Abstract

I. Introduction

II. System Model

We consider a downlink multiple-input multiple-output (MIMO) broadcast channel (BC) scenario in an orthogonal frequency division multiplexing (OFDM) framework with N sub-channels and N_B base stations (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 the set $\mathcal{B}$ holds the coordinating BSs. The serving BS of user k is denoted by $b_k \in \mathcal{B}$. We denote by $\mathcal{C} = \{1, 2, \dots, N\}$ 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 $\mathbf{m}_{l,k,n} \in \mathbb{C}^{N_T \times 1}$ for transmission. 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 signal of the l^{th} spatial stream over sub-channel n^{th} at user k is given by

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

where $\mathbf{H}_{b,k,n} \in \mathbb{C}^{N_R \times N_T}$ with rank $L = \min(N_R, N_T)$ is the channel between the BS b and user k on the 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. Assuming $||\mathbf{w}_{l,k,n}||_2^2 = 1$ and independent detection of data streams, we can write the signal-to-interference-plus-noise ratio (SINR) as

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

With the infinite buffer model assumption, 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^{th} instant for the user k depend on the fresh arrivals at the i^{th} instant and the total number of backlogged packets $Q_k(i-1)$ as given by

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

where t_k denote the transmissions in bits and $[x]^+ = \max\{x, 0\}$. The total number of transmitted bits for the user k is given by $t_k(i) = \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}(i)$, where $t_{l,k,n}$ is the total number of bits that can be transmitted over l^{th} spatial stream and on n^{th} sub-channel at the i^{th} instant. Note that the units of t_k and Q_k are in bits defined per channel use.

III. Queue Minimizing Precoder Designs

In general, the precoder design for the MIMO OFDM problem is highly difficult due to the non-convex 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. In this section, we discuss the precoder design in a centralized way for all BSs in the set $\mathcal{B}$.

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

The weighted sum rate maximization (WSRM) technique is predominantly used to design the precoders for the MIMO BC channels to maximize the weighted sum throughput in the downlink direction. The weights are used to provide different priorities to the users based on their requirements and quality of service (QoS) criterion. Since the objective of this work is mainly concerned with the reduction of the number of queued packets, we use the length of the queued packets for each user as the respective weights. Using the length of the queued packets as the corresponding weights is obtained from the well known algorithm, called the *backpreassure algorithm*, which is used to decide the routing decisions for each node in the network to efficiently send the packets to the desired destination with minimal packet drops and delay. The backpreassure algorithm is the outcome of minimizing the Lyapunov drift conditioned over the current queue states $\mathbf{Q}(i-1)$, where $\mathbf{Q}$ is the vector formed by stacking the number of queued packets of each user in the system. Assuming the queue length grows according to (3) for all the users, the Lyapunov drift is given by [1], [2]

$$L(i) = \frac{1}{2}\left\|\mathbf{Q}(i)\right\|_2^2 = \frac{1}{2}\sum_{k \in \mathcal{U}} Q_k(i)^2. \tag{4}$$

The conditional Lyapunov drift is given by

$$\Delta(i) = \mathrm{E}\left[L(i) - L(i-1) | \mathbf{Q}(i-1)\right]. \tag{5}$$

Now, by substituting (3) in (5), the resulting expression can be upper bounded by

$$\Delta(i) \leq \sum_{k \in \mathcal{U}} \mathrm{E}\left[\mathbf{Q}(i-1)\left(\lambda(i) - \mathbf{t}(i)\right)|\mathbf{Q}(i-1)\right] + \mathrm{E}\left[\lambda^2(i-1)\right] + \mathrm{E}\left[\mathbf{t}^2(i-1)\right], \tag{6}$$

where λ and $\mathbf{t}$ are the stacked arrivals and transmissions of all users. Since the expectation is conditioned over $\mathbf{Q}(i-1)$ and the second order statistics of the arrivals and the transmissions are bounded [2], the drift minimization solution is given as

$$\underset{\mathbf{t}(i)}{\text{minimize}} \sum_{k \in \mathcal{U}} Q_k(i-1)\, t_k(i). \tag{7}$$

Now from (7), the optimization variables $\mathbf{t}(i)$, which is the vectorized transmission rates, are to be identified for each user for the current slot i. Since the current system can exploit both space and frequency resources through MIMO OFDM model, the transmission rates $\mathbf{t}(i)$ for the current instant can be controlled using the transmit precoders $\mathbf{M}_{k,n}$ designed for each user over each sub-channel resource. Since the decision variables $\mathbf{t}(i)$ depends on the queue state at $\mathbf{Q}(i-1)$, we drop the time variable i for the easy representation. The above problem defined in (7) can be rewritten to include the space-frequency allocation variable, namely $\mathbf{M}_{k,n}$, as

$$\begin{aligned}
\underset{\substack{\mathbf{M}_{k,n},\gamma_{l,k,n}\\ t_{l,k,n}}}{\text{maximize}} \quad & \sum_{k\in\mathcal{U}} Q_k \left(\sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right) && \text{(8a)}\\
\text{subject to.} \quad & t_{l,k,n} \le \log_2(1+\gamma_{l,k,n}) && \text{(8b)}\\
& \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b} \text{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\text{H}}\right) \le P_{\max},\ \forall\, b, && \text{(8c)}\\
& \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n} \le Q_k,\ \forall\, k\in\mathcal{U} && \text{(8d)}
\end{aligned}$$

where (8d) is due to the $[x]^+$ operation in (3) and $\gamma_{l,k,n}$ is 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 (8), we can formulate the sum rate maximization problem by setting the weights in (8a) as unity, leading to the problem as

$$\begin{aligned}
\underset{\substack{\mathbf{M}_{k,n},\gamma_{l,k,n}\\ t_{l,k,n}}}{\text{maximize}} \quad & \sum_{k\in\mathcal{U}}\sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n} && \text{(9a)}\\
\text{subject to.} \quad & \text{(8a), (8b), (8c) and (8d).} && \text{(9b)}
\end{aligned}$$

The problem defined in (9) provides an efficient queue minimizing approach as compared to (8), since the resource allocation is driven by the channel conditions as compared to the number of queued packets in the problem defined in (8). In both formulations, the resources allocated to the users are limited by the backlogged packets (8d), thereby handling the problem of over allocation. The resource allocation constraint in (8d) renders the problem difficult to be solved by the iterative algorithm to design the precoders, which will be discussed later.

B. Joint Space Frequency Resource Allocation (JSFRA)

In order to minimize the number of queued packets at the BSs, the precoders $\mathbf{M}_{k,n}$ are designed to distribute the resources in an efficient manner to the users by utilizing both spatial and frequency dimension. The problem defined in (8) raises a question, namely, the linearity of the weights used in the objective function. It can also be a quadratic function, providing more emphasis on the users with large number queued packets as compared to the users with the small queue sizes or some other function of $\mathbf{Q}$. In order to answer this question, in this section, we discuss the joint space-frequency resource allocation (JSFRA) problem formulation, which restricts the solution only to the power functions. For practical and tractability reasons, we impose a constraint that the maximum number of transmitted bits for the user k is limited by the packets available at the transmitter. As a result, the number of backlogged packets remaining in the system is given by

$$v_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} \log_2(1+\gamma_{l,k,n}) \ge 0\ \forall\, k\in\mathcal{U}. \tag{10}$$

The precoder design problem can be given minimizing the absolute value of the deviation in (10) raised to the exponent

q as minimize $\sum_{k\in\mathcal{U}} |v_k|^q$. The exponent q plays a vital role in the final allocation, which will be discussed later. Now, the problem of weighted queued packet minimization formulated as a q-norm minimization, is given by

$$\underset{\substack{\mathbf{m}_{l,k,n},\\ \mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{11a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b} \text{tr}\left(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\text{H}}\right) \le P_{\max},\ \forall\, b, \tag{11b}$$

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), and $\mathbf{M}_{k,n} \triangleq [\,\mathbf{m}_{1,k,n}\,\mathbf{m}_{2,k,n}\ldots\mathbf{m}_{L,k,n}\,]$ comprises the beamformers associated with the user k for n^{th} sub-channel transmission. It can be easily extended for user specific streams L_k instead of using the common L streams for all users. The expression for $t_{l,k,n}$ is due to the assumption of Gaussian signaling, so that the maximum achievable rate is $\log_2(1+\gamma_{l,k,n})$ for a given signal-to-interference-plus-noise ratio (SINR) $\gamma_{l,k,n}$. In (11b), we consider the sum power constraint for each BS across all sub-channels. The proposed solution presented in this section also applies to the sub-channel power constraint by replacing (11b) by corresponding formulations. Before proceeding further, we note that the constraint in (10) is handled implicitly by the definition of the q-norm in the objective of (11). 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. We comment on the choice of the norm q on the objective as below [3], [4].

