I. Response to reviewer comments

A. Reveiwer Comments - 1

In this paper, the authors proposed a traffic aware resource allocation scheme for multi-cell MIMO-OFDM systems, where the precoders at all BSs are chosen to minimize the total user queue deviations. The problem is nonconvex and the authors proposed two centralized algorithms based on the successive approximation (SCA) technique to find a stationary point. Moreover, several distributed algorithms are also proposed using primal decomposition, alternating directions method of multipliers (ADMM), and decomposition via KKT conditions, respectively.

Most sections of this paper are well written. The results and algorithms also seem valid. However, the motivation of minimizing the total user queue deviations is not well justified. The convergence results of some algorithms are not clearly presented. The presentation of the distributed solutions needs significant improvement. Analysis and comparison of the signaling overhead and computational complexity between the centralized and distributed algorithms are also necessary to justify the advantages of distributed algorithms.

Response :We thank the reviewer for reading the paper and providing the comments. The response for the reviewer questions are made in line.

Detailed Comments

1 - **Comment** : In Section II.B, please provides more justifications for the problem formulation in (6). For example, the Queue weighted sum rate maximization (Q-WSRM) is throughput optimal, i.e., if there exists a scheme which can make all queues stable, then the Q-WSRM can also do this. How about the proposed formulation in (6)? Is it also throughput optimal?

Response :We fully agree with the reviewer comment. The Q-WSRM scheme is throughput optimal when the queues associated with the users are significantly large in comparison with the transmission rate (service rate). In order to restrict the over allocation of the available resources to a specific user beyond the total number of queued packets, we have included the additional rate constraint in the Q-WSRME algorithm. It would be ideal to compare with the Q-WSRME algorithm which performs similar to the Q-WSRM when the queued packets are large enough to be emptied by the current transmissions. In Section VI-C, we have discussed the performance of the proposed schemes with the existing queue weighted sum rate maximization (Q-WSRM) extended (Q-WSRME) using the average number of backlogged packets at each instant for different arrival rates. We have also compared using the instantaneous backlogged packets in the system for a arrival rate of $A_k = 5$ bits for each user.

2 - **Comment** : Do the proposed solutions based on (6) achieve better average delay performance than the existing solutions? By the way, in the simulations, you should also add a figure comparing the average delay performance, instead of just comparing the performance metric defined by (6). This will better justify the advantage of the proposed solutions.

Response :We agree with the reviewer comment on the better average delay performance of Q-WSRM(E) approach over JSFRA scheme with $q = 1$ formulation. Note that the average delay can be reduced by including a convex minimum rate constraint in the JSFRA formulation to provide a guaranteed rate to all users or for certain users in the system. More over,

the priority can also be incorporated easily in the formulation by the scaling factor a_k used in the formulation. As mentioned in Section III-B, the delay can also be addressed by using $q = 2$ or $q = \infty$ norm objective in the JSFRA formulation.

3 - **Comment** : In Section III.B, the convergence conditions under Algorithm 1 are not clear. First, you should be more specific about what is the SCA subproblem. Do you mean problem (19)? Second, does the uniqueness of the transmit and receive beamformers mean that the solution of the original problem in (16) is unique, or the solutions of the subproblems in (19) and (20) are unique, respectively?

Response :We understand the reviewer concerns. We have updated the convergence section to include additional information to provide more clarity. Appendix B discusses the convergence proof of the proposed algorithm in detail.

4 - **Comment** : It is better to clearly summarize the convergence conditions and results (i.e., does it converge to a stationary point or the optimal solution) for all algorithms in a theorem/proposition.

Response :We agree with the reviewers view. We have updated the convergence proof for both centralized and distributed algorithm to include the discussions on the stationary point and the locally optimal point in Appendix B and Section V-C.

5 - **Comment** : At the end of Section III, you mentioned that the proposed reduced complexity resource allocation scheme is sensitive to the order in which the subchannels are selected for the optimization problem. Please provide a discussion how to choose this order.

Response :We thank the reviewer for the raising the concern on the selection order. We have elaborated the discussion on the sub-channel wise selection scheme in Section IV-D.

6 - **Comment** : In the distributed algorithms, it is not clear what exact information is exchanged between the BSs or between the BSs and users. Moreover, the signaling overhead should be analyzed and compared with the centralized solution. The proposed distributed algorithms require exchanging over-the-air signaling or backhaul signaling for many times within each channel coherent time (e.g., from Fig. 2, the distributed algorithm requires 20-30 iterations to converge even when there are only 3 subchannels). I dont think this is acceptable in practice. Is the signaling overhead of the distributed algorithm really smaller than the centralized algorithm which only requires exchange the CSI between the BSs for once within each channel coherent time?

Response :We thank the reviewer for the insightful comment on the practicality of the distributed algorithm. We fully agree with the reviewer comment on the information exchange between the BSs in the distributed approach. Please note that the proposed distributed algorithm based on the primal or dual decomposition is provided for the completeness purpose. Note that the distributed approach is performed for the convex subproblem, which leads to the same stationary point asymptotically as that of the centralized solution. In reality, we have to limit the number of iterations required for each distributed algorithm, thereby leading to a point which is not the stationary point when the algorithm is allowed to converge. In Section V-D, we have discussed a practical approach based on the KKT conditions, which attains a local optimal point with few number of iterations.

7 - **Comment** : The convergence analysis of the distributed algorithms is not clear. For example, what is the exact condition to ensure the convergence of the distributed algorithms. Does the distributed algorithms also converge to a stationary point?

Response :We understand the reviewers concern. We have update the text on the convergence of the distributed algorithm

(ADMM). Please not that the ADMM or the primal decomposition algorithm is used for the convex subproblem only. If the distributed algorithm is iterated until convergence, it is guaranteed to attain the same stationary point as that of the centralized algorithm [1]. We have included this discussion in Section V-C.

8 - **Comment** : Im totally confused with the ADMM approach in Section IV.A. Many notations, such as the local interference vector and consensus interference vector are used without formal definition. What is the difference between the local interference vector and consensus interference vector? What are their relationships with the actual interference vector. It seems that you are using the same notation for all of these interference vectors and I cant tell when a notation refers to a local interference vector, a consensus interference vector, or the actual interference vector. These questions should be clarified and perhaps you should choose the notation system more carefully. For example, in (36), there are 3 similar notations and I dont know which one is local interference vector and which one is the actual interference vector.

Response :We understand the concern of the reviewer. We have updated the distributed section to include all the details pointed by the reviewer. Section V-B includes the updated discussions on ADMM scheme.

9 - **Comment** : In the distributed algorithms, it is not clear what information is available at each node. For example, what are your assumption on CSIT (CSI knowledge at each BS) and CSIR (CSI knowledge at each user)? How to 2 obtain the information used to perform the required calculation at each node (such as calculating the actual interference, MMSE receiver and the dual variables)?

Response :We understand the concern of the reviewer. We have updated the distributed section to include the details pointed by the reviewer. Section V-B includes the updated discussions on ADMM scheme. In addition, page 17, line 2 includes information about who knows what in the network.

10 - **Comment** : Do you have any convergence result for the proposed distributed solution based on the KKT conditions in Section IV.B? It seems that the iterative method to solve the KKT conditions is totally heuristic.

Response :We fully agree with the reviewer. It is a heuristic approach since we update the transmit precoder, receive beamformer and the dual variables all at each iteration. Please note that the proposed algorithm is of practical significance, since it has few number of iterations before the actual precoder design. It is surely not a stationary point of the original nonconvex problem but it is guaranteed to provide better performance in the sum rate compared to the distributed approaches presented in Section V-A and Section V-B for the same number of iteration. Note that, if the dual variables are allowed to iterate until convergence, the proposed KKT based scheme achieves the same stationary point for each fixed receive beamformer (similar to the distributed algorithms).

11 - **Comment** : Since queue is a dynamic system evolving according to (3), it doesnt make sense to compare the queue deviations at a given time. You should compare average queue deviations in the simulations. Moreover, you should also compare the average delay performance instead of just comparing the performance metric (queue deviations) defined in this paper. Using the queue deviations as the performance metric also needs more justification.

Response :We thank the reviewer for the insightful comment. We agree that the instantaneous deviation is not the right measure for the comparison. In the current work, we planned to discuss on the precoder design only and since the expectation is maximized by maximizing the function inside the expectation at each instant, we used the snapshot a given instant to compare

different algorithms. Please note that Fig. 4a includes the comparison over 50 slot duration and plot compares the average number of backlogged packets. We thought of presenting the results with the time-correlated fading channel and the precoder design with fixed number of iterations in the future publication. If the reviewer still insist this figure needs to be included, we will include.

12 - **Comment** : What is SRA in the simulation figures?

Response :We have updated the figure.

13 - **Comment** : In the discussion for Fig. 1, you mentioned that JSFRA converge to the optimal point, and all algorithms are Pareto-optimal. Since the problem is non-convex, why these algorithm can find optimal solution or Paretooptimal point?

Response :We thank the reviewer for the comment. We have updated the text to include pareto-suboptimal point.

Reveiewer comments - 2

1 - **Comment** : The logic from (6) to (16) is not clear. The only difference is the two newly introduced NON-CONVEX constraints (16b) and (16c), while the objective function (16a) and the constraint (16d) is the same as (6). The equivalence between (6) and (16) is not straightforward and it is confusing why the reformulation in (16) is beneficial.

Response :We thank the reviewer for the critical comment. We have updated the discussions on the equivalence in Section IV-A. Please note that the discussion on the Q-WSRM formulation is made for the completeness of the problem discussion. For more information on this topic, please refer to the Chapter 3 in [2] book.

2 - **Comment** : The authors use the successive convex approximation framework, but the approximate problem proposed by the authors is actually not convex. Inspecting (19), its objective function is the same as in (6), and the non-convexity of (6) comes exactly from the objective function, so (19) is not a convex problem. The same flaw is repeated several times in the approximate problems proposed by the authors.

Response :Please note that the objective function is a norm function and is convex. Note that the successive convex approximation (SCA) is used to decompose the original nonconvex problem due to the nonconvex constraint in (16b) into a series of convex subproblems that are solved in an iterative manner. We understand the reviewers concern and we have modified the text to bring out the details clearly.

3 - **Comment** : The authors proposed to use block coordinate descent method to solve (16). But as the authors have already pointed out, to apply block coordinate descent method, the constraint sets for different variables should be disjoint (uncoupled), which is however not the case in (16), because receive and transmit precoders (i.e w and m) are coupled in the constraints. It is confusing on its own why the authors made a statement that contradicts the proposed methodology, and the convergence followed is in question.

Response :We thank the reviewer for the pointing out the flaw in the text. We have modified the text with more details.

4 - **Comment** : Regarding the convergence of the SCA, the authors cited [27] for the convergence conditions, but the reference is wrong, because the conditions after the three bullets on page 6 are not mentioned in [27]. In case the authors disagree, please make the citation more specific, for example, specify the theorem/statement/proposition in [27] where those conditions are specified.

Response :We thank the reviewer for the pointing out the citation problem. We have included the text on convergence in Appendix B to address it in detail.

5 - **Comment** : The authors also cited [28] to establish the convergence of SCA. But the techniques of [27] and [28] are different, and the convergence conditions are different too. It is not clear why the authors need two set of convergence conditions for a single problem, and the resulting convergence analysis itself is not solid enough.

Response :We thank the reviewer for the pointing out the citation problem. We have included the text on convergence in Appendix B to address it in detail.

6 - **Comment** : Another comment on reference: to the reviewer's knowledge, the term SCA is never explicitly used in [2]. So please either correct the reference or be more specific (section, theorem, etc.).

Response :We thank the reviewer for the pointing out the citation problem. We have removed it in the revised manuscript.

7 - **Comment** : The authors propose primal decomposition method, ADMM approach to the non-convex problem (19), while their convergence analysis is based on literature that proved convergence for convex problems only, e.g., [13]. So the convergence analysis is not trustworthy.

Response :We apologize for the confusion in the text. We have updated the text accordingly to point out the convex problem for which the decomposition is performed.

8 - **Comment** : The length of the paper is too extensive. Some of the reformulations as mentioned in the previous comment can be skipped. Also, Section III.D. is not deeply explained and does not bring additional value to the paper. The implications of ordering the sub-channels for the iterative approach should be carefully studied and extensively explained in a different publication.

Response :We understand the reviewers concern. We have included the subchannel wise resource allocation or (SRA) in Section IV-D for the completeness. It is a variation to the centralized problem which is included for the completeness. If the reviewer still insists on its removal, we will do.

9 - **Comment** : Information regarding the value of q used to obtain the simulation results is missing (with exception of Fig. 3).

Response :We thank the reviewer for pointing out the important point. We have included the norm used in Figures in the caption.

10 - **Comment** : In Fig. 1 and Fig. 2, the labels for the system model do not fit with the written description. Additionally, the reference scheme Q-WSRM is not optimal, since it over allocates resources if there are few queued packets. Therefore, it is not interesting for comparison purposes.

Response :We understand the reviewers concern. It is included for the comparison purpose only. If reviewer insists on removing, we will do.

11 - **Comment** : Assuming that Fig. 2 and Fig. 3 where obtained based on the same simulation setup, i.e. user queues, number of transmit and receive antennas and number of base stations, it is not clear why results in Fig. 3 are worse than Fig. 2 when comparing JSFRA. Even more, since the number of sub-channels is larger in Fig. 3, the result seems contradictory.

Response :Please note that the number of sub-channels and the number of base stations (BSs) are different in addition to

the number of backlogged packets. If they are same, then both schemes should provide same performance, since the distributed algorithm is performed over the convex subproblem after linearization.

Reveiewer comments - 3

1 - **Comment** : This manuscript focuses on the beamforming and scheduling optimization for IBC MIMO-OFDM system, including the centralized and decentralized optimization methods. This is an interesting and important topic.

Response :We thank the reviewer for reading the manuscript and providing valuable comments.

2 - **Comment** : The number of transmitted packets t_k's are optimization variables, which should be explicitly stated in the problem formulation of (6), (16), (19), (20) and (26) to avoid confusing.

