Reviewers Comments & Authors Replies

Manuscript No.	Paper T-SP-18051-2014, submitted to *"IEEE Transactions on Signal Processing"*
Title	"Traffic Aware Resource Allocation Schemes for Multi-Cell MIMO-OFDM Systems"
Authors	Ganesh Venkatraman, Antti Tölli, Markku Juntti, and Le-Nam Tran

The authors would like to thank the associate editor and the reviewers for their valuable comments on the manuscript of the paper, which have been greatly helpful to improve the paper quality. Based on the comments, we have made several major revisions to the paper, following the suggestions of the reviewers. In what follows, the comments are listed, each followed immediately by the corresponding reply from the authors. The revisions in the revised manuscript are highlighted using blue color and the authors responses are also presented in blue color text. Following is the summary of the revision made on the manuscript in accordance with the reviewers comments.

1. We have provided additional information to improve the continuity in the algorithm formulation from the SINR expression as suggested by the reviewers.
2. We have included the discussion for the MSE based reformulation for different norms in the objective function.
3. We have rewritten the reduced complexity spatial resource allocation (SRA) in Section III-D for better readability.
4. We have shortened the ADMM approach as suggested by the reviewers in Section IV-B.
5. Rigorous convergence analysis is provided for the centralized algorithm in Appendix A on the supplementary document.
6. We have included Section V-C on the queue behavior over multiple time instants. We have added Fig. 4 to show the performance of the proposed scheme over the existing precoder design algorithms.
7. We have addressed the issues in the citations and also provided additional references to prove the convergence of the centralized algorithm as suggested by the reviewers.

Due to the strict page limitation imposed on the resubmitted manuscript, we have included all Appendices on the supplementary document in the manuscript central. In what follows, the comments are listed, each followed by the corresponding reply from the authors. Unless otherwise stated, all the numbered items (figures, equations, references, citations, etc) in this response letter refer to the revised manuscript.

Response to Reviewer - 1's Comments

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.

We thank the reviewer for providing valuable and insightful comments.

1. 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?

Reply: We thank the reviewer for the comment.

   (a) The reformulation for the problem defined in (6) is required, since the SINR expression in (2) cannot be handled directly in the problem. Note that the equality constraint imposed by the SINR expression in (2) is handled by two explicit inequality constraints (16b) and (16c), leading to an approximation for the original problem in (6). We have provided justifications for the problem formulation in (6) and (16) in the paragraph before (16).

   (b) The Q-WSRM scheme and the proposed schemes are all throughput optimal. It can be seen that the proposed extension Q-WSRME and the JSFRA formulations aim at minimizing the number of backlogged packets in addition to avoiding the over allocation of available resources. The JSFRA formulation using ℓ_1 norm as the objective minimizes the number of backlogged packets in a greedy manner at each time instant. By increasing the exponent $q \to \infty$, we obtain fair allocation at every transmission instant. We have included the discussions on the average of number of backlogged packets after each transmission instant for different arrival rates in Section V-C.

   (c) The equivalence between the Q-WSRM scheme and the ℓ_2 norm objective in the JSFRA formulation can be seen when the queue size increases. We have clarified this point in the revised manuscript. Please refer to the first paragraph in Section III-B.

2. 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.

Reply:

   (a) In the previous manuscript, we primarily focused on evaluating the performance of all schemes by comparing the total number of backlogged packets retained in the system after each transmission instant. Since the delay is proportional to the average number of backlogged packets in the system, this implies that the delay performance was indirectly evaluated in the previous manuscript. In addition, the average delay of a particular user can be reduced by controlling the priority factor a_k in (6a), which alters the resources allocated to a particular user. We have included this statement in the manuscript in Section II-B after (6). We can also control the delay by changing the exponent used in the objective ℓ_q of the JSFRA problem as discussed in bullet points before Section III.

(b) We considered the residual number of backlogged packets as a performance measure, since we assume only the instantaneous channel state informations and together with the number of backlogged packets, resources can be allocated only for a given instant.

(c) We agree with the reviewer that the delay should be compared directly to justify the advantages of the proposed schemes. In this regard, we have included a new figure to clarify this point. Please refer to Fig. 4 and the associated text in Section V-C. As can be seen, the proposed methods outperforms the existing schemes.

3. 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?

Reply:

(a) We thank the reviewer for the comment. The convergence analysis has been clearly rewritten in Appendix A on the supplementary document. The discussions are provided for the convergence of the iterative Algorithm 1, *i.e*, the problem defined in (16).

(b) The uniqueness of the convex subproblem (20) can be guaranteed by the linear constraint (19), if the initial feasible operating point is matched, *i.e*, $q_{l,k,n} = \Im\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\} = 0$ while starting the iterative solution. On the other hand, the constraint in (16b) is not unique, since the precoder is inside the absolute value operator. Once the algorithm finds a unique set of transmit precoders, all unitary rotations are also valid for the original problem in (16). The uniqueness of the transmit and the receive beamformers are discussed in detail in Appendix A-C on the supplementary document.