- With $q = 1$, 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 (11) reduces to the queue weighted sum rate maximization (Q-WSRM) problem (9) when the queue size is large enough for all users.
- With $q = 2$, the objective prioritizes users with higher queued packets before considering the users with a smaller number of backlogged packets. This is 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 [2].
- With $q = \infty$, 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.

1) Solution via Successive Convex Approximation SCA: We present an iterative algorithm to solve (11) locally based on the idea of alternating optimization and successive convex approximation. For this purpose, from (2), we can explicitly reformulate (11) as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\\ \beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{12a}$$

$$\text{subject to} \quad \gamma_{l,k,n} \le \frac{\mathbf{w}_{l,k,n}^{\text{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}}{\beta_{l,k,n}} \triangleq f(\tilde{\mathbf{u}}_{l,k,n}), \tag{12b}$$

$$\beta_{l,k,n} \ge N_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\text{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2, \tag{12c}$$

$$\text{(11b)} \tag{12d}$$

where $\beta_{l,k,n}$ denotes the total interference seen by the l^{th} stream of the user k on the n^{th} sub-channel and let $\tilde{\mathbf{u}}_{l,k,n} \triangleq \{\mathbf{w}^{\mathrm{H}}_{l,k,n}, \mathbf{H}_{b_k,k,n}, \mathbf{m}_{l,k,n}, \beta_{l,k,n}\}$ 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 (12b) and (12c). However, this step is without loss of optimality leads to the same solution, since the inequalities in (12b) and (12c) are active for an optimal solution, following the same arguments as those in [5]. Intuitively, (12b) denotes the SINR constraint for $\gamma_{l,k,n}$, and (12c) gives an upper bound for the interference seen by the user $k \in \mathcal{U}_b$. The problem in (12) is known to be NP-hard even for the single antenna case [6], [7]. The reformulation in (12) allows a tractable solution as presented below. First, we note that the constraints (11b) are convex with involved variables. Thus, we only need to deal with (12b) and (12c). Towards this end, we resort to the traditional coordinate descent technique by fixing the transmit beamformers, and find the optimal linear receivers. 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 (11) w.r.t the $\mathbf{w}_{l,k,n}$, and the closed form solution is given by

$$\mathbf{w}_{l,k,n} = \mathbf{R}^{-1}_{l,k,n}\, \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n}, \tag{13a}$$

$$\mathbf{R}_{l,k,n} = \sum_{(j,i)\neq(l,k)} \mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\mathbf{m}^{\mathrm{H}}_{j,i,n}\mathbf{H}^{\mathrm{H}}_{b_i,k,n} + \mathbf{I}_{N_R}, \tag{13b}$$

which is the minimum mean squared error (MMSE) receive beamformers [6]–[8].

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 $\mathbf{w}_{l,k,n}$, (12c) can be written as a second-order cone (SOC) constraint. Thus, the difficulty is due to the non-convexity in (12b). To arrive at a tractable formulation, we adopt the successive convex approximation (SCA) method to handle (12b) 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 (12b) 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 $\beta_{l,k,n}$) [9]. According to the SCA method, we relax (12b) 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 (12b) 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 $\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}$ be represented by

$$p_{l,k,n} \triangleq \Re\left\{\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\}, \tag{14a}$$

$$q_{l,k,n} \triangleq \Im\left\{\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\}, \tag{14b}$$

and hence $f(\tilde{\mathbf{u}}_{l,k,n}) = (p^2_{l,k,n}+q^2_{l,k,n})/\beta_{l,k,n}$. Note that $p_{l,k,n}$ and $q_{l,k,n}$ are just symbolic notation and not the newly introduced optimization variables. In CVX [10], 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 (12b) by the following linear inequality constraint

$$2\frac{\tilde{p}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\left(p_{l,k,n} - \tilde{p}_{l,k,n}\right) + 2\frac{\tilde{q}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\left(q_{l,k,n} - \tilde{q}_{l,k,n}\right) + \frac{\tilde{p}^2_{l,k,n} + \tilde{q}^2_{l,k,n}}{\tilde{\beta}_{l,k,n}}\left(1 - \frac{\beta_{l,k,n} - \tilde{\beta}_{l,k,n}}{2\,\tilde{\beta}_{l,k,n}}\right) \geq \gamma_{l,k,n}. \tag{15}$$