Response :Please note that the objective function uses $v_k = Q_k - t_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L}\log_2(1+\gamma_{l,k,n})$ expression instead of including an additional constraint for the transmitted packets using the rate expression. It is necessary in case of the MSE formulation, where we have explicitly stated the optimization variable t_k. If the reviewer insist on replacing the rate expression using t_k, we can make those changes.

3 - **Comment** : The manuscript states that the inequalities (16b) and (16c) achieve equality at optimality(line 23, page 5). This is not obvious. An easy case to check this statement is that assuming the system has two BS and each BS serves one user. When $Q_1 = 0$ and 2nd BS has sufficiently large power, (16b) and (16c) do not hold equality. Rigorous proof is needed if authors stick to this statement.

Response :We thank the reviewer for the insightful comment. We modified the statement such that it is an under estimator for the actual SINR expression and will be tight when the system is limited by the transmit power.

4 - **Comment** : The solution in (21) is obtianed for MMSE, i.e. for 2-norm(q=2). If $q = 1$ or $q = \infty$, it is actually an equivalent linear programming problem. Details for this solution should be provided.

Response :We thank the reviewer for the critical comment. We have update the manuscript to include the information about the receiver with other q values. Please note that the receive beamformer doesn't depend on the q value explicitly. It depends inherently through the transmit precoders.

5 - **Comment** : The convergence proof need to be rigorous. The inequality of (23a) is opposite to the reference [28]. Also the statement on uniqueness of the transmit and the receive beamformers are not correct. Although we can choose one antenna to be real value, this does not mean the problem has unique solution!

6 - **Comment** : (25) is generally wrong. (25) only holds when the MSE is minimized(by MMSE receiver) and the snr is the optimized(which is obtained by general eignvalue decompostion). This is clearly stated in the reference [5] and [6]. This can also be easily checked by comparing (25) and (2). Consequently the alternative formulation (26) based on this conclusion is questionable.

Response :We thank the reviewer for pointing out the information missed out in the statement. We have provided more detail in the manuscript. The relation is valid if the receiver is based on the MMSE criterion. Note that, the receiver does not depend on the q value used in the objective. It depend only on the transmit precoders which in fact change based on the q value used in the objective function.

7 - **Comment** : For ADMM aproach, the determination of the value of ρ in equation (35a) should be discussed. 1. The numbers of transmitted packets for users t_k's are optimization variables. So they should be explicitly stated in the problem formulation (6), (16), (20) and (26) to avoid confusing.

Reveiewer comments - 4

1 - **Comment** : First, this reviewer is not convinced by the arguments for showing the convergence of the JSFRA method. "The SCA method" is often referred to, but never really defined or referenced. The three required conditions (as stated on p. 6, col. 1, rows 38-40) do not, as far as I can tell, appear in [27]. Indeed, [27] is concerned with optimization problems where the objective function is non-convex, but the constraint set is convex and separable over the blocks of variables. Perhaps you meant to cite [A], wherein non-convex constraints are handled in a similar way? Numerically, the algorithms do converge, and the argument put forward makes sense, but the treatment must be improved to be more rigorous.

Response :We thank the reviewer for We have updated the proof for convergence in a separate section on Appendix B.

2 - **Comment** : Second, the optimization problems formulated only depend on Q_k, the current levels of backlogged packets, and not on the arrival rates. This is due to how the conditional Lyapunov drift is minimized. This approach completely removes the queue dynamics from the optimization problem, essentially leading to greedy one-shot optimization in 7 every time instant. The framework would be more interesting if some sort of optimization (or tracking) is performed over several time instants, rather than the one-shot approach that is currently used for the JSFRA algorithms. Possibly, some expectation over the queues would be optimized then. Even if no analytical treatment of the tracking over several time-steps is added, I would at least highly recommend adding some simulation results where the proposed one-shot algorithms are performed sequentially over several time instants.

Response :We thank the reviewer for pointing out the missing information. We have included a figure on the number of backlogged packets over a period of time in Section VI-C.

3 - **Comment** : Third, the distributed methods (at least the primal decomposition and ADMM) seem to be fairly straightforward applications of existing results. This reviewer recommends spending more space on the convergence, than on the description of the distributed techniques. Still, it would be nice with a direct description of what local CSI is required, and how it is acquired, to perform the local computations for the primal decomposition and ADMM methods. For the description of the signaling of the CSI in Sec. IV-B, are you envisioning a TDD system?

Response :We agree with the reviewer. We have updated the manuscript with more details.

4 - **Comment** : Finally, some readers might be confused by the "joint space-frequency" terminology, believing that the beamforming is performed over a joint space-frequency channel space, where the space-frequency channels are formed by block-diagonal matrices, each block belonging to one subcarrier. This could easily be clarified.

Response :We thank the reviewer for pointing out the misinterpretation of the text. We have included the statement accordingly in the introduction.

5a - **Comment** : - p. 1, col. 1, row 42: "userss"

Response :Modified.

5b - **Comment** : - p. 1, col. 2, row 18: the precoders are used _implicitly_ as decision variables. This is the whole point, to avoid explicitly modeling the hard decisions in the optimization, and instead do soft decisions during the iterations, and then finally hard decisions after convergence.

Response :We thank the reviewer for the insightful comment on the proposed approach. We have modified the text accordingly.

5c - **Comment** : - p. 1, col. 2, row 33: Which chapter in [2] is referred to? With a quick look-through of the table of contents, I can't find and chapter or section treating the SCA method?

Response :We understand the reviewer's concern. We have provided the valid reference for the SCA scheme.

5d - **Comment** : - p. 2, col. 2, row 36: Write rank$(.)$ and $\min$ instead

Response :Changed.

5e - **Comment** : - p. 3, col. 1, row 26: It would be more clear to explicitly write out the dependence of $\mathbf{M}$ and $\mathbf{W}$ in $\tilde{v}$ here

Response :We thank the reviewer for the concern on the grouped variable. We have modified the expressions accordingly.

5f - **Comment** : - p. 3, col. 2, row 26: Which general MIMO-OFDM problem are you talking about here, and what is combinatorial about it? Is it the problem of selecting users to be served on orthogonal subcarriers? There is nothing inherently combinatorial over the problem in (6) as far as I can tell, as the beamformers are used as soft decision variables.

Response :We thank the reviewer for the comment. We have removed the word "combinatorial" from the text.

5g - **Comment** : - p. 4, col. 2, row 40: "In fact, (5) provides similar expression of ..." This sentence is very hard to understand.

Response :It is restructured to include additional details.

5h - **Comment** : - (16d): suggest your write out the power constraints here, in order to be faster be able to interpret the optimization problem. There is hardly any spaced saved by referring back to (6b).

Response :Agreed.

5i - **Comment** : - p. 5, col. 1, rows 27-30: Here you might want to quickly mentioned how one could show the NP-hardness of (16).

Response :We have updated the NP hardness discussion in the revised version.

5j - **Comment** : - p. 5, col. 1, row 50: "According to the SCA method...". I am not sure exactly how you define "_the_ SCA method"? Clarify or cite the definition.

Response :We thank the reviewer for pointing out the mistake. We have modified the text accordingly.

5k - **Comment** : - p. 5, col. 2, row 31: Here is a case where it makes sense to reference earlier optimization constraints. However, are (19d) and (18) not the same??

Response :We thank the reviewer for pointing out the mistake. We have modified accordingly to the comment.

5l - **Comment** : - p. 5, col. 2, row 51: Slightly confusing with the notation between the iterates in (21b) and the MMSE filter in (22b).

Response :We have revised the manuscript to make it clear.

5m - **Comment** : - p. 6, col. 1, row 9: You might want to add somewhere that (22b) can be used instead of the fixed-point of (21b), since the scaling of the receive filters do not matter in the SINRs. However, does it affect the convergence of the algorithm?

Response :We thank the reviewer for the comment. We have revised the manuscript accordingly.

5n - **Comment** : - p. 6, col. 2, rows 8-10: I don't fully understand the reasoning on the relation between the constraint sets in the different iterations. Why is this the case?

Response :We thank the reviewer for raising the concern on the convergence, we have updated the convergence proof in Appendix B to include much detail for clear understanding.

5o - **Comment** : - p. 7, col. 1, row 35: Just because a problem is convex does not mean that it has a unique solution. (Although it seems to me that (26) should have a unique solution.) Is the problem in (26) strictly convex?

Response :We thank the reviewer for informing the important argument. The relaxed sub problem is strongly convex, due to the presence of norm function in the objective. Since the relaxation forced $q_{l,k,n} = 0$, the precoders are unique.

5p - **Comment** : - Table 1: "backpreassure"

Response :Changed in the revised manuscript

5q - **Comment** : - p. 11, col. 1, row 56: "performances". I'm not sure this is a countable noun.

Response :Changed in the revised manuscript

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

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

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

***Index Terms*—Convex approximations, MIMO-IBC, MIMO-OFDM, Precoder design, SCA, WSRM.**

II. Introduction

In a network with multiple base stations (BSs) serving multiple-users (MUs), the main driving factor for the transmission are the packets waiting at each BS corresponding to the different users present in the network. These available packets are transmitted over the shared wireless resources subject to certain system limitations and constraints. We consider the problem of transmit precoder design over the space-frequency resources provided by the multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) framework in the downlink interference broadcast channel (IBC) to minimize the number of queued packets. Since the space-frequency resources are shared by multiple users associated with different BSs, it can be viewed as a resource allocation problem.

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

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

In general, the resource allocation problems are formulated by assigning a binary variable for each user to indicate the presence or the absence in a particular resource [3]. *In contrast, the linear transmit precoders, which are complex vectors, are implicitly used as decision variables, thereby avoiding the explicit modeling of binary decision variables. It is used to determine the transmission rate of a user on a space-frequency resource if the precoder vector is non-zero. A zero transmit precoder indicates the absence of the user on a given resource. In this way, the soft decisions are used in the optimization problem and the hard decisions are made after the algorithm convergence*.

The queue minimizing precoder designs are closely related to the weighted sum rate maximization (WSRM) problem with additional rate constraints determined by the number of backlogged packets for each user in the system. The topics on MIMO IBC precoder design have been studied extensively with different performance criteria in the literature. Due to the nonconvex nature of the MIMO IBC precoder design problems, the successive convex approximation (SCA) method has become a powerful tool to deal with these problems [4]. For example, in [5], the nonconvex part of the objective has been linearized around an operating point in order to solve the WSRM problem in an iterative manner. Similar approach of solving the WSRM problem by using arithmetic-geometric inequality has been proposed in [6].

The relation between the achievable capacity and the mean squared error (MSE) of the received symbol by using fixed minimum mean squared error (MMSE) receivers can be used to solve the WSRM problem [7]. In [8], [9], the WSRM problem is reformulated via MSE, casting the problem as a convex one for fixed linearization coefficients. In this way, the original problem is expressed in terms of the MSE weight, precoders, and decoders. Then the problem is solved using an alternating optimization method, i.e., finding a subset of variables while the remaining others are fixed. The MSE reformulation for the WSRM problem has also been studied in [10] by using the SCA to solve the problem in an iterative manner. Additional rate constraints based on the quality of service (QoS) requirements were included in the WSRM problem and solved via MSE reformulation in [11], [12].

The problem of precoder design for the MIMO IBC system are solved either by using a centralized controller or by using decentralized algorithms where each BS handles the corresponding subproblem independently with the limited information exchange with the other BSs via back-haul. The distributed approaches are based on primal, dual or alternating

directions method of multipliers (ADMM) decomposition, which has been discussed in [1], [13]. In the primal decomposition, the so-called coupling interference variables are fixed for the subproblem at each BS to find the optimal precoders. The fixed interference are then updated by using the subgradient method as discussed in [14]. The dual and ADMM approaches control the distributed subproblems by fixing the *'interference price'* for each BS as detailed in [15].

By adjusting the weights in the WSRM objective properly, we can find an arbitrary rate-tuple in the rate region that maximizes the suitable objective measures. For example, if the weight of each user is set to be inversely proportional to its average data rate, the corresponding problem guarantees fairness on an average among the users. As an approximation, we may assign weights based on the current queue size of the users. More specifically, the queue states can be incorporated to traditional weighted sum rate objective $\sum_k w_k R_k$ by replacing the weight w_k with the corresponding queue state Q_k or its function, which is the outcome of minimizing the Lyapunov drift between the current and the future queue states [2], [16]. In backpressure algorithm, the differential queues between the source and the destination nodes are used as the weights scaling the transmission rate [17].

Earlier studies on the queue minimization problem were summarized in the survey paper [18], [19]. In particular, the problem of power allocation to minimize the number of backlogged packets was considered in [20] using geometric programming. Since the problem addressed in [20] assumed single antenna transmitters and receivers, the queue minimizing problem reduces to the optimal power allocation problem. In the context of wireless networks, the backpressure algorithm mentioned above was extended in [21] by formulating the corresponding user queues as the weights in the WSRM problem. Recently, the precoder design for the video transmission over MIMO system is considered in [22]. In this design, the MU-MIMO precoders are designed by the MSE reformulation as in [8] with the higher layer performance objective such as playback interruptions and buffer overflow probabilities.

Main Contributions: In this paper, we design the precoders jointly across space-frequency resources by minimizing the total number of backlogged packets waiting at the BSs. The proposed design provides better control over the resource allocation strategy by the change of a variable. Since the transmissions are guided by the backlogged packets, the proposed formulation limits the resource allocation beyond the number of backlogged packets without additional rate constraint. Since the problem is nonconvex due to difference of convex (DC) constraints, we adopt SCA to solve by a sequence of convex subproblems using first order approximations. Initially, we propose centralized joint space-frequency resource allocation (JSFRA) algorithms, which employs SCA for the nonconvex DC constraint. First method is by using the direct formulation and the second one is by using the MSE equivalence with the rate expression to solve for an optimal precoders. Then we propose distributed precoder designs based on the primal and the ADMM methods. Finally, we propose a iterative practical algorithm to decouple the precoder design across the coordinating BSs with limited information exchange by solving the Karush-Kuhn-Tucker (KKT) conditions for the MSE reformulation solution. *It is worth noting that the joint space-frequency channel matrix can be formed by stacking the channel of each sub-channel in a block-diagonal form for all users*.

