $\alpha_{l,k,n}^{(i)} < \alpha_{l,k,n}^{(i-1)}$ to reduce the priority for the corresponding user in the next iteration. The performance of the closed form solution using the KKT solution performs similar to the centralized schemes with $q = 2$ by limiting the rates beyond the available queued packets.

ACKNOWLEDGMENT

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

VI. Conclusions

Appendix A

primal decomposition approach

By fixing the interference values ζ_{l',k',n,b_k} corresponding to the interference from the BS b_k, the constraint involving the coupling variables (18b) can be relaxed using the equivalent formulation in (28). 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

$$\underset{\substack{\gamma_{l,k,n},\mathbf{M}_{k,n}\\ \mathbf{W}_{k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_{b_k}\|_q \tag{38a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_{b_k}} \text{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\text{H}}\right) \leq P_{\max}, \tag{38b}$$

$$\beta_{l,k,n} \geq \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\text{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\text{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} \; + \; N_0\|\mathbf{w}_{l,k,n}\|^2 \tag{38c}$$

$$\zeta_{l',k',n,b_k} \geq \sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} |\mathbf{w}_{l',k',n}^{\text{H}}\mathbf{H}_{b_k,k',n}\mathbf{m}_{l,k,n}|^2, \; \forall k' \in \bar{\mathcal{U}}_{b_k}, \; \forall n \in \mathcal{C} \tag{38d}$$

$$\text{and (17)}, \tag{38e}$$

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

$$\boldsymbol{\zeta}_{k,n,b} = \left[\, \zeta_{1,k,n,b}, \ldots, \zeta_{L,k,n,b} \,\right] \tag{39a}$$

$$\boldsymbol{\zeta}_n^{\{b_k\}} = \Big[\, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(1)}, \ldots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \ldots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(|\mathcal{U}_{b_k}|),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \ldots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(1),n,b_k}, \ldots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(|\bar{\mathcal{U}}_{b_k}|),n,b_k} \,\Big] \tag{39b}$$

$$\boldsymbol{\zeta}^{\{b_k\}} = \Big[\, \boldsymbol{\zeta}_1^{(b_k)}, \ldots, \boldsymbol{\zeta}_N^{(b_k)} \,\Big], \tag{39c}$$

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

$$n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| = \left(|\bar{\mathcal{B}}_{b_k}||\mathcal{U}_{b_k}| + |\bar{\mathcal{U}}_{b_k}|\right) LN \tag{40}$$

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

The local subproblem (38) 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

$$\underset{\boldsymbol{\zeta}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \alpha_b^{\star}(\boldsymbol{\zeta}^{\{b\}}) \tag{41a}$$

$$\text{subject to} \quad \boldsymbol{\zeta}^{\{b\}} \in \mathbb{R}_{+}^{n_b}, \forall b \in \mathcal{B}, \tag{41b}$$

where $\alpha_b^{\star}(\boldsymbol{\zeta}^{\{b\}})$ denotes the optimal solution for (38) 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 $\boldsymbol{\zeta} = \left[\boldsymbol{\zeta}^{\{\mathcal{B}(0)\}}, \boldsymbol{\zeta}^{\{\mathcal{B}(1)\}}, \ldots, \boldsymbol{\zeta}^{\{\mathcal{B}(|\mathcal{B}|)\}}\right]$.

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

$$\zeta_{l,k,n,b}^{\{b_k\}(i)} = \left[\zeta_{l,k,n,b}^{\{b_k\}(i-1)} - \rho\, s_{l,k,n,b}^{\{b_k\}(i-1)}\right]^{+}, \forall b \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b, \tag{42}$$

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 (41) 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 (39) as discussed in [10]. Now the primal and the dual variables $\mu_{l,k,n}^{\{b_k\}}$ and $\mu_{l',k',n}^{\{b_k\}}$ corresponding to the constraints (38c) and (38d) 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, the 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 (42). The algorithmic representation of the PD approach is detailed in Algorithm. 5.

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 (34) are obtained by differentiating the Lagrangian by assuming the equality constraint for (36a) and

Algorithm 5: Decentralization via Primal Decomposition for JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall\, b \in \mathcal{B},\ \forall\, k \in \mathcal{U},\ \forall\, n \in \mathcal{C}$
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 (20) and $\tilde{\mathbf{u}}_{l,k,n}$ with (17) 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 (39)
    exchange $\mu_{l,k,n}^{\{b\}}$ across the coordinating BSs in $\mathcal{B}$
    update $\zeta_{l,k,n,b}^{\{b\}(j+1)}$ using (42) locally
    $j = j + 1$
  until *convergence or* $j \geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (20) with 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 (16) and (15c) for the SCA approach (or)
  update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (25c) with equality for the MSE formulation approach
  $i = i + 1$
until *convergence or* $i \geq I_{\max}$

(36b) as

$$\nabla_{t_{l,k,n}}: \quad -q\left[a_k\left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^{(q-1)}\right] + \sigma_{l,k,n}\log(2) = 0 \tag{43a}$$

$$\nabla_{\epsilon_{l,k,n}}: \quad -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \tag{43b}$$

$$\nabla_{\mathbf{m}_{l,k,n}}: \sum_{y\in\mathcal{U}}\sum_{x=1}^{L} \alpha_{x,y,n}\mathbf{H}_{b_k,y,n}^{\mathrm{H}}\mathbf{w}_{x,y,n}\mathbf{w}_{x,y,n}^{\mathrm{H}}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n} + \delta_b\mathbf{m}_{l,k,n} = \alpha_{l,k,n}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}, \tag{43c}$$

$$\nabla_{\mathbf{w}_{l,k,n}}: \sum_{(x,y)\neq(l,k)} \mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}_{x,y,n}^{\mathrm{H}}\mathbf{H}_{b_y,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n} + \mathbf{I}_{N_R}\mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n} \tag{43d}$$

in addition to the primal constraints given in (36a), (36b) and (36c), the complementary slackness criterions are given by

$$\underbrace{\alpha_{l,k,n}\Big(\epsilon_{l,k,n} - \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 - \sum_{(x,y)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}\right|^2 - N_0\left\|\mathbf{w}_{l,k,n}\right\|^2\Big)}_{=0} = 0 \tag{44a}$$

$$\underbrace{\sigma_{l,k,n}\Big(\log(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\tilde{\epsilon}_{l,k,n}} + t_{l,k,n}\,\log(2)\Big)}_{=0} = 0 \tag{44b}$$

$$\delta_b\Big(\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}\right) - P_{\max}\Big) = 0. \tag{44c}$$

In the expressions (44a) and (44b), the value inside the braces are zero due to the equality constraints. Now, the dual variables corresponding to the inequality constraint (36b) satisfies $\sigma_{l,k,n} \geq 0$. The total power constraint in (44c) 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 (44c). Upon solving the KKT expressions in (43) and (44), 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. 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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 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. Tölli, and M. Juntti, "Weighted sum rate maximization for interfering broadcast channel via successive convex approximation," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2012, pp. 3838–3843.