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

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

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

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

$$\text{and (15).} \tag{16d}$$

The proposed algorithm is referred as queue minimizing (QM) JSFRA scheme with a sum 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 (11b)
update $\mathbf{w}_{l,k,n}$ with (13) and $\tilde{\mathbf{u}}_{l,k,n}$ with (15) using $\tilde{\mathbf{m}}_{l,k,n}$
repeat
    repeat
        solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (16)
        update the constraint set (15) with $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$ using (14) 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 (13) with the recent precoders $\mathbf{m}_{l,k,n}$
    $j=j+1$
until Queue convergence or $j\geq J_{\max}$
```

In the proposed solution, we replaced (12b) 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 [11]. In our case, we solve the sub problem inexactly, so the convergence proof of BCDM does not apply to our problem. Recall that [6] only approximates the non-convex objective in each iteration. 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. In a single receive antenna case, the proposed solution is guaranteed to converge to a Karush-Kuhn-Tucker (KKT) point as discussed in [5].

2) Solution via MSE reformulation: In this section, we provide an alternative formulation for the QM JSFRA scheme using the mean squared error (MSE) equivalence with the capacity as in [6], [12] for the WSRM problem. The proposed formulation uses the well known relation between the capacity and the MSE given by $C(\gamma) = \log_2(1/\epsilon)$, where $\epsilon = 1/(1+\gamma)$ is the MSE expression for decoding the transmitted symbols using MMSE receivers and C being the capacity of the system. Using the MSE relation, (12) is written as

$$\begin{aligned} \text{minimize} \quad & \|\tilde{\mathbf{v}}\|_q && \text{(17a)} \\ \text{subject to} \quad & t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n}), && \text{(17b)} \\ & \epsilon_{l,k,n} \geq \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 + N_0 \|\mathbf{w}_{l,k,n}\|^2 && \text{(17c)} \\ & \text{and (11b)}, && \text{(17d)} \end{aligned}$$

where (17c) bounds the MSE by $\epsilon_{l,k,n}$ and (17b) relaxes the transmitted rate $\mathbf{t}_{l,k,n}$ using the the MSE relation.

The QM JSFRA problem using alternative MSE formulation given by (17) is non-convex due to the set defined by the constraint (17b). In order to solve this efficiently, we use the SCA method as discussed earlier in Section III-B by using the linear under estimator for the convex function on the r.h.s of (17b). The first order Taylor approximation around a fixed MSE value $\tilde{\epsilon}_{l,k,n}$ for (17b) is given by

$$-\log_2(\tilde{\epsilon}_{l,k,n}) - \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\,\tilde{\epsilon}_{l,k,n}} \geq t_{l,k,n}. \tag{18}$$

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 can be given as

$$\begin{aligned} \text{minimize} \quad & \|\tilde{\mathbf{v}}\|_q && \text{(19a)} \\ \text{subject to} \quad & \text{(11b), (18)} && \text{(19b)} \\ & \epsilon_{l,k,n} \geq \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}\right|^2 + \sum_{(j,i)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,i,n} \mathbf{m}_{j,i,n}\right|^2 + N_0 \|\mathbf{w}_{l,k,n}\|^2. && \text{(19c)} \end{aligned}$$

The algorithmic representation is similar to Algorithm 1, with (16) is replaced by (19). The convergence of the MSE reformulation algorithm follows the similar comments drawn in [6], [8]. Since at each KKT point, the gradient of the original problem and the MSE reformulation problem are made equal by the weights in the MSE reformulation, the problem is guaranteed to converge to the optimal solution by following the same stationary KKT points.

C. Per Sub-Channel Resource Allocation Schemes

In this method, titled as queue minimizing (QM) spatial resource allocation (SRA) scheme, we limit the resource allocation over the spatial dimension only by performing JSFRA design over each sub-channel in an instant. This scheme can be considered in a system employing the fractional frequency reuse (FFR) like power restrictions on certain sub-channels. It can also be used to reduce the problem dimension by limiting the allocation over the spatial dimension only. It is well suited for the persistent scheduling like schemes, where certain sub-channels are dedicated to only certain users. Imposing the user restrictions in the JSFRA scheme adds a non-convex constraint which will be difficult to solve.

For a scheduling slot, the precoders are designed over each sub-channel in a sequential manner with the corresponding user queues are updated with the total transmission allocations made in the previous sub-channels on the same slot.

The queue update is common for SRA and band-wise Q-WSRM schemes, and the queue length controls the design of precoders for the allocation of spatial resources for each sub-channel. The queues are updated before designing the precoders for each sub-channel n as

$$Q_{k,n} = \max\Big\{Q_k - \sum_{j=1}^{n-1} \sum_{l=1}^{L} t_{l,k,j}, 0\Big\},\ \forall\, k \in \mathcal{U} \tag{20}$$

where Q_k is given by (3) for the user k. The weight for the sub-channel n is given by (20), which uses the allotted transmission bits $t_{l,k,j}$ evaluated from the earlier sub-channels $j < n$. The QM SRA scheme depends on the permutation pattern of the sub-channel selection order for the precoder design and the performance.

IV. Distributed Solutions

This section addresses the distributed precoder designs for the proposed JSFRA scheme. The formulation in (16) 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 (PD) and the alternating directions method of multipliers (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 defined by (16) can be equivalently written as

$$\underset{\substack{t_{l,k,n},\gamma_{l,k,n}\\ \mathbf{m}_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{21a}$$

$$\text{subject to} \quad \text{(16b)} - \text{(16d)}, \tag{21b}$$

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] and [7], the interference constraint given by (16b) can be expressed by grouping the interference contribution from each BSs in the system as

$$\beta_{l,k,n} \geq \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2 + \sum_{i\in\mathcal{U}_{b_k}\backslash\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \underbrace{\sum_{b\in\bar{\mathcal{B}}_{b_k}} \sum_{i\in\mathcal{U}_b} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2}_{\text{neighboring BSs interference}} + N_0, \forall k \in \mathcal{U}. \tag{22}$$

In (22), the neighboring BSs signals can be written explicitly using the relaxed expression as

$$\beta_{l,k,n} \geq \sum_{\substack{j=1\\j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2$$

$$+ \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} \;+\; N_0, \tag{23a}$$

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

$$\zeta_{l,k,n,b} \geq \sum_{i\in\mathcal{U}_b} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2, \forall b \in \bar{\mathcal{B}}_{b_k}. \tag{24}$$

The expression in (23) can be solved either by primal or by dual decomposition based on the usage of the coupling variable $\zeta_{l,k,n,b}$ in the optimization problem. In both approaches, the distributed problem is solved by a two level optimization approach in which the local problems are solved at the BS level and the master problem controls the variables involved in each local problem [15].

A. PD Approach

The centralized problem described in (21) can be solved in a distributed approach using the PD method. The decentralized approach decomposes the centralized problem as a two level optimization problem in which the multiple sub problems are coordinated using a master problem. The coordination of the multiple subproblems can be carried out by exchanging the information across the coordinating BSs in $\mathcal{B}$ via backhaul. The distributed solution is achieved by fixing the interference level $\zeta_{l,k,n,b}, \forall k \in \mathcal{U}_{b_k}$ from the BSs $b \in \bar{\mathcal{B}}_{b_k}$ so as to decouple the variables $\zeta_{l,k,n,b}$ [13], [16]. Now, the subproblems are governed by the master problem by updating the interference thresholds at each iteration using the subgradients.

As discussed in Section IV, the constraint involving the coupling variables (16b) can be relaxed using the equivalent formulation in (23). 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

$$\underset{\substack{\gamma_{l,k,n}\\ \mathbf{m}_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_{b_k}\|_q \tag{25a}$$

$$\text{subject to} \quad \sum_{n=1}^{N} \sum_{k\in\mathcal{U}_{b_k}} \operatorname{tr}(\mathbf{M}_{k,n}\mathbf{M}_{k,n}^{\mathrm{H}}) \leq P_{\max}, \tag{25b}$$

$$\beta_{l,k,n} \geq \sum_{\substack{j=1\\j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2$$

$$+ \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} \;+\; N_0 \tag{25c}$$

$$\zeta_{l',k',n,b_k} \geq \sum_{k\in\mathcal{U}_b} \sum_{l=1}^{L} |\mathbf{w}_{l',k',n}^{\mathrm{H}} \mathbf{H}_{b_k,k',n} \mathbf{m}_{l,k,n}|^2, \; \forall k' \in \bar{\mathcal{U}}_{b_k}, \; \forall n \in \mathcal{C} \tag{25d}$$

$$\text{and } (15), \tag{25e}$$

where ζ_{l',k',n,b_k} corresponds to the fixed maximum interference caused by the transmission from the BS b_k to the l'^{th} spatial stream of the k'^{th} user. Let $\boldsymbol{\zeta}^{\{b_k\}}$ be the vector representing the fixed interference levels relevant to the BS b_k in a fully connected network (in practice it will be less), which is given by

$$\begin{aligned}
\boldsymbol{\zeta}_{k,n,b} &= [\ \zeta_{1,k,n,b}, \ldots, \zeta_{L,k,n,b}\ ] && \text{(26a)}\\
\boldsymbol{\zeta}_n^{\{b_k\}} &= \Big[\ \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(1)}, \ldots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(1),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \\
&\qquad \ldots, \boldsymbol{\zeta}_{\mathcal{U}_{b_k}(|\mathcal{U}_{b_k}|),n,\bar{\mathcal{B}}_{b_k}(|\bar{\mathcal{B}}_{b_k}|)}, \ldots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(1),n,b_k}, \ldots, \boldsymbol{\zeta}_{\bar{\mathcal{U}}_{b_k}(|\bar{\mathcal{U}}_{b_k}|),n,b_k}\ \Big] && \text{(26b)}\\
\boldsymbol{\zeta}^{\{b_k\}} &= \Big[\ \boldsymbol{\zeta}_1^{(b_k)}, \ldots, \boldsymbol{\zeta}_N^{(b_k)}\ \Big], && \text{(26c)}
\end{aligned}$$

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

$$n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| = \big(|\bar{\mathcal{B}}_{b_k}||\mathcal{U}_{b_k}| + |\bar{\mathcal{U}}_{b_k}|\big)\, LN \tag{27}$$

The problem defined in (25) is already decoupled across the BSs by fixing the inter-cell interference vector (26) for all BSs. For a fixed interference threshold $\boldsymbol{\zeta}^{\{b\}}$, the precoders are designed using (25). The local subproblem (25) 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 and the subproblems follows the classical BCDM approach to achieve the optimal solution by alternating the variable set while keeping other set constant in an iterative manner. The master problem controlling multiple subproblems is given by

$$\begin{aligned}
&\underset{\boldsymbol{\zeta}}{\text{minimize}} && \sum_{b\in\mathcal{B}} \alpha_b^\star(\boldsymbol{\zeta}^{\{b\}}) && \text{(28a)}\\
&\text{subject to} && \boldsymbol{\zeta}^{\{b\}} \in \mathbb{R}_+^{n_b}, \forall b \in \mathcal{B}, && \text{(28b)}
\end{aligned}$$

where $\alpha_b^\star(\boldsymbol{\zeta}^{\{b\}})$ denotes the optimal solution for (25) 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 $\boldsymbol{\zeta} = \Big[\boldsymbol{\zeta}^{\{\mathcal{B}(0)\}}, \boldsymbol{\zeta}^{\{\mathcal{B}(1)\}}, \ldots, \boldsymbol{\zeta}^{\{\mathcal{B}(|\mathcal{B}|)\}}\Big]$.

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 [16] as

$$\zeta_{l,k,n,b}^{\{b_k\}(i)} = \Big[\zeta_{l,k,n,b}^{\{b_k\}(i-1)} - \rho\, s_{l,k,n,b}^{\{b_k\}(i-1)}\Big]^+, \forall b \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b, \tag{29}$$

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 (28) 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 Lagrangian of the primal problem (26) as

$$\begin{aligned}
&\underset{\substack{\gamma_{l,k,n}, \mathbf{m}_{l,k,n},\\ \beta_{l,k,n}, \mu_{l,k,n}^{\{b_k\}}}}{\text{minimize}} && \|\tilde{\mathbf{v}}_{b_k}\|_q + \Bigg(\beta_{l,k,n} - \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,k,n}|^2 \\
& && + \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}} \sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b}^{(i-1)} + N_0\Bigg) \mu_{l,k,n}^{\{b_k\}} \\
& && + \Bigg(\zeta_{l',k',n,b_k}^{(i-1)} - \sum_{k\in\mathcal{U}_b} \sum_{l=1}^{L} |\mathbf{w}_{l',k',n}^{\mathrm{H}} \mathbf{H}_{b_k,k',n} \mathbf{m}_{l,k,n}|^2\Bigg) \mu_{l',k',n}^{\{b_k\}},\ \forall k' \in \bar{\mathcal{U}}_{b_k},\ \forall n \in \mathcal{C} && \text{(30a)}\\
&\text{subject to} && \text{(25b) and (15).} && \text{(30b)}
\end{aligned}$$

Now the primal and the dual variables $\mu^{\{b_k\}}_{l,k,n}$ and $\mu^{\{b_k\}}_{l',k',n}$ corresponding to the constraints (25c) and (25d) are obtained by solving (30).

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*, $s^{\{b_k\}(i)}_{l,k,n,b} = \mu^{\{b_k\}}_{l,k,n} - \mu^{\{b\}}_{l,k,n}$. With the newly estimated subgradient value $s^{\{b_k\}(i)}_{l,k,n,b}$, the interference terms corresponding to the BS b_k are updated using (29).

Once the interference vector converges $\boldsymbol{\zeta}^{\{i\}} \approx \boldsymbol{\zeta}^{\{i-1\}}$ or iterated for certain times, say, $J_{\max}$, the SCA update is performed by exchanging the transmit and the receive precoders to update the respective variables $\tilde{p}_{l,k,n}$, $\tilde{q}_{l,k,n}$ and $\tilde{\beta}_{l,k,n}$. The SCA update can also be performed until convergence or for certain threshold $I_{\max}$. The algorithmic representation of the PD approach is detailed in Algorithm. 2. In this case, we perform SCA and the receiver update at the same instant compared with the update of the recieve beamformers after the SCA convergence. The approach followed in the algorithm provides the same solution as that of the separate update, since the receive beamformers are optimal for the given transmit precoders.

Algorithm 2: PD based decentralized 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 (11b)
update $\mathbf{w}_{l,k,n}$ with (13) and $\tilde{\mathbf{u}}_{l,k,n}$ with (15) using $\tilde{\mathbf{m}}_{l,k,n}$
initialize the interference threshold $\zeta^{\{0\}}_{l,k,n,b} \forall b \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_{b_k}, \forall l, n$
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 $\mu^{\{b\}}_{l,k,n}$ using (30)
        exchange $\mu^{\{b\}}_{l,k,n}$ across the coordinating BSs in $\mathcal{B}$
        update $\zeta^{\{b\}(j+1)}_{l,k,n,b}$ using (29) locally
        $j = j + 1$
    until convergence or $j \geq J_{\max}$
    update the receive beamformers $\mathbf{w}_{l,k,n}$ using (13) with the recent precoders $\mathbf{m}_{l,k,n}$
    exchange the transmit and the receive precoders $\mathbf{M}_{k,n}$ and $\mathbf{W}_{k,n}$ $\forall k \in \mathcal{U}_b$ 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 (14) and (12c) for the SCA approach (or)
    update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (19c) with equality for the MSE formulation approach
    $i = i + 1$
until convergence or $i \geq I_{\max}$
```