The paper is organized as follows. In Section III, we introduce the system model and the problem formulation for the queue minimizing precoder design. The existing and the proposed centralized precoder designs are presented in Section IV. The distributed solutions are provided in Section V followed by the simulation results in Section VI. Conclusions are drawn in Section VII.

III. System Model and Problem Formulation

A. System Model

We consider a downlink MIMO IBC scenario in an OFDM framework with N sub-channels and N_B BSs each equipped with N_T transmit antennas, serving in total K users each with N_R receive antennas. The set of users associated with BS b is denoted by $\mathcal{U}_b$ and the set $\mathcal{U}$ represents all users in the system, i.e., $\mathcal{U} = \underset{b\in\mathcal{B}}{\cup}\mathcal{U}_b$, where $\mathcal{B}$ is the set of indices of all coordinating BSs. Data for user k is transmitted from only one BS which is denoted by $b_k \in \mathcal{B}$. We denote by $\mathcal{N} = \{1, 2, \ldots, N\}$ the set of all sub-channel indices available in the system.

We adopt linear transmit beamforming technique at BSs. Specifically, the data symbols $d_{l,k,n}$ for user k on the l^{th} spatial stream over the sub-channel n is multiplied with beamformer $\mathbf{m}_{l,k,n} \in \mathbb{C}^{N_T\times 1}$ before being transmitted. In order to detect multiple spatial streams at the user terminal, receive beamforming vector $\mathbf{w}_{l,k,n}$ is employed for each user. Consequently, the received data symbol estimate corresponding to the l^{th} spatial stream over sub-channel n at user k is given by

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

where $\mathbf{H}_{b,k,n} \in \mathbb{C}^{N_R\times N_T}$ is the channel between BS b and user k on sub-channel n, and $\mathbf{n}_{k,n} \sim \mathcal{CN}(0, N_0)$ is the additive noise vector for the user k on the n^{th} sub-channel and l^{th} spatial stream. In (1), $L = \text{rank}(\mathbf{H}_{b,k,n}) = \min(N_T, N_R)$ is the maximum number of spatial streams[1]. Assuming independent detection of data streams, we can write the signal-to-interference-plus-noise ratio (SINR) as

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

where $\widetilde{N}_0 = N_0\,\text{tr}(\mathbf{w}_{l,k,n}\mathbf{w}_{l,k,n}^{\mathrm{H}})$ denotes the equivalent noise variance. *To reduce the overhead involved in feeding back the*

[1]It can be easily extended for user specific streams L_k instead of using common L streams for all users. L streams are initialized but after solving the problem, only $L_{k,n} \leq L$ non-zero data streams are transmitted

user channels, we consider a time division duplexing (TDD) system, which uses channel reciprocity.

Let Q_k be the number of backlogged packets destined for the user k at a given scheduling instant. The queue dynamics of the user k are modeled using the Poisson arrival process with the average number of packet arrivals of $A_k = \mathbf{E}_i\{\lambda_k\}$ packets/bits, where $\lambda_k(i) \sim \text{Pois}(A_k)$ represents the instantaneous number of packets arriving for the user k at the i^{th} time instant[2]. The total number of queued packets at the $(i+1)^{\text{th}}$ instant for the user k, denoted as $Q_k(i+1)$, is given by

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

where $[x]^+ \equiv \max\{x, 0\}$ and t_k denotes the number of transmitted packets or bits for user k. At the i^{th} instant, transmission rate of the user k is given by

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

where $t_{l,k,n}$ denotes the number of transmitted packets or bits over l^{th} spatial stream on the n^{th} sub-channel. The maximum rate achieved over the (l,n) space-frequency resource is given by $t_{l,k,n} \leq \log_2(1+\gamma_{l,k,n})$ for the signal-to-interference-plus-noise ratio (SINR) of $\gamma_{l,k,n}$[3]. Note that the units of t_k and Q_k are in bits defined per channel use.

B. Problem Formulation

To minimize the total number of backlogged packets, we consider minimizing the weighted ℓ_q-norm of the queue deviation given by

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

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

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

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

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

where $\tilde{v}_k \triangleq a_k^{1/q}v_k$ is the element of vector $\tilde{\mathbf{v}}$, and a_k is the weighting factor which is incorporated to control user priority based on their respective QoS. In (6b), BS specific sum power constraint for all sub-channels is considered.

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

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

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

Before proceeding further, we show that the constraint in (7) is handled implicitly by the definition of norm ℓ_q in the objective of (6). Suppose that $t_k > Q_k$ for certain k at the optimum, i.e., $-v_k = t_k - Q_k > 0$. Then there exists $\delta_k > 0$ such that $-v_k' = t_k' - Q_k < -v_k$ where $t_k' = t_k - \delta_k$. Since $\|\tilde{\mathbf{v}}\|_q = \||\tilde{\mathbf{v}}|\|_q = \||-\tilde{\mathbf{v}}|\|_q$, this means that the newly created vector $\mathbf{t}'$ achieves a smaller objective which contradicts with the fact that the optimal solution has been obtained. The choice of the norm ℓ_q used in the objective function [18], [20] alters the priorities for the queue deviation function as follows

- ℓ_1 results in greedy allocation *i.e*, emptying the queue of users with good channel conditions before considering the users with worse channel conditions. As a special case, it is easy to see that (6) reduces to the WSRM problem when the queue size is large enough for all users.
- ℓ_2 prioritizes users with higher number of queued packets before considering the users with a smaller number of backlogged packets. For example, it could be more ideal for the delay limited scenario when the packet arrival rates of the users are similar, since the number of backlogged packets is proportional to the delay in the transmission following the Little's law [2].
- ℓ_∞ minimizes the maximum number of queued packets among users with the current transmission, thereby providing queue fairness by allocating the resources proportional to the number of backlogged packets.

IV. Proposed Queue Minimizing Precoder Designs

In general, the precoder design for the MIMO OFDM problem is difficult due to its nonconvex nature. In addition, the objective of minimizing the number of the queued packets over space-frequency dimensions adds further complexity. Since the scheduling of users in each sub-channel attained by allocating zero transmit power over certain sub-channels, our solutions perform joint precoder design and user scheduling. Before discussing the proposed solutions, we consider the existing algorithm to minimize the number of backlogged packets with additional constraints required by the problem.

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

The queue minimizing algorithms are discussed extensively in the networking literature to provide congestion-free routing between any two nodes in the network. One such algorithm is the *backpressure algorithm* [2], [16], [17]. It determines an optimal control policy in the form of rate or resource allocation for the nodes in the network by considering the differential backlogged packets between the source and the destination nodes. Even though the algorithm is primarily designed for the wired infrastructure, it can be extended to the wireless

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

[3]Upper bound is achieved by using Gaussian signaling

networks by designing the user rate variable t_k in accordance to the wireless network.

The queue weighted sum rate maximization (Q-WSRM) formulation extends the *backpressure algorithm* to the downlink MIMO-OFDM framework, in which the multiple BSs act as the source nodes and the user terminals as the receiver nodes. The control policy in the form of transmit precoders aims at minimizing the number of queued packets waiting in the BSs. In order to find the optimal strategy, we resort to the Lyapunov theory, which is predominantly used in the control theory to achieve system stability. Since at each time slot, the system is described by the channel conditions and the number of backlogged packets of each user, the Lyapunov function is used to provide a scalar measure, which grows large when the system moves toward the undesirable state. By following [2], the scalar measure for the queue stability is given by

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

where $\mathbf{Q}(i) = [Q_1(i), Q_2(i), \ldots, Q_K(i)]^{\mathrm{T}}$ and $\frac{1}{2}$ is used for the convenience. It provides a scalar measure of congestion present in the system [2, Ch. 3].

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

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

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

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

by using the following inequality

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

The total number of backlogged packets at any given instant i is reduced by minimizing the conditional expectation of the Lyapunov drift expression (10) given the current number of queued packets $\boldsymbol{Q}(i)$ waiting in the system. The expectation is taken over all possible arrival and transmission rates of the users to obtain the optimal rate allocation strategy.

Now, the conditional Lyapunov drift, denoted by $\Delta(\boldsymbol{Q}(i))$, is given by the infimum over the transmission rate as

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

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

where the subscripts $\mathbf{t}$ and $\boldsymbol{\lambda}$ represents the vector formed by stacking the transmission and the arrival rate of all users in the system. Since the transmission and the arrival rates are bounded, the second order moments in the first term of (12b) can be bounded by a constant B without affecting the optimal solution of the problem [2]. The second term in (12b) follows from the Poisson arrival process.

The expression in (12) looks similar to the WSRM formulation if the weights in the WSRM problem are replaced by the number of backlogged packets corresponding to the users. The above discussed approach is extended for the wireless networks in [21], in which the queues are used as weights in the WSRM formulation to determine the transmit precoders. Since the expectation is minimized by minimizing the function inside, the Q-WSRM formulation is given by

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

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

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

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

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

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

B. JSFRA Scheme via SCA approach

The problem defined in (13) ignores the second order term arising from the Lyapunov drift minimization objective by limiting it to a constant value. *In fact, using $\ell_{q=2}$ in (5), we obtain the following objective*

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

which is similar to the objective in (13). Note that the equivalence can be seen either by removing t_k^2 from (15) or when the total number of queued packets is significantly large for all users such that t_k^2 has no impact on the objective function.

By limiting t_k^2 with a constant value, the Q-WSRM formulation requires an explicit rate constraint (14) to avoid over-

allocation of the available resources. In the proposed queue deviation formulation, the explicit rate constraint is not needed, since it is handled by the objective function (5) itself. It makes the problem simpler and allows us to employ efficient algorithms to distribute the precoder design problem across each BSs independently by exchanging minimal information exchange [1]. In contrast to the WSRM formulation, the JSFRA and the Q-WSRME problems include the sub-channels jointly to obtain an efficient allocation by identifying the optimal space-frequency resource for the users.

We now present an iterative algorithm to solve problem (6) by using alternating optimization technique in conjunction with successive convex approximation (SCA) [23]. The problem is to determine the transmit precoders $\mathbf{m}_{l,k,n}$ and the receive beamformers $\mathbf{w}_{l,k,n}$ to minimize the total number of backlogged packets in the system. Since the SINR expression in (2) cannot be handled directly to formulate the problem, we relax the equality constraint in (2) by the inequality constraints to solve for the transmit and the receive precoders as

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

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

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

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

The SINR expression in (2) is relaxed by the inequalities (16b) and (16c) in the above formulation. Note that (16b) is an under estimator for SINR $\gamma_{l,k,n}$, and (16c) provides an upper bound for the total interference seen by user $k \in \mathcal{U}_b$, denoted by the variable $\beta_{l,k,n}$. The constraints are tight when each BS objective is non zero at the optimal point, as discussed in Appendix A. Since the JSFRA formulation can be modeled as a WSRM problem, which is known to be NP-hard [24], it also belongs to the class of NP-hard problems.

In order to find a tractable solution for (16), we note that (16d) is the only convex constraint with the involved variables. Thus, we only need to deal with (16b) and (16c). We resort to the alternating optimization (AO) technique by fixing the linear receivers, to solve for the transmit beamformers. For a fixed receivers $\mathbf{w}_{l,k,n}$, the problem now is to find the optimal transmit beamformers $\mathbf{m}_{l,k,n}$ which is still a challenging task. We note that for a fixed $\mathbf{w}_{l,k,n}$, (16c) can be written as a second-order cone (SOC) constraint. Thus, the difficulty is due to the non-convexity of the DC constraint in (16b). Let us define a function,

$$f(\mathbf{u}_{l,k,n}) \triangleq \frac{|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{l,k,n}\mathbf{m}_{l,k,n}|^2}{\beta_{l,k,n}}, \tag{17}$$

where $\mathbf{u}_{l,k,n} \triangleq \{\mathbf{w}_{l,k,n}, \mathbf{m}_{l,k,n}, \beta_{l,k,n}\}$ is the vector which needs to be identified for the optimal allocation. Note that the function $f(\mathbf{u}_{l,k,n})$ is convex for a fixed $\mathbf{w}_{l,k,n}$, since it is in fact the ratio between a quadratic form of $\mathbf{m}_{l,k,n}$ over an affine function of $\beta_{l,k,n}$ [25]. *The nonconvex set defined by the DC constraint (16b) can be decomposed by a series of convex subsets by linearizing the convex function $f(\mathbf{u}_{l,k,n})$ with its first order Taylor approximation around a fixed operating point $\tilde{\mathbf{u}}_{l,k,n}$ [26], [27], also called as SCA method in [23]. By using the reduced convex subset for (16b), the problem defined in (16) can be solved at each operating point iteratively.*

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

$$p_{l,k,n} \triangleq \Re\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{18a}$$

$$q_{l,k,n} \triangleq \Im\left\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right\} \tag{18b}$$

and hence $f(\mathbf{u}_{l,k,n}) = (p_{l,k,n}^2 + q_{l,k,n}^2)/\beta_{l,k,n}$[4]. 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 first order Taylor approximation around the local point $[\tilde{p}_{l,k,n}, \tilde{q}_{l,k,n}, \tilde{\beta}_{l,k,n}]^T$, we can approximate (16b) by the following linear inequality

$$2\frac{\tilde{p}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(p_{l,k,n} - \tilde{p}_{l,k,n}) + 2\frac{\tilde{q}_{l,k,n}}{\tilde{\beta}_{l,k,n}}(q_{l,k,n} - \tilde{q}_{l,k,n}) + \frac{\tilde{p}_{l,k,n}^2 + \tilde{q}_{l,k,n}^2}{\tilde{\beta}_{l,k,n}}\left(1 - \frac{\beta_{l,k,n} - \tilde{\beta}_{l,k,n}}{\tilde{\beta}_{l,k,n}}\right) \geq \gamma_{l,k,n}. \tag{19}$$

In summary, for the fixed linear receivers $\mathbf{w}_{l,k,n}$ and the operating point $[\tilde{p}_{l,k,n}, \tilde{q}_{l,k,n}, \tilde{\beta}_{l,k,n}]^T$, the relaxed convex subproblem to find transmit beamformers is given by