B. ADMM approach

In contrast to the primal decomposition problem, the dual decomposition (DD) performs the distributed precoder design by relaxing the interference constraints by including it in the objective function of each subproblem with a penalty pricing [14], [15]. In order to decouple the problem (21), the global coupling variables $\zeta_{l,k,n,b}$ in (23) are replaced by the local copies at each BS b denoted by $\zeta^{\{b\}}_{l,k,n,b}$, $\forall\, b \in \mathcal{B}$, which is then used in the problem as an optimization variable.

Let $\boldsymbol{\zeta}^{\{b_k\}}$ be the locally maintained vector formed by stacking the interference entries relevant to the BS b_k. Let $\boldsymbol{\zeta}$ be the complete interference entries of all BSs in the set $\mathcal{B}$ stacked as a vector as

$$\begin{aligned}\boldsymbol{\zeta} = \big[&\zeta_{1,\bar{\mathcal{U}}_1(1),1,1}, \dots, \zeta_{L,\bar{\mathcal{U}}_1(1),1,1}, \dots, \zeta_{L,\bar{\mathcal{U}}_1(|\bar{\mathcal{U}}_1|),1,1},\\ &\dots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),1,N_B}, \dots, \zeta_{L,\bar{\mathcal{U}}_{N_B}(|\bar{\mathcal{U}}_{N_B}|),N,N_B}\Big] \qquad (31a)\\ n_{b_k} = |\boldsymbol{\zeta}^{\{b_k\}}| = NL\sum_{b\in\mathcal{B}} \left|\bar{\mathcal{U}}_b\right|, \qquad (31b)\end{aligned}$$

where $\boldsymbol{\zeta}(b_k)$ denote the entries corresponding to the BS b_k and the vector $\boldsymbol{\nu}^{\{b_k\}}$ stacks the dual variables corresponding to the equality condition $\boldsymbol{\zeta}^{\{b_k\}} = \boldsymbol{\zeta}(b_k)$. The equality constraint for the local and the global interference vector $\zeta^{\{b_k\}}_{l,k,n,b} = \zeta_{l,k,n,b}, \ \forall\, b \in \bar{\mathcal{B}}_{b_k}, k \in \mathcal{U}$ and $\forall k \in \bar{\mathcal{U}}_{b_k}, b = b_k$ is relaxed by the partial Lagrangian in the objective as

$$\begin{aligned}\underset{\substack{\gamma_{l,k,n},\nu^{\{b_k\}}_{l,k,n,b}\\ \mathbf{m}_{l,k,n},\beta_{l,k,n},\zeta^{\{b_k\}}_{l,k,n,b}}}{\text{minimize}} \quad & \|\tilde{\mathbf{v}}_{b_k}\|_q + \boldsymbol{\nu}^{\{b_k\}T}\left(\boldsymbol{\zeta}^{\{b_k\}} - \boldsymbol{\zeta}(b_k)\right) \qquad (32a)\\ \text{subject to} \quad & \beta_{l,k,n} \geq \sum_{\substack{j=1\\ j\neq l}}^{L} |\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2\\ & \quad + \sum_{i\in\mathcal{U}_{b_k}\setminus\{k\}}\sum_{j=1}^{L} |\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta^{\{b_k\}}_{l,k,n,b} \ + \ N_0 \qquad (32b)\\ & \zeta^{\{b_k\}}_{l',k',n,b_k} \geq \sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} |\mathbf{w}^{\mathrm{H}}_{l',k',n}\mathbf{H}_{b_k,k',n}\mathbf{m}_{l,k,n}|^2, \ \forall k' \in \bar{\mathcal{U}}_{b_k}, \ \forall n \in \mathcal{C} \qquad (32c)\\ & \boldsymbol{\nu}^{\{b_k\}} \in \mathbb{R}^{n_{b_k}}_{+}, \ \text{(15) and (25b).} \qquad (32d)\end{aligned}$$

It can be seen from the objective (32a) that the global interference $\boldsymbol{\zeta}(b_k)$ can be dropped from the objective (32a) without affecting the optimal solution, since it is constant for the subproblems.

Now, the problem defined in (32) can be decoupled to solve for the precoders at each BS by using interference vector $\boldsymbol{\zeta}^{\{b_k\}}$ as a optimization variable for the fixed interference price (dual variable) $\boldsymbol{\nu}^{\{b_k\}}$. The convergence of the problem with the objective function (32) is slower due to the linear penalty term for the interference assumptions at each BS, thereby providing more emphasis only when the interference deviation is significantly large.

In order to bound the interference assumptions $\zeta^{\{b_k\}}_{l,k,n,b}$ and $\zeta^{\{b\}}_{l,k,n,b}$ between the BSs b_k and b, adding a scaled quadratic penalty of the interference deviation provides better convergence properties without affecting the optimal solution [16], [17]. At the optimal point, the actual and the assumed (local) interference values are equal, thereby providing no contribution to the objective. This way of adding a scaled quadratic penalty term in the objective function is known as ADMM method for solving the dual decomposition problem. In addition to the faster convergence, it also identifies the optimal step size for the dual variable update. It can be shown that the optimal step size is equal to the scaling factor ρ used for the penalty term in the objective function [16], [17]. Now, the subproblem at each BS, obtained via ADMM decomposition, is given by

$$\begin{aligned} \underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\\ \beta_{l,k,n},\zeta_{l,k,n,b}^{\{b_k\}(i)}}}{\text{minimize}} \quad & \|\tilde{\mathbf{v}}_{b_k}\|_q + \boldsymbol{\nu}^{\{b_k\}(i)T}\boldsymbol{\zeta}^{\{b_k\}(i)} + \frac{\rho}{2}\left\| \boldsymbol{\zeta}^{(i)}(b_k) - \boldsymbol{\zeta}^{\{b_k\}(i)} \right\|_2^2 && \text{(33a)} \\ \text{subject to} \quad & \text{(32b) -- (32d)}, && \text{(33b)} \end{aligned}$$

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

Once the subproblems are solved at each BS, the update for the global interference vector and the dual variables can be performed at the BSs locally by exchanging the local copies of the interference vector $\boldsymbol{\zeta}^{\{b\}}, \forall b \in \mathcal{B}$. Since the entries in $\boldsymbol{\zeta}^{(i)}$ relates exactly two BSs only, each entry in the $\boldsymbol{\zeta}^{(i)}$ can be updated using a function operating on the local copies from the corresponding BSs. For instance, the entry $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)}$ depends on the local interference value $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b_k\}(i-1)}$ assumed by the BS b_k and the actual interference caused by the BS b as in $\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i-1)}$. It can be updated by either using

$$\begin{aligned} \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)} &= \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i-1)} \text{ (or)} && \text{(34a)} \\ \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{(i)} &= \frac{1}{2}\left(\zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b\}(i-1)} + \zeta_{l,\mathcal{U}_{b_k}(1),n,b}^{\{b_k\}(i-1)} \right), && \text{(34b)} \end{aligned}$$

as discussed briefly in [14]. 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

$$\nu_{l,k,n,b}^{\{b_k\}(i)} = \nu_{l,k,n,b}^{\{b_k\}(i-1)} + \rho\left(\zeta_{l,k,n,b}^{\{b_k\}(i-1)} - \zeta_{l,k,n,b}^{\{b\}(i-1)} \right), \forall b, b_k \in \mathcal{B}, \forall k \in \bar{\mathcal{U}}_b. \tag{35}$$

The ADMM approach uses the quadratic penalty term in the objective for better convergence properties compared to the plain DD scheme. The addition of the quadratic term in the objective translates the objective function into a strict convex function. The subproblem defined by (33) is convex and the objective has the convex function, therefore the convergence is guaranteed as discussed in [17], [18]. The algorithmic representation of the ADMM based approach for decentralization is given in Algorithm 3.

C. Distributed MSE Formulation

So far, we have discussed the problem of minimizing the number of backlogged packets at the BSs using the precoders by formulating it as an optimization problem, which can be handled by the existing solvers CVX [10]. In this section, we provide an iterative precoder design for the JSFRA formulation approached via MSE reformulation as discussed in III-B2. Even though the JSFRA scheme using SCA approach and the MSE reformulation approach obtained the solutions by relaxing the nonconvex constraints by the half spaces, the structure of the MSE reformulation problem allows us to decouple the problem readily without further approximations.

In order to come up with the iterative solution, we need to find the KKT equations for the the MSE reformulation problem defined in (19). It can be seen that the objective function (19a) has the norm function, which requires the sum of absolute values of the vector entries raised to the power q. Since the absolute of a function is not differentiable, the problem is usually