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

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

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}, \forall b, \tag{20c}$$

$$\text{and (19).} \tag{20d}$$

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

$$\underset{\substack{\gamma_{l,k,n},\\ \mathbf{w}_{l,k,n},\beta_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}\|_q \tag{21a}$$

$$\text{subject to} \quad \beta_{l,k,n} \geq \widetilde{N}_0 + \sum_{(j,i)\neq(l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}|^2 \tag{21b}$$

$$\text{and (19).} \tag{21c}$$

Solving (21) using the KKT conditions, we obtain the following iterative expression for the receiver $\mathbf{w}_{l,k,n}^*$ as

$$\mathbf{A}_{l,k,n} = \sum_{(j,i)\neq(l,k)} \mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\mathbf{m}_{j,i,n}^{\mathrm{H}}\mathbf{H}_{b_i,k,n}^{\mathrm{H}} + N_0\,\mathbf{I}_{N_R} \tag{22a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \left(\frac{\tilde{\beta}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)}}{\|\mathbf{w}_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\|^2}\right)\mathbf{A}_{l,k,n}^{-1}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}, \tag{22b}$$

where $\mathbf{w}_{l,k,n}^{(i-1)}$ is the receive beamformer from the previous iteration, upon which the linear relaxation is performed for the

[4]Note that $p_{l,k,n}$ and $q_{l,k,n}$ are just symbolic notation and not the newly introduced optimization variables. In CVX [28], for example, we declare $p_{l,k,n}$ and $q_{l,k,n}$ with the '*expression*' qualifier

nonconvex constraint in (21). *The optimal receiver $\mathbf{w}^{*}_{l,k,n}$ is obtained by either iterating (22b) until convergence or for fixed number of iterations. Note that the receiver has no explicit relation with the choice of ℓ_q norm used in the objective function. The dependency is implicitly implied by the transmit precoders $\mathbf{m}_{l,k,n}$, which in deed depend on the q value.*

It can be seen that the optimal receiver in (22b) is in fact a scaled version of the MMSE receiver, which is given by

$$\mathbf{R}_{l,k,n} = \sum_{i\in\mathcal{U}}\sum_{j=1}^{L} \mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\mathbf{m}^{\mathrm{H}}_{j,i,n}\mathbf{H}^{\mathrm{H}}_{b_i,k,n} + N_0\,\mathbf{I}_{N_R} \quad (23a)$$

$$\mathbf{w}_{l,k,n} = \mathbf{R}^{-1}_{l,k,n}\,\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}. \quad (23b)$$

Since the scaling present in the optimal receiver (22b) has no impact on the received SINRs, the MMSE receiver in (23b) can also be used without compromising the performance or the convergence behavior.

The proposed subproblems in (20) and (21) are solved in an iterative manner by updating the operating point from the previous iteration. The iterative algorithm is referred to as queue minimizing JSFRA scheme with a per BS power constraint, and it is outlined in Algorithm 1. The iterative procedure repeats until the improvement on the objective is less than a predetermined tolerance parameter or the maximum number of iterations is reached. Instead of initializing $\mathbf{u}_{l,k,n}$ arbitrarily to a feasible point, transmit precoders can also be initialized with some feasible point $\tilde{\mathbf{m}}_{l,k,n}$, which is then used to find $\mathbf{u}_{l,k,n}$ as briefed in Algorithm 1. For a fixed receive beamformer $\mathbf{w}_{l,k,n}$, the SCA iteration is carried out until convergence or for the predefined iterations, say, $J_{\max}$ for the optimal transmit precoders $\mathbf{m}_{l,k,n}$. Next, the receive beamformers are updated based on either (22b) or (23b) using the fixed transmit precoders $\mathbf{m}_{l,k,n}$. This procedure is carried out until convergence of the queue deviation or for fixed number of iterations by $I_{\max}$ as outlined in Algorithm 1. The convergence proof is discussed in Appendix B

Algorithm 1: Algorithm of JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall l \in \{1, 2, \ldots, L\}$
Initialize: $i = 0$ and transmit precoders $\tilde{\mathbf{m}}_{l,k,n}$ randomly satisfying the total power constraint (6b)
update $\mathbf{w}_{l,k,n}, \mathbf{u}_{l,k,n}$ using (23b) and (19) with $\mathbf{m}_{l,k,n}$
repeat
  initialize $j = 0$
  repeat
    solve for the transmit precoders $\mathbf{m}_{l,k,n}$ using (20)
    update the constraint set (19) with $\mathbf{u}_{l,k,n}$ and $\mathbf{m}_{l,k,n}$ using (18)
    $j = j + 1$
  until *SCA convergence or* $j \geq J_{\max}$
  update the receive beamformers $\mathbf{w}_{l,k,n}$ using (21) or (23b) with the updated precoders $\mathbf{m}_{l,k,n}$
  $i = i + 1$
until *Queue convergence or* $i \geq I_{\max}$

C. JSFRA Scheme via MSE Reformulation

In this, we solve the JSFRA problem by exploiting the equivalence between the MSE and the achievable sum rate for the receivers designed based on the MMSE criterion [7], [8]. The MSE $\epsilon_{l,k,n}$, for a data symbol $d_{l,k,n}$ is given by

$$\mathbb{E}\big[(d_{l,k,n} - \hat{d}_{l,k,n})^2\big] = \big|1 - \mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\big|^2 + \sum_{(j,i)\neq(l,k)} \big|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\big|^2 + \widetilde{N}_0 = \epsilon_{l,k,n}, \quad (24)$$

where $\hat{d}_{l,k,n}$ is the estimate of the transmitted symbol. *Using the MMSE receive beamformer (23b) in the MSE expression (24) and in the SINR expression (2), we can arrive at the following relation between the MSE and the SINR as*

$$\epsilon_{l,k,n} = (1 + \gamma_{l,k,n})^{-1}. \quad (25)$$

The above equivalence is valid only if the receivers are based on the MMSE criterion. Using the equivalence in (25), the WSRM objective can be reformulated as the weighted minimum mean squared error (WMMSE) equivalent to obtain the precoders for the MU-MIMO scenario as discussed in [8]–[10]. *Note that the receiver is invariably based on the MMSE criterion irrespective of the ℓ_q norm used in the objective function to obtain the optimal transmit precoders $\mathbf{m}_{l,k,n}$.*

Let $v'_k = Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ denote the queue deviation corresponding to user k and $\tilde{v}'_k \triangleq a_k^{1/q} v'_k$ represents the weighted equivalent. By using the relaxed MSE expression in (24), the problem in (6) can be expressed as

$$\underset{\substack{t_{l,k,n},\mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}'\|_q \quad (26a)$$

$$\text{subject to } t_{l,k,n} \leq -\log_2(\epsilon_{l,k,n}) \quad (26b)$$

$$\sum_{(j,i)\neq(l,k)} \big|\mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_i,k,n}\mathbf{m}_{j,i,n}\big|^2 + \widetilde{N}_0 + \big|1 - \mathbf{w}^{\mathrm{H}}_{l,k,n}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\big|^2 \leq \epsilon_{l,k,n} \quad (26c)$$

$$\sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}^{\mathrm{H}}_{l,k,n}\right) \leq P_{\max}, \forall b. \quad (26d)$$

The alternative MSE formulation given by (26) is non-convex even for the fixed $\mathbf{w}_{l,k,n}$ due to the constraint (26b), which is in fact a DC constraint. We resort to the SCA approach [23] by relaxing the constraint by a sequence of convex subsets using first order Taylor series approximation around a fixed MSE point $\tilde{\epsilon}_{l,k,n}$ as

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

Using the above approximation for the rate constraint, the problem defined in (26) is solved for optimal transmit precoders $\mathbf{m}_{l,k,n}$, MSEs $\epsilon_{l,k,n}$, and the user rates over each sub-channel $t_{l,n,k}$ for a fixed receive beamformers. The optimization subproblem to find the transmit precoders for a fixed receive beamformers $\mathbf{w}_{l,k,n}$ is given by

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

$$\text{subject to} \quad (26c), (26d), \text{ and } (27). \quad (28b)$$

The optimal transmit precoders for a fixed receivers are obtained by solving the subproblem (28) iteratively by updating the fixed MSE point $\tilde{\epsilon}_{l,k,n}$ with $\epsilon_{l,k,n}$ from the previous iteration until termination as discussed in Section IV-B.

D. Reduced Complexity Spatial Resource Allocation (SRA)

The complexity of the JSFRA algorithm scales with the number of sub-channels considered. In addition, the iterations required for the algorithm convergence increases with the problem size. To overcome the complexity, we can decouple the problem to sub-channel specific subproblems by using distributed approaches presented in [1], [13] to design the precoders using primal or dual decomposition.

As an alternative sub-optimal solution, we present queue minimizing spatial resource allocation (SRA), which solves for the precoders using JSFRA formulation only for a specific sub-channel i with a fixed transmit power $P_{\max,i}$. The sharing of power can be equal or based on a predetermined pattern as in partial frequency reuse for the sub-channels by adhering to

$$\sum_{i=1}^{N} P_{\max,i} = P_{\max}. \tag{29}$$

Even though N sub-channels are present at any given scheduling instant, precoders are solved for each sub-channel in a sequential manner with the sub-channel specific total power constraint $P_{\max,i}$ and the backlogged packets. Let $Q_{k,i}$ number of backlogged packets associated with user k before solving for precoders specific to the sub-channel i. Since the precoder design is sequential, for the initial sub-channel, the number of backlogged packets is given by $Q_{k,1} = Q_k$ and for the consecutive sub-channels, it is given by

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

where $t_{l,k,j}$ denotes the rate corresponding to the user k on the j^{th} sub-channel and l^{th} spatial stream. Note that the proposed scheme is sensitive to the order of the sub-channel selection due to the sequential precoder design for each sub-channel. However, the SRA approach provides faster convergence in contrast to the JSFRA formulation due to the substantial reduction in the optimization variables for each sub-channel problem. As the number of user increases, the SRA formulation will be insusceptible to the sub-channel ordering due to the multi-user diversity.

V. Distributed Solutions

The distributed precoder designs for the proposed JSFRA scheme are discussed in this section. The convex formulation in (20) *or* (28) *requires a centralized controller to perform the precoder design for all users belonging to the coordinating BSs. In order to design the precoders independently at each BS with the minimal information exchange via backhaul, iterative decentralization methods are addressed. In particular, the primal decomposition and the ADMM based dual decomposition approaches are considered.*

Let us consider the convex subproblem with the fixed receive beamformers $\mathbf{w}_{l,k,n}$ presented in (20) *based on the Taylor series approximation for the nonconvex constraint. The following discussions are equally valid for the MSE based solution outlined in* (28) *as well. Since the objective of* (20) *can be decoupled across each BS, the centralized problem can be equivalently written as*

$$\underset{\gamma_{l,k,n},\mathbf{m}_{l,k,n},\beta_{l,k,n}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}} \|\tilde{\mathbf{v}}_b\|_q \tag{31a}$$

$$\text{subject to} \quad (20\text{b}) - (20\text{d}), \tag{31b}$$

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

To begin with, let $\bar{\mathcal{B}}_b$ denote the set $\mathcal{B}\backslash\{b\}$ and $\bar{\mathcal{U}}_b$ represents the set $\mathcal{U}\backslash\mathcal{U}_b$. Following similar approaches presented in [14], [15], the coupling constraint (20b) *or* (26c) *can be expressed by grouping the interference contribution from each BS in $\mathcal{B}$ as*

$$\widetilde{N}_0 + \sum_{j=1,j\neq l}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,k,n}|^2 + \sum_{b\in\bar{\mathcal{B}}_{b_k}} \zeta_{l,k,n,b} + \sum_{i\in\mathcal{U}_{b_k}\backslash\{k\}}\sum_{j=1}^{L} |\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{j,i,n}|^2 \leq \beta_{l,k,n}, \tag{32}$$

where $\zeta_{l,k,n,b}$ is the total interference caused by the transmission of BS b to user $k \in \mathcal{U}_{b_k}$ in the spatial stream l and sub-channel n. It is given by the following upper bound as

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

The decentralization is achieved by decomposing the original convex problem in (31) *by a parallel iterative subproblems coordinated by either primal or dual decomposition update. The coupling variables are updated in each iteration by exchanging limited information among the subproblems. Before proceeding further, let $\bar{\boldsymbol{\zeta}}_b$ be the vector formed by stacking interference terms* (33) *from the neighboring BSs to the users of BS b and $\hat{\boldsymbol{\zeta}}_b$ be the stacked interference terms caused by BS b to all users in the neighboring BSs $\bar{\mathcal{B}}_b$, represented as*

$$\bar{\boldsymbol{\zeta}}_b = \left[\zeta_{l,k,n,\bar{\mathcal{B}}_b(1)}, \ldots, \zeta_{l,k,n,\bar{\mathcal{B}}_b(|\bar{\mathcal{B}}_b|)}\right]^{\mathrm{T}}, \forall k \in \mathcal{U}_b, \tag{34a}$$

$$\hat{\boldsymbol{\zeta}}_b = \left[\zeta_{l,\bar{\mathcal{U}}_b(1),n,b}, \zeta_{l,\bar{\mathcal{U}}_b(2),n,b}, \ldots, \zeta_{l,\bar{\mathcal{U}}_b(|\bar{\mathcal{U}}_b|),n,b}\right]^{\mathrm{T}}. \tag{34b}$$

Let us define the vector $\boldsymbol{\zeta}_b$, formed by stacking the interference terms corresponding to the BS b as

$$\boldsymbol{\zeta}_b = \left[\hat{\boldsymbol{\zeta}}_b^{\mathrm{T}}, \bar{\boldsymbol{\zeta}}_b^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{35}$$

Since the decentralization solution is an iterative procedure, we represent the i^{th} iteration index as $x^{(i)}$. Let $\boldsymbol{\zeta}_b(b_k)$ denote the interference terms corresponding to the BS b_k in BS b as

$$\boldsymbol{\zeta}_b(b_k) = \left[\zeta_{l,\mathcal{U}_b(1),n,b_k}, \ldots, \zeta_{l,\mathcal{U}_b(|\mathcal{U}_b|),n,b_k}\right]. \tag{36}$$

Now, to decouple the problem in (31), *the BS specific vector $\boldsymbol{\zeta}_b$ in* (35), *which include all interference terms relevant for the transmission of BS b, can either be fixed or*

treated as a variable in each iteration in accordance to the decomposition method. Since the BS specific precoders are solved independently, we assume local channel information and queues are available at each BS b_k together with the cross channel knowledge $\mathbf{H}_{b_k,k,n}, \forall k \in \bar{\mathcal{U}}_{b_k}$ through precoded uplink sounding using the receive beamformers $\mathbf{w}_{l,k,n}$ [29].

A. Primal Decomposition

In primal decomposition, the convex problem in (31) is solved for the optimal transmit precoders in an iterative manner for a fixed BS specific interference terms $\boldsymbol{\zeta}_{b_k}$ using master-slave model [14]. The slave subproblem is solved in each BS for the optimal transmit precoders only for the associated users by assuming fixed interference terms $\boldsymbol{\zeta}_{b_k}^{(i)}$ in each i^{th} iteration. Upon finding the optimal associated transmit precoders by each slave subproblems, the master problem is used to update the BS specific interference terms $\boldsymbol{\zeta}_{b_k}^{(i+1)}$ for the next iteration by using dual variables corresponding to the interference constraint (32) as discussed in [14]. In this manner, the interference variables are updated until the global consensus is obtained. The primal approach is similar to the minimum power precoder design presented in [14]. Note that the master problem treats $\boldsymbol{\zeta}_b$ as a variable and the slave subproblems assumes it to be a constant for each iteration to find the transmit precoders.

B. Alternating Directions Method of Multipliers (ADMM)

In this section, we discuss the ADMM approach to decouple the precoder design across multiple BSs to solve the convex problem in (31). The ADMM is preferred over the dual decomposition (DD) approach in [15] for its robustness and improved convergence behavior [1]. In contrast to the primal decomposition, the ADMM approach relaxes the interference constraints by including in the objective function of each subproblem with a penalty pricing [1], [13]. Similar decomposition for the precoder design in the minimum power context is considered in [30].

Using the formulation presented in [1], [30], we can write the BS b specific ADMM subproblem for the i^{th} iteration as

$$\underset{\substack{\gamma_{l,k,n},\mathbf{m}_{l,k,n}\\ \beta_{l,k,n},\boldsymbol{\zeta}_b}}{\text{minimize}} \quad \|\tilde{\mathbf{v}}_b\|_q + \boldsymbol{\nu}_b^{(i)\,\mathrm{T}}\left(\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(i)}\right) + \frac{\rho}{2}\left\|\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(i)}\right\|^2 \tag{37a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \le P_{\max} \tag{37b}$$

$$\sum_{\bar{b}\in\bar{\mathcal{B}}_b}\zeta_{l,k,n,\bar{b}} + \sum_{\{\bar{l},\bar{k}\}\neq\{l,k\}}|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}\mathbf{m}_{\bar{l},\bar{k},n}|^2 + \widetilde{N}_0 \le \beta_{l,k,n}, \tag{37c}$$

$$\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L}|\mathbf{w}_{\bar{l},\bar{k},n}^{\mathrm{H}}\mathbf{H}_{b,\bar{k},n}\mathbf{m}_{l,k,n}|^2 \le \zeta_{\bar{l},\bar{k},n,b}, \forall\bar{k}\in\bar{\mathcal{U}}_b, \forall n \tag{37d}$$

and (19), (37e)

where $\boldsymbol{\zeta}_b^{(i)}$ denotes the interference vector updated from the earlier iteration and $\boldsymbol{\nu}_b^{(i)}$ represents the dual vector corresponding to the equality constraint at the i^{th} iteration as

$$\boldsymbol{\zeta}_b = \boldsymbol{\zeta}_b^{(i)}. \tag{38}$$

Upon solving (37) for $\boldsymbol{\zeta}_b \forall b$ in the i^{th} iteration, the next iterate is updated by exchanging the corresponding interference terms between two BSs b and b_k as

$$\boldsymbol{\zeta}_{b_k}(b)^{(i+1)} = \boldsymbol{\zeta}_b(b_k)^{(i+1)} = \frac{\boldsymbol{\zeta}_b(b_k) + \boldsymbol{\zeta}_{b_k}(b)}{2}. \tag{39}$$

The dual vector for the next iteration is updated by using subgradient to maximize the dual objective as

$$\boldsymbol{\nu}_b^{(i+1)} = \boldsymbol{\nu}_b^{(i)} + \rho\left(\boldsymbol{\zeta}_b - \boldsymbol{\zeta}_b^{(i+1)}\right), \tag{40}$$

where the step size parameter ρ is chosen in accordance with [1], which depend on the system model under consideration for the convergence rate. The above iteration is performed until convergence or for certain accuracy in the variation of the objective value between two consecutive updates. The distributed precoder design using ADMM approach is shown in Algorithm 2.

Algorithm 2: Distributed JSFRA scheme using ADMM

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall l$
Initialize: $i = 0$ and $\mathbf{m}_{l,k,n}$ randomly satisfying total power constraint (37b)
update $\mathbf{w}_{l,k,n}$ with (23b) and $\tilde{\mathbf{u}}_{l,k,n}$ using (16c) and (18)
initialize the interference vectors $\boldsymbol{\zeta}_b^{(0)} = \mathbf{0}^{\mathrm{T}}, \forall b \in \mathcal{B}$
initialize the dual vectors $\boldsymbol{\nu}_b^{(0)} = \mathbf{0}^{\mathrm{T}}, \forall b \in \mathcal{B}$
foreach *BS* $b \in \mathcal{B}$ **do**
  repeat
    initialize $j = 0$
    repeat
      solve for $\mathbf{m}_{l,k,n}$ and $\boldsymbol{\zeta}_b$ with (37) using $\boldsymbol{\zeta}_b^{(j)}$
      exchange $\boldsymbol{\zeta}_b$ among BSs in $\mathcal{B}$
      update interference vector $\boldsymbol{\zeta}_b^{(j+1)}$ using (39)
      update dual variables in $\boldsymbol{\nu}_b^{(j+1)}$ using (40)
      $j = j + 1$
    until *convergence or* $j \ge J_{\max}$
    downlink precoded pilot transmission with $\mathbf{m}_{l,k,n}$
    update $\mathbf{w}_{l,k,n}$ and notify all BSs in $\mathcal{B}$ using uplink precoded pilots [29]
    update $\tilde{\mathbf{u}}_{l,k,n}$ using (16c) and (18) for SCA point or $\tilde{\epsilon}_{l,k,n}$ using (26c) for MSE operating point
    $i = i + 1$
  until *convergence or* $i \ge I_{\max}$
end

C. Convergence Analysis for Distributed Algorithms

The convergence of the distributed algorithm outlined in Algorithm 2 follows the same discussion in Appendix B if the subproblem (31) converge to the centralized solution. Since the subproblem (31) is convex, each BS specific slave subproblem is also convex for a fixed interference vector $\boldsymbol{\zeta}_{b_k}^{(i)}$ [13]. The master subproblem in the primal decomposition uses subgradient to update the coupling interference vectors in consensus with the objective function, it is guaranteed to

converge to the centralized solution as the iteration $i \to \infty$ [4] for a diminishing step size. It can be seen that the subproblem (31) *satisfies Slater's constraint qualification by having non empty interior and bounded due to the total power constraint for the transmit precoders.*

To prove the convergence of the ADMM approach, we use the argument presented in [31] Proposition 4.2. If the problem is written as

$$\underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad G(\mathbf{x}) + H(\mathbf{y}) \tag{41a}$$

$$\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{z} \tag{41b}$$

$$\mathbf{x} \in \mathcal{C}_1, \mathbf{z} \in \mathcal{C}_2, \tag{41c}$$

following conditions are required for the convergence if ADMM is used.

- G, H *should be convex*
- $\mathcal{C}_1, \mathcal{C}_2$ *should be a convex set and bounded*
- $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ *should be invertible.*

Note that the equality constraint (41b) *is identical to* (38) *used in the ADMM subproblem* (37)*. It is evident from the equality constraint* (38) *that* $\mathbf{A} = \mathbf{I}$*, which is an identity matrix and is invertible. The objective function G, H are ℓ_q norm in* (31) *are convex and the set defined by the constraints of the problem* (31) *are all convex sets and has a nonempty interior. The feasibility of the interior point is verified by having a non zero precoder for only one user. Therefore, by following [31, Prop. 4.2], it can be seen that the ADMM approach converges to the centralized solution as $i \to \infty$.*

D. Decomposition via KKT Conditions for MSE Formulation

In this section, we discuss an alternative way to decentralize the precoder design across the coordinating BSs in $\mathcal{B}$ based on the MSE reformulation method discussed in Section IV-C. In contrast to Section V-A and V-B, the problem is solved using the KKT conditions in which the transmit precoders, receive beamformers and the subgradient updates are performed at the same instant to minimize the global queue deviation objective with few number of iterations. The proposed methods in this section provide algorithms that can be of practical importance owing to the limited signaling requirements. *We consider an idealized TDD system due to the knowledge of complete channel information at the transmitter*. Similar work has been considered for the WSRM problem with minimum rate constraints in [11], [12]. Since the formulation in [11], [12] are similar to the Q-WSRME scheme with an additional maximum rate constraint (14), it requires explicit dual variables to handle the maximum rate constraint, thereby making the problem difficult to solve in an iterative manner.

In the proposed JSFRA formulation, the maximum rate constraints are implicitly handled by the objective function without the need of explicit constraints. However, the KKT conditions cannot be formulated due to the non-differential objective function. The non-differentiability is due to the absolute value operator present in the norm function. In order to make the objective function differentiable, we consider the following two cases for which the absolute operator can be ignored without affecting the optimal solution, namely,

- when the exponent q is even, or
- when the number of backlogged packets of each user is large enough, *i.e*, $Q_k \gg \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}$ to ignore the absolute operator, which means also ignoring the queues in the first place as well.

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

$$\underset{\substack{t_{l,k,n},\mathbf{m}_{l,k,n},\\ \epsilon_{l,k,n},\mathbf{w}_{l,k,n}}}{\text{minimize}} \quad \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{U}_b} a_k \left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\right)^q \tag{42a}$$

subject to

$$\alpha_{l,k,n} : \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\right|^2 + \widetilde{N}_0 + \sum_{(x,y)\neq(l,k)} \left|\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\right|^2 \leq \epsilon_{l,k,n} \tag{42b}$$

$$\sigma_{l,k,n} : \log_2(\tilde{\epsilon}_{l,k,n}) + \frac{(\epsilon_{l,k,n} - \tilde{\epsilon}_{l,k,n})}{\log(2)\tilde{\epsilon}_{l,k,n}} \leq -t_{l,k,n} \tag{42c}$$

$$\delta_b : \sum_{n=1}^{N}\sum_{k\in\mathcal{U}_b}\sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}, \forall b, \tag{42d}$$

where $\alpha_{l,k,n}, \sigma_{l,k,n}$ and δ_b are the dual variables corresponding to the constraints defined in (42b), (42c) and (42d).

The problem in (42) is solved using the KKT expressions, which are obtained by the derivative of the Lagrangian function w.r.t the primal and the dual variables, complementary slackness conditions, and the primal, dual feasibility requirements as shown in Appendix C. Upon solving, we obtain the iterative solution as

$$\mathbf{m}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L} \alpha_{y,x,n}^{(i-1)}\mathbf{H}_{b_k,x,n}^{\mathrm{H}}\mathbf{w}_{y,x,n}^{(i-1)}\mathbf{w}_{y,x,n}^{\mathrm{H}\,(i-1)}\mathbf{H}_{b_k,x,n} + \delta_b\mathbf{I}_{N_T}\Big)^{-1}\alpha_{l,k,n}^{(i-1)}\mathbf{H}_{b_k,k,n}^{\mathrm{H}}\mathbf{w}_{l,k,n}^{(i-1)} \tag{43a}$$

$$\mathbf{w}_{l,k,n}^{(i)} = \Big(\sum_{x\in\mathcal{U}}\sum_{y=1}^{L} \mathbf{H}_{b_x,k,n}\mathbf{m}_{y,x,n}^{(i)}\mathbf{m}_{y,x,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_x,k,n}^{\mathrm{H}} + \mathbf{I}_{N_R}\Big)^{-1}\mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}^{(i)}. \tag{43b}$$

$$\epsilon_{l,k,n}^{(i)} = \left|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}^{(i)}\right|^2 + N_0\left\|\mathbf{w}_{l,k,n}^{(i)}\right\|^2 + \sum_{(x,y)\neq(l,k)}\left|\mathbf{w}_{l,k,n}^{\mathrm{H}\,(i)}\mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}^{(i)}\right|^2 \tag{43c}$$

$$t_{l,k,n}^{(i)} = -\log_2(\epsilon_{l,k,n}^{(i-1)}) - \frac{\left(\epsilon_{l,k,n}^{(i)} - \epsilon_{l,k,n}^{(i-1)}\right)}{\log(2)\,\epsilon_{l,k,n}^{(i-1)}} \tag{43d}$$

$$\sigma_{l,k,n}^{(i)} = \left[\frac{a_k\,q}{\log(2)}\left(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}^{(i)}\right)^{(q-1)}\right]^+ \tag{43e}$$

$$\alpha_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i-1)} + \rho\left(\frac{\sigma_{l,k,n}^{(i)}}{\epsilon_{l,k,n}^{(i)}} - \alpha_{l,k,n}^{(i-1)}\right) \tag{43f}$$

Since the dual variables $\alpha^{(i)}$ and $\sigma^{(i)}$ are interdependent in (43), one has to be fixed to optimize for the other. So, $\alpha^{(i)}$ is fixed to evaluate $\sigma^{(i)}$ using (43). *At each iteration, the dual variables $\alpha^{(i)}$ are linearly interpolated with any point between the fixed iterate $\alpha^{(i-1)}$ and $\frac{\sigma^{(i)}}{\epsilon^{(i)}}$ using a step size $\rho \in (0, 1)$. The choice of ρ depends on the system model and it affects*

the convergence behavior of the algorithm. It is used to reduce the oscillations in the objective function when $\sigma^{(i)}$ is negative due to over allocation.