Algorithm 3: ADMM based decentralized 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 (11b)
update $\mathbf{w}_{l,k,n}$ with (13) and $\tilde{\mathbf{u}}_{l,k,n}$ with (15) using $\tilde{\mathbf{m}}_{l,k,n}$
initialize the 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 (33)
        exchange $\boldsymbol{\zeta}^{\{b\}(j)}$ across the coordinating BSs in $\mathcal{B}$
        update the dual variables in $\boldsymbol{\nu}^{\{b\}(j+1)}$ using (35)
        update the interference vector $\boldsymbol{\zeta}^{\{b\}(j+1)}$ using (34a) or (34b)
        $j = j + 1$
    until convergence or $j \geq J_{\max}$
    update the receive beamformers $\mathbf{w}_{l,k,n}$ using (13) with the recent precoders $\mathbf{m}_{l,k,n}$
    exchange the transmit and the receive precoders $\mathbf{M}_{k,n}$ and $\mathbf{W}_{k,n}$ $\forall k \in \mathcal{U}_b$ 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 (14) and (12c) for the SCA approach (or)
    update $\tilde{\epsilon}_{l,k,n}$ with the recent precoders using (19c) with equality for the MSE formulation approach
    $i = i + 1$
until convergence or $i \geq I_{\max}$
```

solved by the subgradient approach. Since the subgradients are not unique and the selection of a subgradient and a step size are not guaranteed to be optimal, in this approach, we drop the absolute value operation for the queue deviation function so as to obtain a unique supporting vector. It is valid in the following special cases without affecting the optimal solution

- when the exponent q is odd and $q > 1$ or,
- when the number of backlogged packets for each user is larger than the available resources for the transmission $Q_k \gg \sum_{n=1}^{N} \sum_{l=1}^{L} t_{l,k,n}$.

The first condition is from the fact that the differential of a function with power q leaves the power to $q - 1$, which helps us to drop the absolute value condition when $q - 1$ is even. This condition cannot be applied for $q = 1$ since the rate constraint is eliminated by the differential operation.

By removing the absolute value operator from (19), the problem can be rewritten as

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

$$\text{subject to}$$

$$\alpha_{l,k,n} : \quad \epsilon_{l,k,n} = \left| 1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n} \right|^2 + \sum_{(x,y) \neq (l,k)} \left| \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_y,y,n} \mathbf{m}_{x,y,n} \right|^2 + N_0 \left\| \mathbf{w}_{l,k,n} \right\|^2 \tag{36b}$$

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

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

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (36b), (36c) and (36d). The equality condition is used for (36b) and (36c), since the variables $\epsilon_{l,k,n}$ and $t_{l,k,n}$ are used for the corresponding expressions on the other side.

Now the Lagrangian of the problem expressed in (36), with the corresponding dual variables, is used obtain the KKT expressions by partially differentiating with respect to the variables $t_{l,k,n}$, $\epsilon_{l,k,n}$, $m_{l,k,n}$, $w_{l,k,n}$ and the dual variables $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b. Since the problem of precoder design using MSE reformulation is solved by the iterative method, we use the iteration index i in the superscript to represent the variables at the respective iteration instants. The discussions on the KKT expressions are provided in the Appendix A. Upon solving the KKT expressions, we obtain the following iterative solution with the variables update as

$$\alpha_{l,k,n}^{(i)} = \frac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i-1)}} \tag{37a}$$

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

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

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

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

$$\mathbf{w}_{l,k,n}^{(i)} = \left(\sum_{(x,y)\neq(l,k)} \mathbf{H}_{b_y,k,n} \mathbf{m}_{x,y,n}^{(i-1)} \mathbf{m}_{x,y,n}^{\mathrm{H}\,(i-1)} \mathbf{H}_{b_y,k,n}^{\mathrm{H}} + \mathbf{I}_{N_R}\right)^{-1} \mathbf{H}_{b_k,k,n}\, \mathbf{m}_{l,k,n}^{(i-1)}. \tag{37f}$$

The KKT solutions provided in (37) are solved in an iterative manner by initializing the MSE variables $\epsilon_{l,k,n}^{(0)}$ and the throughput variables $t_{l,k,n}^{(0)}$ randomly. It can be initiated by solving (37d) and (37e) for a random transmit precoders $\mathbf{M}_{k,n}$ satisfying (36d). It can be noted that the proposed iterative algorithm converges faster than the primal decomposition and the ADMM based dual decomposition scheme, since the later schemes are based on the subgradient approach and the former one achieves the KKT points at each iteration. In (37), all expressions provide the closed form solution for the respective variables except the transmit precoders in (37c), which depends on the variable δ_b. It can be seen that the variable δ_b are associated only to the BS b, which can be efficiently solved by the bisection method satisfying the power constraint (36d). After each iteration instant, the transmit and the receive precoders are exchanged across the coordinating BSs in $\mathcal{B}$ in order to reach the optimal point in the next iteration.

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 $\mathbf{W}_{k,n}$ evaluated at the receivers. Upon receiving the uplink precoded pilots by the BS, the effective channel $\mathbf{W}_{k,n}^{\mathrm{H}\,(i-1)} \mathbf{H}_{b,k,n}$ can be used in the expression (37c) to update the transmit precoders at the respective BSs [19]. 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 $\mathbf{w}_{l,k,n}\ \forall\, l \in \{1, 2, \dots, L\}$
Initialize: $i = 1$ and the transmit precoders $\mathbf{m}^{(0)}_{l,k,n}$ randomly satisfying the total power constraint (36d)
Initialize: $\epsilon^{(0)}_{l,k,n}$ randomly, $\mathbf{w}^{(0)}_{l,k,n}$ with (37f) and $t^{(0)}_{l,k,n}$ using (37e)
set the maximum iteration counter $I_{\max}$ to a valid number
for each BS $b \in \mathcal{B}$, perform the following procedure
repeat
    solve for the dual variables $\alpha^{(i)}_{l,k,n}$ and $\sigma^{(i)}_{l,k,n}$ using (37a) and (37b)
    update the transmit precoders $\mathbf{M}^{(i)}_{k,n}$ with $\delta_b$ using (37c) by the bisection method satisfying (36d)
    update the MSE variable $\epsilon^{(i)}_{l,k,n}$ and the throughput variable $t^{(i)}_{l,k,n}$ using (37d) and (37e)
    evaluate the receive beamforming vector $\mathbf{W}_{k,n}$ by (37f)
    exchange the transmit and the receive precoders $\mathbf{M}^{(i)}_{k,n}, \mathbf{W}^{(i)}_{k,n}$ across the coordinating BSs in $\mathcal{B}$
    $i = i + 1$
until convergence or $i \geq I_{\max}$
```

Users	Queued Packets	Channel Gains			Joint Q-WSRM w/o Queue constraints (8d)			JSFRA Scheme			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	6	2.26	0.69	0.73	0	0	0	5.56	0	0	0	0	2.68
2	6	0.53	1.93	1.40	0	5.29	0	0	5.33	0	0	5.26	0
3	6	2.45	1.31	2.42	5.84	0	5.95	0	0	5.95	5.93	0	0
Remaining backlogged packets (χ)					6.70 bits			1.14 bits			4.11 bits		

TABLE I: Sub channel wise allocation for a scheduling instant

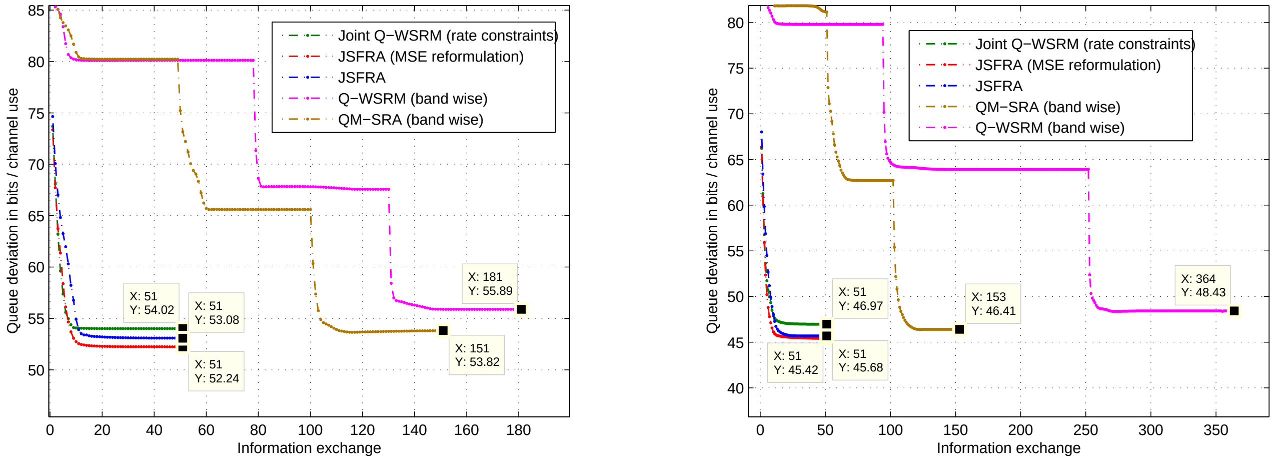
V. Simulation Results

A. Centralized Solutions

In this section, we discuss the performance of the above mentioned algorithms in different scenarios on minimizing the number of queued packets at the transmitter. To begin with, we consider a single cell single-input single-output (SISO) system 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 is kept as $Q_k = 6$ bits. With these assumptions, the resources allocated for the users on each sub-channel by various algorithms are listed in Table I.

Table I shows the channel seen by the users over each sub-channel followed by the rates allocated over each sub-channel by three different algorithms, joint Q-WSRM allocation without the maximum rate constraints (8d), JSFRA scheme and the band-wise Q-WSRM scheme. Since the total allowed transmission constraint is present only on the JSFRA scheme, the total number of packets remained after the current transmission slot is minimum for the JSFRA scheme with 1.14 bits, where $\chi = \sum_{k=1}^{K} \, [Q_k - t_k]^+$. The precoder design for the optimal allocation in band-wise Q-WSRM depends on the order of selecting the sub-channels, which leads to an exhaustive search. In this scenario, the formulation in (8) and (9) performs the same as compared to the JSFRA scheme, *i.e*, with the maximum rate constraints, the joint Q-WSRM scheme performs the same as the JSFRA scheme.