When the allocated rate $t_k^{(i-1)}$ is greater than the number of queued packets Q_k for a user k, the corresponding dual variable $\sigma^{(i)}$ will be negative and due to the projection operator $[x]^+$ in (43e), it will be zero, thereby forcing $\alpha_k^{(i)} < \alpha_k^{(i-1)}$ as in (43f). Once the $\alpha_k^{(i)}$ is reduced, the precoder weight in (43a) is lowered to make the rate $t_k^{(i)} < t_k^{(i-1)}$ eventually. *The choice of ρ is susceptible to the system model under consideration, which affects the convergence speed of the iterative algorithm. In our simulations, we fixed $\rho = 0.1$ irrespective of the model under consideration.*

The KKT expressions in (43) are solved in an iterative manner by initializing the transmit and the receive beamformers $\mathbf{m}_{l,k,n}, \mathbf{w}_{l,k,n}$ with the single user beamforming and the MMSE vectors. The dual variable α's are initialized with ones to have equal priorities to all the users in the system. Then the transmit and the receive beamformers are evaluated using the expressions in (43). The transmit precoder in (43a) depends on the BS specific dual variable δ_b, which can be found by bisection search satisfying the total power constraint (42d). Note that the fixed SCA operating point is given by $\tilde{\epsilon}_{l,k,n} = \epsilon_{l,k,n}^{(i-1)}$, which is considered in the expression (43).

To devise an algorithm for a practical implementation, we assume the cross channels $\mathbf{H}_{b,k,n}, \forall k \in \bar{\mathcal{U}}_b$ and the receive beamformers $\mathbf{w}_{l,k,n}$ of all users in the system are known through uplink signaling. We extend the decentralization methods discussed in [29], for the current problem as follows. After receiving the updated transmit precoders from all BSs in $\mathcal{B}$, each user evaluates the MMSE receiver in (43b) and notify them to the BSs via uplink precoded pilots. On receiving pilot signals, BSs update the MSE in (24) as

$$\epsilon_{l,k,n}^{(i)} = 1 - \mathbf{w}_{l,k,n}^{(i)\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}^{(i)}. \tag{44}$$

Using the current MSE value, $t_{l,k,n}^{(i)}, \sigma_{l,k,n}^{(i)}$, and $\alpha_{l,k,n}^{(i)}$ are evaluated using (43d), (43e) and (43f), and the updated dual variables $\alpha_{l,k,n}$ are exchanged between the BSs to evaluate the transmit precoders $\mathbf{m}_{l,k,n}^{(i+1)}$ for the next iteration. The SCA operating point is also updated with the current MSE value.

To avoid the back-haul exchanges between BSs, as an alternative approach, users may perform all processing required and BSs will update the precoders based on the feedback information from the users. Upon receiving the transmit precoders from BSs, each user will update the receive beamformer $\mathbf{w}_{l,k,n}$, the MSE $\epsilon_{l,k,n}$, and the dual variables $\lambda_{l,k,n}$ and $\alpha_{l,k,n}$. The updated $\alpha_{l,k,n}$ and $\mathbf{w}_{l,k,n}$ are notified to the BSs using two separate precoded uplink pilot symbols with $\tilde{\mathbf{w}}_{l,k,n}^{(i)} = \sqrt{\alpha_{l,k,n}^{(i)}} \mathbf{w}_{l,k,n}^{*(i)}$ and $\bar{\mathbf{w}}_{l,k,n}^{(i)} = \alpha_{l,k,n}^{(i)} \mathbf{w}_{l,k,n}^{*(i)}$ as the precoders. On receiving the precoded uplink pilots, each BS use the effective channel $\mathbf{H}_{b,k,n}^{\mathrm{T}} \tilde{\mathbf{w}}_{l,k,n}^{(i)}$ and $\mathbf{H}_{b,k,n}^{\mathrm{T}} \bar{\mathbf{w}}_{l,k,n}^{(i)}$ in (43a) to update the transmit precoders, where $\mathbf{x}^*$ is the complex conjugate of $\mathbf{x}$. Finally, Algorithm 3 outlines the distributed precoder design using the KKT based MSE reformulated JSFRA problem.

The Algorithm 3 outlines a practical way of implementing the transmit precoders in a distributed manner using over-the-air (OTA) signaling of the transmit precoders and the receive beamformers for certain iterations before the actual transmission of data is performed with it. Unlike primal decomposition (PD) or ADMM approach, all variables are updated at once, i.e, the SCA point of $\epsilon^{(i-1)}$, AO update of $\mathbf{w}_{l,k,n}$ and the dual variable α using subgradient update, it is not guaranteed to obtain the same point as that of the centralized problem. Since the problem solution in (43) is equivalent to the centralized formulation in (28) if the receive beamformers $\mathbf{w}_{l,k,n}$ and the MSE operating point $\epsilon_{l,k,n}^{(i-1)}$ are fixed and optimized for the transmit precoders $\mathbf{m}_{l,k,n}$ and the dual variable $\alpha_{l,k,n}$, the problem is guaranteed to converge to the centralized solution. Note that it requires four iterations to obtain the centralized solution, namely, the receive beamformer loop, MSE operating point loop, dual variable update loop and the bisection method for finding the transmit precoders. In order to avoid this, the proposed method performs the group update of all variables to obtain the transmit and the receive beamformers with the limited number of iterations. If the step size $\rho < 1$, the algorithm for ℓ_2 norm will converge by using the arguments on controlling overallocation and the nonincreasing objective function, since the earlier iterates are the operating point for the current iteration.

Algorithm 3: KKT approach for the JSFRA scheme

Input: a_k, Q_k, $\mathbf{H}_{b,k,n}$, $\forall b \in \mathcal{B}$, $\forall k \in \mathcal{U}, \forall n \in \mathcal{N}$
Output: $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n} \forall l \in \{1, 2, \ldots, L\}$
Initialize: $i = 1$, $\mathbf{w}_{l,k,n}^{(0)}, \tilde{\epsilon}_{l,k,n}$ randomly, dual variables $\alpha_{l,k,n}^{(0)} = 1$, and $I_{\max}$ for certain value
foreach *BS* $b \in \mathcal{B}$ **do**
  initialize $i = 0$
  repeat
    update $\mathbf{m}_{l,k,n}^{(i)}$ using (43a), and perform downlink transmission
    find $\mathbf{w}_{l,k,n}^{(i)}$ using (43b) at each user
    evaluate $\epsilon_{l,k,n}^{(i)}$, $t_{l,k,n}^{(i)}$, $\sigma_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ using (43c) and (43d), (43e) and (43f) at each user with the updated $\mathbf{w}_{l,k,n}^{(i)}$
    using precoded uplink pilots, $\mathbf{m}_{l,k,n}^{(i)}$ and $\alpha_{l,k,n}^{(i)}$ are notified to all BSs in $\mathcal{B}$
    $i = i + 1$
  until *until convergence or* $i \geq I_{\max}$
end

VI. Simulation Results

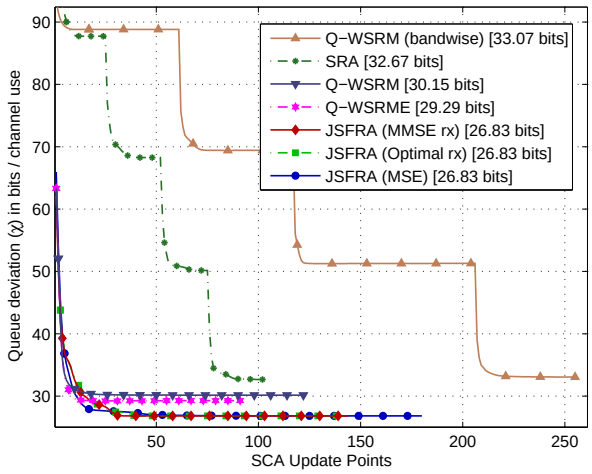
The simulations carried out in this work consider the path loss varying uniformly across all users in the system with the channels drawn from the *i.i.d.* samples. The queues are generated based on the Poisson process with the average values specified in each section presented.

A. Centralized Solutions

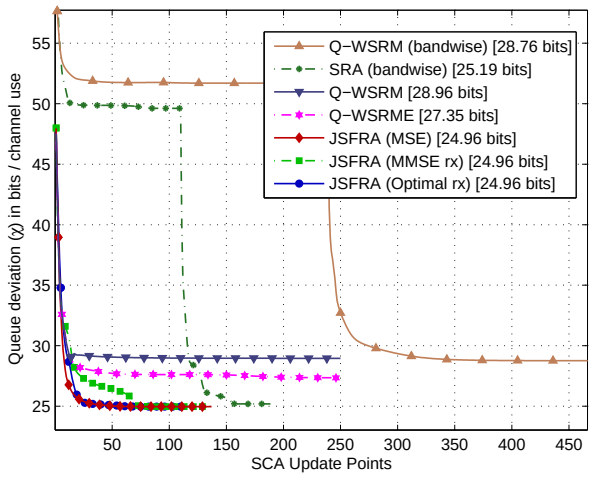
We discuss the performance of the centralized algorithms in Section IV for some system configurations. To begin with, we

TABLE I
SUB-CHANNEL-WISE LISTING OF CHANNEL GAINS AND RATE ALLOCATIONS BY DIFFERENT ALGORITHMS FOR A SCHEDULING INSTANT

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



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



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

Fig. 1. Total number of backlogged packets χ present in the system after each SCA updates *using $\ell_1(q = 1)$ norm in objective*

TABLE II
NUMBER OF BACKLOGGED BITS ASSOCIATED WITH EACH USER FOR A SYSTEM $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$.

q	user indices								χ
	1	2	3	4	5	6	7	8	
1	15.0	3.95	5.26	8.95	7.0	11.9	12.0	9.7	25.15
2	11.2	3.9	10.76	10.65	10.27	9.68	8.77	5.9	27.77
∞	11.4	4.4	10.4	10.4	10.4	8.4	8.4	6.4	28.68
Q_k	15.0	8.0	14.0	14.0	14.0	12.0	12.0	10.0	

consider a single cell single-input single-output (SISO) model operating at 10 dB signal-to-noise ratio (SNR) with $K = 3$ users sharing $N = 3$ sub-channel resources. The number of packets waiting at the transmitter for each user is given by $Q_k = 4, 8$ and 4 bits, respectively.

Table I tabulates the channel seen by the users over each sub-channel followed by the rates assigned by three different algorithms, Q-WSRME allocation, JSFRA approach and the band-wise Q-WSRM scheme using the WMMSE design [9]. The performance metric used for the comparison is the total number of backlogged bits left over at each slot after the allocation, which is denoted as $\chi = \sum_{k=1}^{K} [Q_k - t_k]^+$. Even though $\mathcal{U}(1)$ and $\mathcal{U}(3)$ has equal number of backlogged packets of $Q_1 = Q_3 = 4$ bits, user $\mathcal{U}(3)$ is scheduled in the first sub-channel due to the better channel condition. In contrast, the JSFRA approach assigns the first user on the first sub-channel, which reduces the total number of backlogged packets waiting at the transmitter. The rate allocated for $\mathcal{U}(2)$ on the second sub-channel is higher in JSFRA scheme compared to the other schemes. It is due to the efficient allocation of the total power shared across the sub-channels.

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

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

For both scenarios in Fig. 1, the Q-WSRME performs marginally inferior to the JSFRA algorithms due to the weights used in the algorithm. The performance loss is attributed to

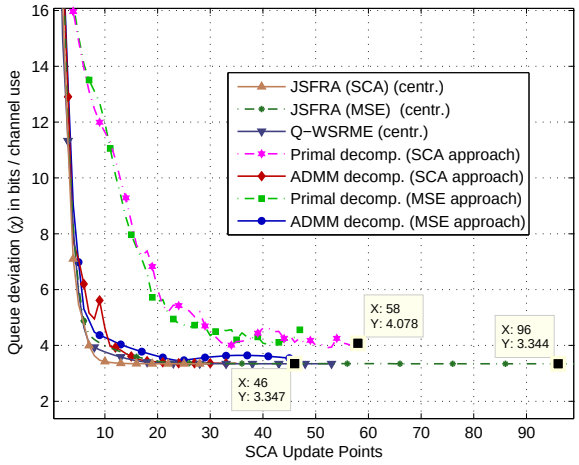


Fig. 2. Convergence behaviour of the centralized and the distributed algorithms for a system $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ *using* ℓ_1 *norm*

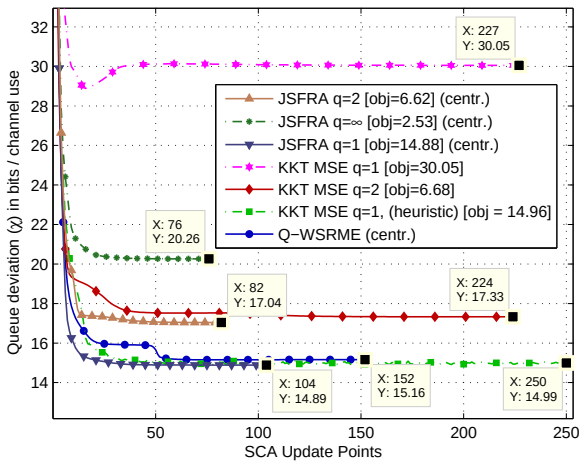


Fig. 3. Impact of varying q in the total number of backlogged packets after each SCA update for a system $\{N, N_B, K, N_R\} = \{5, 2, 8, 1\}$

the fact that the Q-WSRME algorithm favors the users with the large number of backlogged packets as compared to the users with better channel conditions. Fig. 1(b) compares the algorithms for $N_R = 2$ receive antenna case. In all figures, the receivers are updated along with the SCA update instants *i.e*, $J_{\max} = 1$ in Algorithm 1. It is also noted that the degradation by performing combined update is marginal, since the receiver minimizes the objective for a fixed transmit precoders thereby leading to a monotonic convergence.