In order to understand the behavior in a MIMO framework, we consider a system with $N = 3$ sub-channels and $N_B = 2$ BSs, each equipped with $N_T = 4$ transmit antennas operating at 10dB SNR, serving $K = 8$ users with N_R antennas each. The users are assumed to be at the cell-edge with the maximum interference seen from the neighboring BSs is limited to -3



(a) 4×1 system

(b) 4×2 system

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

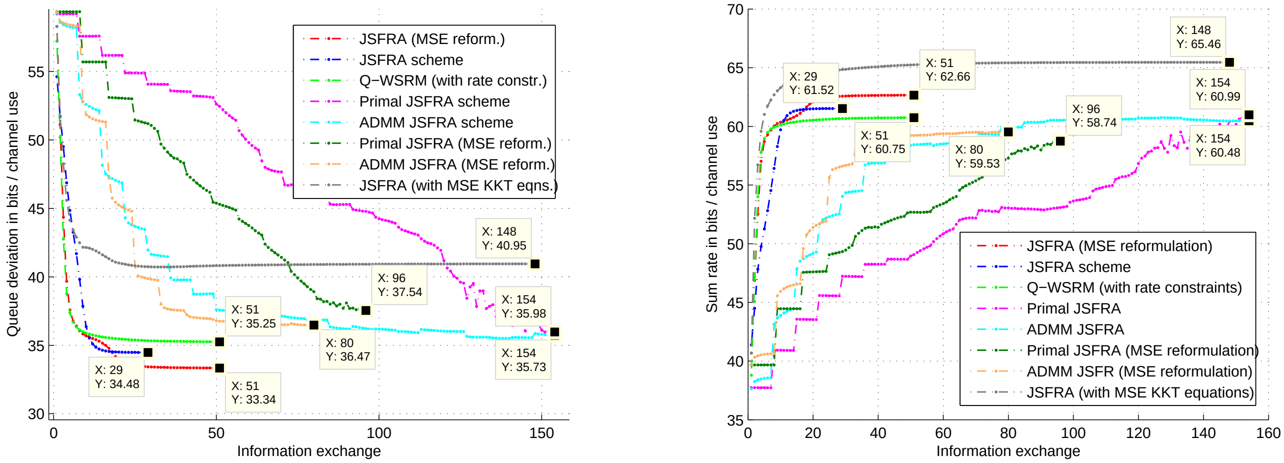
q	user indices								χ
1	12.0	6.15	5.32	12.0	6.95	11.1	11.9	10.8	19.71
2	11.9	7.3	5.9	10.1	9.19	10.1	10.8	10.3	20.48
∞	9.15	9.15	9.15	9.15	9.16	9.15	9.15	9.15	22.75

TABLE II: Queue information for $N = 5$ sub-channels

dB and each BS are associated with $|\mathcal{U}_b|$ users respectively.

Fig. 1a shows the performance of the above discussed schemes for a single receive antenna system. The figure compares the total number of SCA iterations required by the JSFRA, JSFRA with sum power constraint, SRA and Q-WSRM schemes to achieve the optimal resource allocation to minimize the number of backlogged packets during the given scheduling instant. The convergence of the proposed JSFRA is much quicker compared to the band-wise allocation schemes like the Q-WSRM and the SRA schemes. The waterfall like behavior for the band-wise allocation schemes suggests the rapid convergence when there is a band switch over, *i.e*, when the queues are updated from the earlier sub-channels to find the precoders for the current sub-channel. The convergence at each sub-channel is iterated for the accuracy of $\approx 10^{-4}$ or for a predetermined count $I_{\max}$, which creates the flat region between each waterfall behavior. The backlogged bits and the iteration count by various schemes at the convergence point are marked in the figures using data tips.

Fig. 1b compares the convergence behavior of the precoder designs, which allocates the space-frequency resources to the users to minimize the number of backlogged packets. Different values for the exponent q are compared in Table II for the system configuration $\{N_B, K, N_T, N_R\} = \{2, 8, 4, 1\}$. The number of backlogged packets at each user before the current scheduling is fixed to be 12 bits. It is evident that the exponent $q = 1$ shows the greedy resource allocation in comparison with the fair scheduling achieved using $q = \infty$. The $q = \infty$ norm does the fair scheduling even the path loss experienced by the users are different. It tries to minimize the variance across the backlogged packets of each user present in the system.



(a) Queue deviation

(b) Sum rate performance

Fig. 2: Convergence plot for $\{N, N_B, K, N_R\} = \{5, 2, 12, 1\}$ model

B. Distributed Solutions

In this section, we discuss the simulation performances of the algorithms briefed in Section III and IV. The comparison studies are made using the queue deviation and sum rate convergence of the algorithms discussed so far using different system models. The distributed algorithms for the JSFRA scheme and the equivalent MSE representation are compared using the excess backlogged packets present after the current iteration instant.

Fig. 2 demonstrates the performance of the distributed algorithms in an OFDM framework modeled with $N = 5$ sub-channels with $N_B = 2$ BSs, each equipped with $N_T = 4$ transmit antennas. Each BS serves $|\mathcal{U}_b| = 4, \forall b \in \{1, 2\}$ users in a coordinated manner so as to minimize the interference caused to the neighboring BS users. Fig. 2a shows the total number of backlogged packets at the end of each iteration and Fig. 2b plots the total rate achieved by various algorithms at each iteration. As discussed earlier in the distributed section, the rate of convergence of any distributed algorithm depends on the proper choice of a step size. In particular, the primal decomposition method is more susceptible to the choice of the step size compared to the ADMM approach. Since the primal decomposition iterates at each instant by fixing the interference thresholds, it may lead to infeasible solutions if the initial interference values are not feasible and also with the updated interference value as in (29).

Fig. 2 also compares the performance of the iterative algorithm for JSFRA scheme via MSE reformulation using the KKT conditions discussed in Section IV-C. The performance is inferior compared to the other algorithms due to the violation in the assumptions made in the algorithm design *i.e*, $Q_k \gg t_k$. Since in this case, we are comparing the $q = 1$ case, the queue condition $Q_k \gg t_k$ needs to be satisfied in order to remove the absolute operator from the expression in (16a) to form (36a). The total number of backlogged packets after the each iteration reduces monotonically in all cases and converges to the points closer to the centralized solution $\chi = 33$ bits. As seen in Fig. 2, the ADMM approach provides better convergence compared to the primal approach in the initial phase due to the quadratic penalty term included in the objective, which gets pronounced initially due to the large deviation between the assumed and the actual interference across the coordinating BSs.

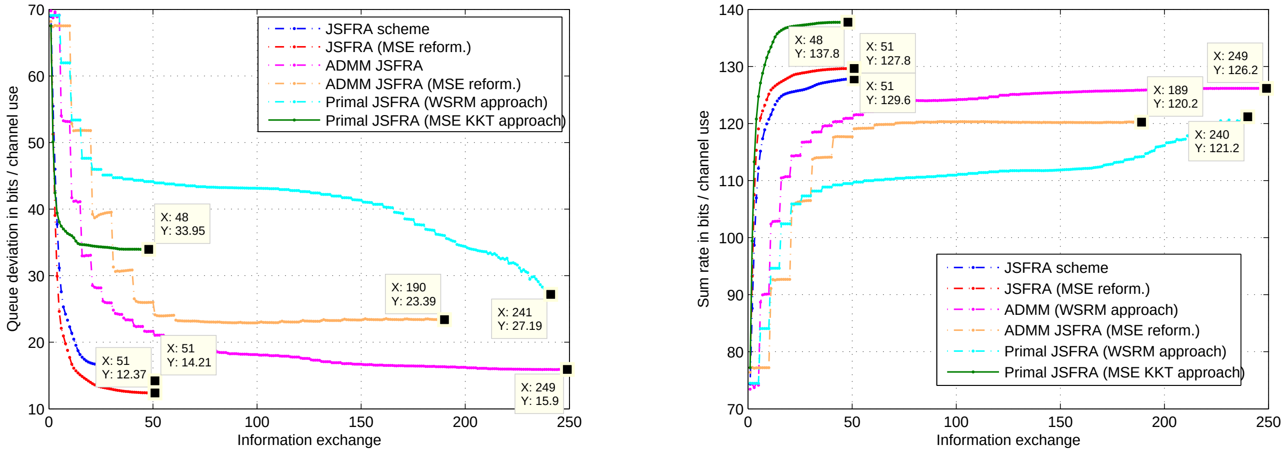
In Fig. 3, the total backlogged packets and the sum throughput offered to minimize the number of queued packets for $K = 12$

users across $N = 5$ sub-channels are plotted for the analysis. The system considers $N_B = 3$ BSs, each having $N_T = 4$ transmit antennas serving $|\mathcal{U}_b| = 4$ users mounted 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 independently. The sum rate and queue minimizing behavior follows similar pattern as in Fig. 2 for $N_B = 2$ BSs scenario.