The behavior of the JSFRA algorithm for different exponents q is outlined in the Table II for the users located at the cell-edge of the system employing $N_T = 4$ transmit antennas. It is evident that the JSFRA algorithm minimizes the total number of queued bits for the ℓ_1 norm compared to the ℓ_2 norm, which is shown in the column displaying the total number of left over packets χ in bits. The ℓ_∞ norm provides fair allocation of the resources by making the left over packets to be equal for all users to $\chi_k = 3.58$ bits.

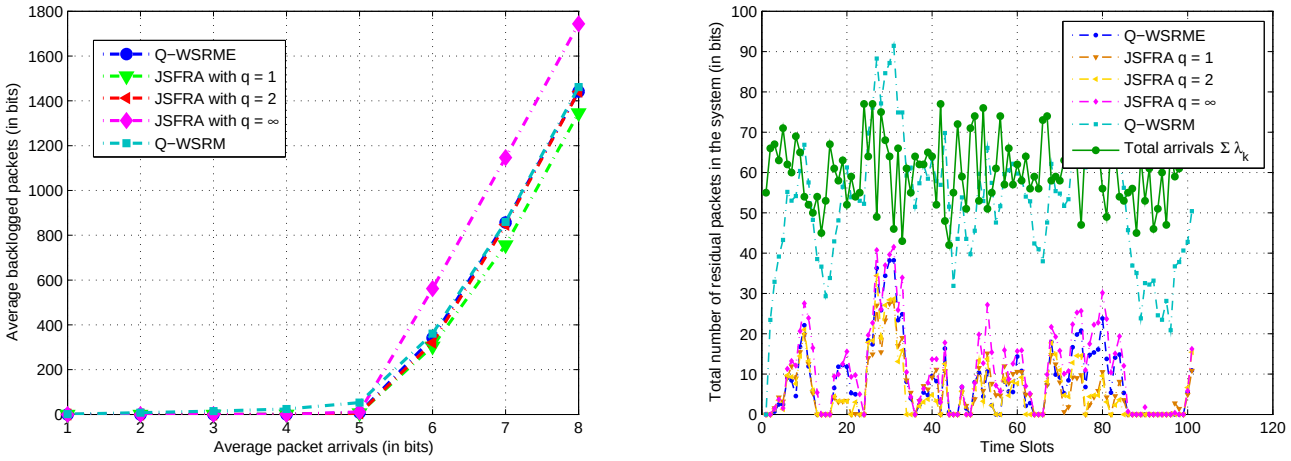
B. Distributed Solutions

The performance of the distributed algorithms are compared using the total number of backlogged packets after each SCA update points. Fig. 2 compares the performance of the algorithms for the system configuration $\{N, N_B, K, N_R\} = \{3, 2, 8, 1\}$ with $N_T = 4$ transmit antennas at the BSs. Each BS serves $|\mathcal{U}_b| = 4$ users in a coordinated manner to reduce the total number of backlogged packets at each BS. The total number of queued packets assumed for both figures is $Q_k = [5, 7, 9, 11, 8, 12, 5, 4]$ bits. As pointed out in Section V, the performance and the convergence speed of the distributed algorithms are susceptible to the step size used in the subgradient update. Due to the fixed interference levels in the primal approach, it may lead to infeasible solutions if the initial or any intermediate update is not feasible.

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

Fig. 3 compares the performance of the centralized and the KKT algorithm in Section V-D for different exponents by plotting the total number of backlogged packets at each SCA update point. The ℓ_1 norm JSFRA scheme provides better performance over other schemes due to the greedy objective. The KKT approach for ℓ_1 norm is not defined due to the non-differentiability of the objective as discussed in the Section V-D. If used for ℓ_1 norm, the problem of over-allocation will not affect the dual variables $\sigma_{l,k,n}$ and $\alpha_{l,k,n}$ since the queue deviation is raised to the power zero in (43e), which will always be equal to one. A heuristic method based on subdifferential calculus in [4] is proposed in Fig. 3 by assigning zero for $\sigma_{l,k,n}$ when the queue deviation is negative, *i.e*, $Q_k - t_k < 0$. It is required to address the problem of over-allocation in the ℓ_1 norm for dropping the absolute value operator from the objective function. It can be seen that the heuristic method oscillates near the stationary point with the deviation determined by the factor ρ used in (43f).

The objective values are mentioned in the legend for all the schemes and the objective of the ℓ_2 norm is not the same as that of the ℓ_1 norm used for plotting. For simulations, we update all variables in (43) at once at each iteration, *i.e*, $J_{\max} = 1$, which is well justified for the practical implementations due to the signaling overheads. The ℓ_2 norm for the JSFRA and the KKT approach achieves nearly the same value of 6.62 with different χ, due to the limited number of iterations for the dual variable convergence between each SCA update. Fig. 3 also shows the effect of dropping the squared rate variable from the objective in the Q-WSRME scheme compared to the ℓ_2 norm which includes it. By dropping it, the



(a). Average backlogged packets in the system after 100 transmission instants

(b). Total backlogged packets at each transmission slot for $A_k = 5$ bits

Fig. 4. Time analysis of the Queue dynamics for a system $\{N, N_B, K, N_R\} = \{4, 2, 12, 1\}$

Q-WSRME scheme minimizes the number of queued packets in a prioritized manner based on the respective queues. On contrary, the ℓ_2 norm allocate rates to the users with the higher number of queued packets before addressing the users with the smaller number of queued packets.

C. Average Backlogged Packets Over Slots

We discuss performance of the JSFRA algorithm for different values of ℓ_q over multiple transmission slots. It is compared with the existing Q-WSRME scheme by varying the average arrival rate A_k of all users. Fig. 4 demonstrates the performance of the centralized algorithms for different ℓ_q values. Even though A_k's are constant for all users, the instantaneous arrivals are random and is based on Poisson arrival process. We considered a 4×1 MIMO system with $N = 4$ sub-channels and $N_B = 2$ BSs. The path loss is modeled using a uniform random variable $[0, -3]$ dB with the maximum SINR seen by any user is 6 dB.

Fig. 4a compares various schemes with the average number of backlogged packets present in the system after each transmission slot. The performance of the JSFRA scheme using ℓ_2 and Q-WSRME approach are similar in the average number of residual packets after each transmission slot. Note that the additional rate constraints in the Q-WSRME scheme is the reason for the equivalence. Both Q-WSRM and Q-WSRME performs similar to ℓ_2 JSFRA scheme when the arrival rates are significantly greater than the actual transmissions. It can be seen from Fig. 4a and Fig. 4b that the number of backlogged packets are noticeably less for the ℓ_1 JSFRA formulation due to the greedy allocation by serving users with better channel conditions. Fig. 4 shows that the ℓ_∞ JSFRA scheme performs worst in terms of the average number of backlogged packets due to the instantaneous fairness constraints.

VII. Conclusions

In this paper, we addressed the problem of allocating downlink space-frequency resources to the users in a multi-cell MIMO IBC system using OFDM. The resource allocation is considered as a joint space-frequency precoder design problem since the allocation of a resource to a user is obtained by a non-zero precoding vector. We proposed the JSFRA scheme by relaxing the nonconvex DC constraint by a sequence of convex subsets using SCA for designing the precoders to minimize the total number of user queued packets. Additionally, an alternative MSE relaxation approach is also proposed by using SCA to address the nonconvex DC constraints for a fixed MMSE receivers. We also proposed distributed precoder designs for the JSFRA problem using primal and ADMM methods.Finally, we proposed a practical iterative algorithm to obtain the precoders in a decentralized manner by solving the KKT conditions of the MSE reformulated JSFRA method. The proposed iterative algorithm requires few iterations and limited signaling exchange between the coordinating BSs to obtain the efficient precoders for a given number of iterations. Numerical results are used to compare the performance of the proposed algorithms with the existing solutions. The distributed precoder design for the time correlated fading will be considered in future.

Appendix A
Tightness of SINR relaxation

For the constraints (16b) *and* (16c) *to be active, there should be at least one user in each BS with enough backlogged packets that cannot be served with the given power budget. On the other hand, to make the constraints active in all cases, the objective of the JSFRA formulation should be regularized with the transmit power without affecting the solution as*

$$\|\tilde{\mathbf{v}}\|_q + \varphi \sum_{k\in\mathcal{U}} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right),$$

where $\varphi \approx 0$. Note that the modified objective will relax the power constraint by making the constraints (16b) *and* (16c) *active at the final solution.*

APPENDIX B
CONVERGENCE PROOF FOR CENTRALIZED ALGORITHM

To prove the convergence of the centralized algorithm in (16) *and* (26)*, we need to show that the following conditions are to be satisfied by the objective function sequence to prove the convergence of the transmit and the receive beamformers.*

(a) *The function should be coercive and bounded below*
(b) *The feasible set should be a compact set*
(c) *The sequence should be monotonically decreasing for each iteration*
(d) *Uniqueness of the mapping, i.e, there should be one-to-one correspondence between the objective value set and the feasible domain.*

Using [4, Prop. A.8], the existence of a global minimizer in the feasible set can be guaranteed if the conditions (a) and (b) are satisfied. Since the feasible set is not fixed in each iteration, we require additional conditions (c) and (d) to prove the global convergence of the objective function and the corresponding arguments namely, the transmit and the receive precoders.

Assuming the above conditions are satisfied by the objective function of the iterative algorithm, using Bolzano-Weierstrass theorem [32], we can show that any bounded monotonically decreasing sequence has a unique limiting point. The one-to-one mapping between the feasible domain and the objective function value is required to prove the convergence of the transmit and the receive beamformers upon convergence of the objective function value iterates. Note that the relaxed iterative problem has the unique solution but not the original nonconvex problem. Since the transmit precoders in problem in (16) *and* (26) *are invariant to any unitary matrix rotations due to the absolute value operator in the SINR expression, the limiting point is a set of precoder vectors with different phase rotations. The limiting point of the iterative algorithm is susceptible to the initial feasible point, since it is the operating point for linearizing the nonconvex DC constraint.*

A. Boundedness

The feasible set of the relaxed nonconvex problem in (16) *and* (26) *is bounded due to the total power constraint on the transmit precoders* (16d)*. Note that the other variables are also bounded in accordance with this constraint. Since the feasible set also includes the boundary value due to the inequality constraint, it is closed and compact.*

To prove the boundedness, it is enough to show that the objective function is bounded below due to the minimization objective. Since the minimum value of the norm operator is zero, i.e, the limiting point is $< -\infty$*, it is bounded below. Note that the objective function is Lipschitz continuous over the feasible set, and therefore, it is bounded from above as well, since the feasible set is bounded. The objective function in* (16) *and* (26) *is continuous and approaches* ∞ *as* $\mathbf{t}_k \to \infty$*, therefore it is coercive. Note that the norm operator is not differentiable at the minimum point for* ℓ_1 *and* $\ell_i nfty$*. Using conditions (a) and (b), we can guarantee the existence of a minimizer to the nonconvex problem in* (16) *and* (26)*.*

B. Monotonicity

Let us express the centralized problem in (16) *and* (26) *as*

$$\underset{\mathbf{m},\mathbf{w},\boldsymbol{\gamma}}{\textit{minimize}} \quad f(\mathbf{m},\mathbf{w},\boldsymbol{\gamma}) \tag{45a}$$

$$\textit{subject to} \quad h(\boldsymbol{\gamma}) - g_0(\mathbf{m},\mathbf{w}) \leq 0 \tag{45b}$$

$$g_1(\mathbf{m},\mathbf{w}) \leq 0, \tag{45c}$$

$$g_2(\mathbf{m}) \leq 0, \tag{45d}$$

where g_2, f *are convex functions and* h *is a linear function. Let* g_0, g_1 *are convex functions only on* $\mathbf{m}$ *or* $\mathbf{w}$ *as the variable but not on both. Note that the* (45b) *correspond to the constraints in* (16b) *and* (26b) *and* (45c) *correspond to the constraints in* (16c) *and* (26c)*. Other convex constraints are addressed by the constraint* (45d)*. With this, the feasible set of the problem* (45) *is given by*

$$\mathcal{F} = \{\, \mathbf{m},\mathbf{w},\boldsymbol{\gamma} \;\big|\; h(\boldsymbol{\gamma}) - g_0(\mathbf{m},\mathbf{w}) \leq 0, \; g_1(\mathbf{m},\mathbf{w}) \leq 0, g_2(\mathbf{m}) \leq 0 \,\}$$

To solve (45)*, we adopt AO by fixing a block of varibles and optimize for others [33]. In* (45)*, even after fixing the variable* $\mathbf{w}$*, the problem is nonconvex due to the DC constraint* (45b)*. We adopt SCA approach presented in [26], [27], [34] by relaxing the nonconvex set by a sequence of convex subsets. Since the proposed method involves two level of iterations, we denote the AO iteration index by a superscript* (i) *and the DC constraint relaxations by a subscript* k*. Let* $\mathcal{X}_k^{(i)}$ *be the feasible set for the* i^{th} *AO iteration and the* k^{th} *SCA point for a fixed* $\mathbf{w}$ *and* $\mathcal{Y}_k^{(i)}$ *denotes the feasible set for a fixed* $\mathbf{m}$*. Since the SCA iterations are performed until convergence, let* $\mathbf{m}_*^{(i)}$ *denotes the converged point of* $\mathbf{m}$ *in the* i^{th} *AO iteration. For the sake of clarity, we define the optimal value of* $\boldsymbol{\gamma}$ *obtained for the* i^{th} *AO iterate for the fixed* $\mathbf{w}$ *variable as* $\boldsymbol{\gamma}_{*|y}^{(i)}$*.*

To begin with, let us consider the variable $\mathbf{w}$ *is fixed for the AO* i *with the optimal value achieved from the previous iteration* $i-1$ *as* $\mathbf{w}_*^{(i-1)}$*. In order to solve for* $\mathbf{m}$ *in the SCA iteration* k*, we linearize the nonconvex function* g_0 *using previous SCA iterate of* $\mathbf{m}$ *as*

$$\hat{g}_o(\mathbf{m},\mathbf{w}_*^{(i-1)};\mathbf{m}_k^{(i)}) = g_0(\mathbf{m}_k^{(i)},\mathbf{w}_*^{(i-1)}) + \nabla g_0(\mathbf{m}_k^{(i)},\mathbf{w}_*^{(i-1)})^{\mathrm{T}}(\mathbf{m}-\mathbf{m}_k^{(i)}). \tag{46}$$