Fig. 3 demonstrates the convergence behavior of the different distributed algorithms discussed so far. It can be seen from Fig. 3, that the distributed algorithms depend on the step size used in the subgradient update procedure. The primal decomposition algorithm, which distributes the precoder design by fixing the interference thresholds at each iteration, converges to the optimal resource allocations gradually as compared to the ADMM based dual approach. The ADMM scheme depends on the step size for the convergence, where the larger step size makes the algorithm to reach the optimal point quickly but takes more iterations to converge. The choice of $I_{\max}$, which denotes the maximum SCA iterations, and the $J_{\max}$, which denotes the subgradient convergence limit, are fixed at 100 and 8 respectively. The information exchange is made at each instant and the performances are evaluated as if the actual transmission is happened with the precoders $\mathbf{m}_{l,k,n}^{(i)}$ designed at the i^{th} instant in a distributed manner. In this paper, we assume that the actual transmission will happen at the end of the precoder convergence or when the maximum number of iterations reached.

The performance of the joint Q-WSRM scheme discussed in the Section III-A performs slightly worse than ($\approx$ 2 bits) the JSFRA scheme in reducing the number of queued packets, even though the performance were similar in Table I. As stated earlier, the joint Q-WSRM scheme and the JSFRA scheme with the exponent $q = 1$ performs the same when the number of queued packets are identical for all the users. In this case, the joint Q-WSRM becomes the sum rate maximization problem which is equivalent to the greedy $q = 1$ JSFRA scheme. In the current scenario, since the number of queued packets are different for the users, the performance of the JSFRA scheme is better due to its greedy resource allocation. It can be seen from Fig. 3, that the KKT based approach provides inferior result in the queue minimization perspective as compared with the sum rate plot. The performance degradation is due to the unconstrained rate increase for the users beyond their actual number of queued packets available to transmit. Since the objective uses the norm-1 minimization, the KKT solution is not optimal unless the queues are far greater than the transmission rate available for each user. The number of queued packets are initialized in a way to analyze the effect of the unconstrained behavior in the KKT scheme.

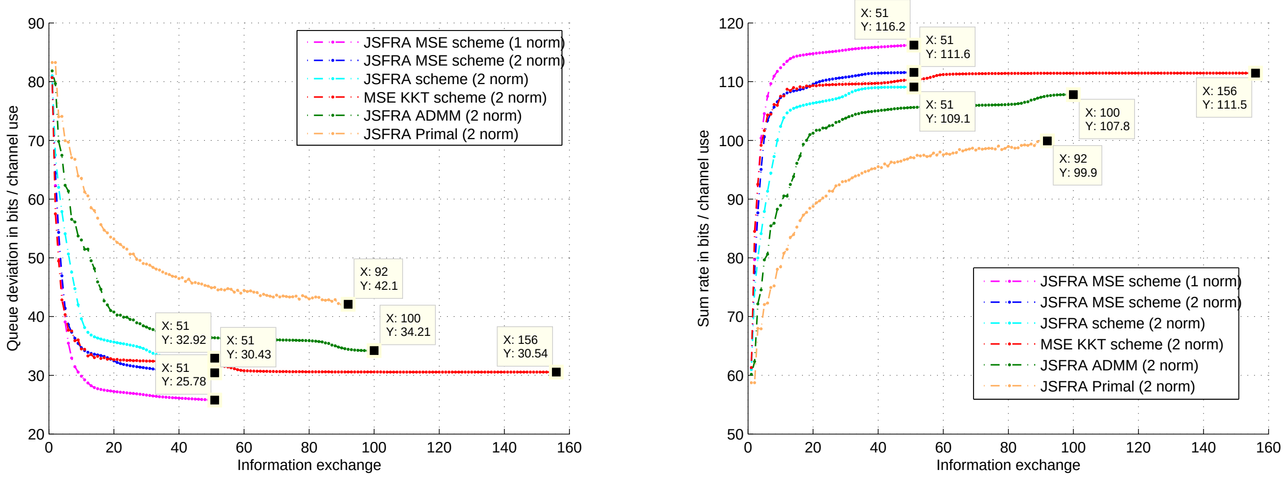
The performance of the MSE-KKT scheme is shown in comparison with the other schemes discussed so far in Fig. 4. It can be seen that the KKT based approach performs similar to the optimization problem given in (16), which is solved using the solver [10]. In the two norm case, the objective is differentiable and hence can be differentiable for any number of queued packets as compared to the one norm case requiring larger queue sizes in order to drop the absolute function occurring in the norm function. The performance of the closed form solution using the KKT solution achieves similar performance without allocating rates beyond the available queued packets. In order to bring out the performance difference, one norm objective plot is also provided in Fig. 4 providing better reduction in the number of queued packets in the current transmission slot. Fig. 4



(a) Queue deviation

(b) Sum rate performance

Fig. 3: Convergence plot for $\{N, N_B, K, N_R\} = \{5, 3, 12, 2\}$ model



(a) Queue deviation

(b) Sum rate performance

Fig. 4: Convergence plot for $\{N, N_B, K, N_R\} = \{6, 3, 12, 1\}$ model

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

KKT expressions for the Distributed MSE Formulation

In order to solve for an iterative precoder design algorithm, the KKT expressions for the problem in (36) are obtained by differentiating the Lagrangian by assuming the equality constraint for (36b) and (36c) as

$$\nabla_{t_{l,k,n}} : \quad -q\left[a_k\left(Q_k-\sum_{n=1}^{N}\sum_{l=1}^{L}t_{l,k,n}\right)^{(q-1)}\right]+\sigma_{l,k,n}\,\log(2) \quad =0 \tag{38a}$$

$$\nabla_{\epsilon_{l,k,n}} : \quad -\alpha_{l,k,n}+\frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} \quad =0 \tag{38b}$$

$$\nabla_{\mathbf{m}_{l,k,n}} : \sum_{y\in\mathcal{U}}\sum_{x=1}^{L}\alpha_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_k,y,n}\mathbf{w}_{x,y,n}\mathbf{w}^{\mathrm{H}}_{x,y,n}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n}+\delta_b\mathbf{m}_{l,k,n}=\alpha_{l,k,n}\mathbf{H}^{\mathrm{H}}_{b_k,k,n}\mathbf{w}_{l,k,n}, \tag{38c}$$

$$\nabla_{\mathbf{w}_{l,k,n}} : \sum_{(x,y)\neq(l,k)}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}^{\mathrm{H}}_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_y,k,n}\mathbf{w}_{l,k,n}+\mathbf{I}_{N_R}\mathbf{w}_{l,k,n} \quad =\mathbf{H}_{b_k,k,n}\;\mathbf{m}_{l,k,n} \tag{38d}$$

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

$$\alpha_{l,k,n}\underbrace{\left(\left|1-\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2+\sum_{(x,y)\neq(l,k)}\left|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_y,y,n}\mathbf{m}_{x,y,n}\right|^2+N_0\left\|\mathbf{w}_{l,k,n}\right\|^2-\epsilon_{l,k,n}\right)}_{=0}=0 \tag{39a}$$

$$\sigma_{l,k,n}\underbrace{\left(\log(\tilde{\epsilon}_{l,k,n})+\frac{(\epsilon_{l,k,n}-\tilde{\epsilon}_{l,k,n})}{\tilde{\epsilon}_{l,k,n}}+t_{l,k,n}\,\log(2)\right)}_{=0}=0 \tag{39b}$$

$$\delta_b\left(\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\mathrm{tr}\left(\mathbf{M}_{k,n}\mathbf{M}^{\mathrm{H}}_{k,n}\right)-P_{\max}\right)=0. \tag{39c}$$

In the expressions (39a) and (39b), the value inside the braces are $=0$ due to the equality constraints. Now, the dual variables corresponding to the equality constraints satisfies $\alpha_{l,k,n}\geq 0$ and $\sigma_{l,k,n}\geq 0$. The total power constraint in (39c) 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 (39c). Upon solving the KKT expressions in (38) and (39), we obtain the iterative algorithm defined in the Section IV-C.

References

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

[2] 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.

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

[4] 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.

[5] L. Tran, M. Hanif, A. Tolli, 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.

[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. Tolli, 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] 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.

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

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

[11] 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.

[12] J. Kaleva, A. Tolli, 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.

[13] H. Pennanen, A. Tolli, 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. Tolli, 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] 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.

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

[17] 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.

[18] 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.

[19] P. Komulainen, A. Tolli, 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.