Using (46)*, the convex subproblem for* i^{th} *AO iteration and* k^{th} *SCA point for the variable* $\mathbf{m}$ *and* $\boldsymbol{\gamma}$ *is given by*

$$\underset{\mathbf{m},\boldsymbol{\gamma}}{\textit{minimize}} \quad f(\mathbf{m},\mathbf{w}_*^{(i-1)},\boldsymbol{\gamma}) \tag{47a}$$

$$\textit{subject to} \quad h(\boldsymbol{\gamma}) - \hat{g}_0(\mathbf{m},\mathbf{w}_*^{(i-1)};\mathbf{m}_k^{(i)}) \leq 0 \tag{47b}$$

$$g_1(\mathbf{m},\mathbf{w}_*^{(i-1)}) \leq 0, \tag{47c}$$

$$g_2(\mathbf{m}) \leq 0, \tag{47d}$$

Let the feasible set defined by the problem in (47) *be represented as* $\mathcal{X}_k^{(i)} \subset \mathcal{F}$*. In order to prove the convergence of the convex subproblem* (47) *for a fixed* $\mathbf{w} = \mathbf{w}_*^{(i-1)}$ *operating at* $\mathbf{m}_k^{(i)}$*, let us consider that* (47) *yields* $\mathbf{m}_{k+1}^{(i)}$ *and* $\boldsymbol{\gamma}_{k+1}^{(i)}$ *as the solution for the* k^{th} *iteration. Note that the point* $\mathbf{m}_{k+1}^{(i)}$ *and* $\boldsymbol{\gamma}_{k+1}^{(i)}$*, which minimizes the objective function, is also feasible*

for (47) *using the following inequality*

$$h(\boldsymbol{\gamma}_{k+1}^{(i)}) - g_0(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}) \leq -\hat{g}_0(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}; \mathbf{m}_k^{(i)}) + h(\boldsymbol{\gamma}_{k+1}^{(i)}) \leq h(\boldsymbol{\gamma}_k^{(i)}) - \hat{g}_0(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}; \mathbf{m}_k^{(i)}) \leq 0. \quad (48)$$

Using (48)*, we can prove that the solution* $\mathbf{m}_{k+1}^{(i)}$ *and* $\boldsymbol{\gamma}_{k+1}^{(i)}$ *are feasible, since the initial point of* $\mathbf{m} = \mathbf{m}_*^{(i-1)}$ *was chosen to be feasible from the earlier AO iteration* $i-1$*. At each SCA iteration, the feasible set includes the optimal point from the previous iteration as* $\{\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k+1}^{(i)}\} \in \mathcal{X}_{k+1}^{(i)} \subset \mathcal{F}$*, thereby, leading to the monotonic decrease in the objective function [27], [34], [35] as*

$$f(\mathbf{m}_0^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_0^{(i)}) \geq f(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)}) \geq f(\mathbf{m}_{k+1}^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{k+1}^{(i)}) \geq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|y}^{(i)}). \quad (49)$$

Thus the sequence $f(\mathbf{m}_k^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_k^{(i)})$ *is nonincreasing and approaches limiting point as* $k \to \infty$*. Note that feasible point* $(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|y}^{(i)})$ *need not be a stationary point for* (45)*, since it is the minimizer only over the feasible set* $\mathcal{X}_*^{(i)} \subset \mathcal{F}$*, for a fixed* $\mathbf{w}$*.*

Once the solution is found for a fixed $\mathbf{w}$*, we fix* $\mathbf{m}$ *as* $\mathbf{m}_*^{(i)}$ *and optimize for* $\mathbf{w}$*. Even after treating* $\mathbf{m}$ *as a constant, the problem is still nonconvex due to the DC constraint. Following similar approach, we can find the minimizer* $\mathbf{w}_k^{(i)}$ *and* $\boldsymbol{\gamma}_k^{(i)}$ *for a similar convex subproblem* (47) *at each iteration* k*. Note that* $\boldsymbol{\gamma}_k^{(i)}$ *is reused since the variable* $\mathbf{m}$ *is fixed for the* i^{th} *AO iteration. The convergence and the nonincreasing behavior of the problem follows similar arguments as above. Now, the optimal solution of the converged subproblems with* $\mathbf{w}$ *as variable are* $\mathbf{w}_*^{(i)}$ *and* $\boldsymbol{\gamma}_*^{(i)}$*. Note that the limiting point* $(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i)}, \boldsymbol{\gamma}_{*|x}^{(i)})$ *is the unique minimizer in the set* $\mathcal{Y}_*^{(i)}$*.*

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

$$f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i)}, \boldsymbol{\gamma}_{*|x}^{(i)}) \leq f(\mathbf{m}_*^{(i)}, \mathbf{w}_0^{(i)}, \boldsymbol{\gamma}_0^{(i)}) \leq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|y}^{(i)}). \quad (50)$$

Let us consider an AO iteration i *in which the optimal value for* $\mathbf{m}$ *and* $\boldsymbol{\gamma}$ *are obtained as* $\mathbf{m}_*^{(i)}$ *and* $\boldsymbol{\gamma}_{*|y}^{(i)}$ *using fixed* $\mathbf{w} = \mathbf{w}_*^{(i-1)}$*. To find* $\mathbf{w}_0^{(i)}$*, we fix* $\mathbf{m}$ *as* $\mathbf{m}_*^{(i)}$ *and optimize for* $\mathbf{w}$*. Since we linearize the convex function in* (45b)*, the fixed operating point is also included in the feasible set* $\{\mathbf{w}_*^{(i-1)}, \mathbf{m}_*^{(i)}, \boldsymbol{\gamma}_{*|y}^{(i)}\} \in \mathcal{Y}_0^{(i)}$ *using* (48)*. Using this, we can show the monotonicity of the objective value sequence as*

$$f(\mathbf{m}_*^{(i)}, \mathbf{w}_0^{(i)}, \boldsymbol{\gamma}_0^{(i)}) \leq f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i-1)}, \boldsymbol{\gamma}_{*|x}^{(i)}).$$

Note that the feasible set follows $\{\mathbf{w}_*^{(i-1)}, \mathbf{m}_*^{(i)}, \boldsymbol{\gamma}_{*|y}^{(i)}\} \in \{\mathcal{X}_*^{(i)} \cap \mathcal{Y}_0^{(i)}\}$ *in each AO iteration.*

C. Uniqueness

The uniqueness of the function mapping, i.e, one-to-one correspondence between the feasible solution and the objective value, can be attributed to the linearization performed on the DC constraint in problem (20) *and the MSE expression in* (28)*. The linearized* (19) *can be written as*

$$\gamma_{l,k,n} + \tilde{\beta}_{l,k,n}^{-2} |\tilde{\mathbf{H}}_{b_k,k,n} \tilde{\mathbf{m}}_{l,k,n}|^2 (\beta_{l,k,n} - \tilde{\beta}_{l,k,n}) - \tilde{\beta}_{l,k,n}^{-1} \tilde{\mathbf{m}}_{l,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n} (\mathbf{m}_{l,k,n} - \tilde{\mathbf{m}}_{l,k,n}) \leq 0, \quad (51)$$

where $\tilde{\mathbf{H}}_{b_k,k,n} = \mathbf{w}_{l,k,n}^{\mathrm{H}} \tilde{\mathbf{H}}_{b_k,k,n}$ *and the MSE expression as*

$$|1 - \mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_k,k,n} \mathbf{m}_{l,k,n}|^2 + \sum_{(j,i) \neq (l,k)} |\mathbf{w}_{l,k,n}^{\mathrm{H}} \mathbf{H}_{b_i,k,n} \mathbf{m}_{j,i,n}|^2 + \widetilde{N}_0 \leq \epsilon_{l,k,n}. \quad (52)$$

Note that the receive beamformer $\mathbf{w}_{l,k,n}$ *is unique for the given set of transmit precoders.*

When the objective function is zero, the uniqueness of the transmit precoders using (51) *and* (52) *are not valid due to the inactive constraints* (16b) *and* (26b)*. To obtain a unique set of transmit precoders when the objective is zero for a single BS, we can regularize the objective with the total transmit power expression without affecting the optimal solution as*

$$\|\tilde{\mathbf{v}}\|_q + \varphi \sum_{k \in \mathcal{U}} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n} \mathbf{m}_{l,k,n}^{\mathrm{H}}\right),$$

for some arbitrarily small value of the scaling factor $\varphi \approx 0$*.*

D. Stationary Point

To show the limiting point of the iterative algorithm is the stationary point of (45)*, it must satisfy the KKT conditions of the nonconvex problem. Since the converged point is a minimizer satisfying*

$$f(\mathbf{m}_*^{(i)}, \mathbf{w}_*^{(i)}, \boldsymbol{\gamma}_{*|x}^{(i)}) = f(\mathbf{m}_*^{(i+1)}, \mathbf{w}_*^{(i)}, \boldsymbol{\gamma}_{*|y}^{(i+1)}) = f(\mathbf{m}_*^{(i+1)}, \mathbf{w}_*^{(i+1)}, \boldsymbol{\gamma}_{*|x}^{(i+1)}), \quad (53)$$

the solution is inside the feasible set $\mathcal{F}$ *and* $(\mathbf{m}_*^{(i+1)}, \mathbf{w}_*^{(i+1)}, \boldsymbol{\gamma}_{*|x}^{(i+1)})$ *is the minimizer of the objective function* f_0 *in the feasible set* $\mathcal{X}_*^{(i+1)} \subset \mathcal{F}$*. Using the discussions in [23], we can easily show that the feasible point* $\mathcal{F}$ *and* $(\mathbf{m}_*^{(i+1)}, \mathbf{w}_*^{(i+1)}, \boldsymbol{\gamma}_{*|x}^{(i+1)})$*, which is the minimizer in the local neighborhood, is a stationary point of the non convex problem in* (45) *satisfying the constraint qualifications and the KKT expressions for the set* $\mathcal{X}_*^{(i+1)} \subset \mathcal{F}$*. The non differentiability of the objective function in* (16) *and* (26) *requires the subdifferential set of the objective function to include* $0 \in \partial f_0(\boldsymbol{\gamma}_*)$ *to satisfy the KKT conditions. The monotonic decrease in the objective is still valid if the variables* $\mathbf{m}, \mathbf{w}$ *and* $\boldsymbol{\gamma}$ *are updated together, since the update in* $\mathbf{w}$ *increases the objective function for a fixed point* $\mathbf{m}$ *and* $\boldsymbol{\gamma}$*. Using these arguments, we can claim that the proposed JSFRA centralized solution achieves a stationary point of the original nonconvex problem with the nonincreasing objective value at each iteration.*

APPENDIX C
KKT CONDITIONS FOR MSE APPROACH

In order to solve for an iterative precoder design algorithm, the KKT expressions for the problem in (42) are obtained by

differentiating the Lagrangian by assuming the equality constraint for (42b) and (42c). At the stationary points, following conditions are satisfied.

$$\nabla_{\epsilon_{l,k,n}} : -\alpha_{l,k,n} + \frac{\sigma_{l,k,n}}{\tilde{\epsilon}_{l,k,n}} = 0 \tag{54a}$$

$$\nabla_{t_{l,k,n}} : -q\, a_k \Big(Q_k - \sum_{n=1}^{N}\sum_{l=1}^{L} t_{l,k,n}\Big)^{(q-1)} + \frac{\sigma_{l,k,n}}{\log_2(e)} = 0 \tag{54b}$$

$$\nabla_{\mathbf{m}_{l,k,n}} : \sum_{y\in\mathcal{U}}\sum_{x=1}^{L} \alpha_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_k,y,n}\mathbf{w}_{x,y,n}\mathbf{w}^{\mathrm{H}}_{x,y,n}\mathbf{H}_{b_k,y,n}\mathbf{m}_{l,k,n} + \delta_b\mathbf{m}_{l,k,n} = \alpha_{l,k,n}\mathbf{H}^{\mathrm{H}}_{b_k,k,n}\mathbf{w}_{l,k,n}, \tag{54c}$$

$$\nabla_{\mathbf{w}_{l,k,n}} : \sum_{(x,y)\neq(l,k)} \mathbf{H}_{b_y,k,n}\mathbf{m}_{x,y,n}\mathbf{m}^{\mathrm{H}}_{x,y,n}\mathbf{H}^{\mathrm{H}}_{b_y,k,n}\mathbf{w}_{l,k,n} + \mathbf{I}_{N_R}\mathbf{w}_{l,k,n} = \mathbf{H}_{b_k,k,n}\,\mathbf{m}_{l,k,n}. \tag{54d}$$

In addition to the primal constraints given in (42b), (42c) and (42d), the complementary slackness criterion must also be satisfied at the stationary point. Upon solving the above expressions in (54) with the complementary slackness conditions, we obtain the iterative algorithm to determine the transmit and the receive beamformers as shown in (43).

REFERENCES

[1] 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.

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

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

[5] 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.

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

[7] 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.

[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] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan 2007.

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

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

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

[23] B. R. Marks and G. P. Wright, "A General Inner Approximation Algorithm for Nonconvex Mathematical Programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

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

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

[26] T. Lipp and S. Boyd, "Variations and extensions of the convex-concave procedure," 2014.

[27] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Advances in neural information processing systems*, 2009, pp. 1759–1767.

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

[29] 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.

[30] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed Robust Multicell Coordinated Beamforming With Imperfect CSI: An ADMM Approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2988–3003, June 2012.

[31] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.

[32] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill New York, 1964, vol. 3.

[33] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Advances in Soft Computing AFSS*. Springer, 2002, pp. 288–300.

[34] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed Methods for Constrained Nonconvex Multi-Agent Optimization – Part I: Theory." [Online]. Available: http://arxiv.org/abs/1410.4754v1

[35] T. D. Quoc and M. Diehl, "Sequential convex programming methods for solving nonlinear optimization problems with DC constraints," *arXiv preprint arXiv:1107.5841*, 2011. [Online]. Available: /http://arxiv.org/abs/1107.5841v1