(c) The solution for the problem (16) is not unique, since all unitary transformations on the precoders identified by the iterative algorithm (20) are indeed the solutions for (16).

(d) The uniqueness of the transmit precoders can be guaranteed by adding a strongly convex function in the objective as discussed in Appendix A-C on the supplementary document.

4. 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.

Reply: The discussions on the convergence of the centralized algorithms are provided in Appendix A and on the distributed algorithms in Appendix B. The discussion on the convergence to a stationary point is presented in Appendix A-E on the supplementary document.

5. 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.

Reply: Since the sub-channel wise resource allocation considers each sub-channel at a given time for designing the precoders, the performance of this scheme depend on the selection order of the sub-channels. For instance, to design the precoders for the sub-channel $j+1$, we assume the transmit precoders and the rates of all users are all evaluated up until previous sub-channels $\{1, 2, \ldots, j\}$, considering the normal ordering for selection. At this point, we have to estimate the number of backlogged packets left over in the system, if the sub-channels $\{1, 2, \ldots, j\}$ are transmitted with the precoders designed so far. Now, the number of backlogged packets for the $j+1^{\text{th}}$ sub-channel is given by

$$\max\left(Q_k - \sum_{i=1}^{j}\sum_{l=1}^{L} t_{l,k,i}, 0\right). \tag{1}$$

Since (1) depends on the rates of the already evaluated sub-channels $\{1, 2, \ldots, j\}$, the overall achievable throughput is susceptible to the ordering used to determine the precoders in each sub-channels. It has

been discussed in detail in Section III-D. Note that the performance will be closer to the JSFRA centralized scheme, if we select the best sub-channel ordering through an exhaustive search. In this case, the performance loss is mainly due to the maximum power constraint imposed on each sub-channel.

6. 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?

Reply:

(a) The distributed algorithms are derived for the convex subproblem, which leads to the same stationary point asymptotically as that of the centralized solution, but indeed we would require a large number of iterations for the convergence. In reality, we have to limit the number of iterations required for each distributed algorithm, thereby leading to a point which may not be the same as when the algorithm is allowed to converge. The number of ADMM or primal updates can be controlled by $J_{\max}$ in the Algorithm 2.

(b) Note that the size of the channel state information depends on the number of transmit antennas. Thus exchanging channels is not efficient when the number of transmit antennas is large. On the other hand, the amount of exchanging interference vector only depends on the number of base stations. In practice, we can limit the primal or ADMM update for a few iterations, then the distributed algorithms are still be efficient.

(c) To address this concern, we propose a practical scheme in Section IV-D. Note that in the time-correlated case, it is often enough to update the precoders once per radio frame. The decentralized schemes are not necessary to converge until the end, it is only important to follow the fading process when $J_{\max} = 1$. The performance of the distributed algorithm based on dual decomposition scheme is discussed for the time-correlated fading in Section C of [13], which shows that it is enough for the distributed precoder design to follow the fading process to provide desired performance. The distributed algorithm for the time correlated case is not provided in the current manuscript, since it is not in the scope of the precoder design algorithms considered in this manuscript.

7. 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?

Reply: We have updated the manuscript to include the discussions on the convergence of the distributed algorithm in Appendix B. The distributed algorithm attains the same stationary point as that of the centralized algorithm if we let the inner loop in Algorithm 2 to converge, *i.e*, the dual or the coupling variables are updated until convergence.

8. 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.

Reply:

(a) We have defined all the variables mentioned by the reviewer in the revised manuscript on Section VI-B

(b) Since the coupling between the distributed precoder designs are the interference between the BSs and the users, in the ADMM approach, the interference is treated as a local variable, which is then included in the precoder design problem for each coordinating BS. This is treated as a local variable for a specific BS. Note that the local variable is an assumption made by the BS on the actual interference caused by the neighboring BSs. Since the actual interference caused is different, the consensus has to be made between the local interference variable maintained at each BS with the global consensus interference variable, which is nothing but the average between the corresponding BSs interference. These discussions have been made in the revised manuscript in Section IV-B. For further details we have also referred the interested reader to [11], which discusses exclusively about the ADMM approach.

(c) We have revised the manuscript to avoid the subscript confusions in the ADMM approach.

9. 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)?

Reply: We thank the reveiwer for the valuable comment. For the distributed precoder designs, we assume that each base station (BS) b knows the equivalent downlink channel $\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}$ of all users in the system by using precoded uplink pilots, where the precoders are the MMSE receiver of all the users. Note that it includes the equivalent downlink channels of all the users in the system. To update the MMSE receiver, the equivalent channel for the k^{th} user $\mathbf{H}_{b,k,n}\mathbf{m}_{l,k',n}, \forall k' \in \mathcal{U}_b, \forall b \in \mathcal{B}$ is obtained from the BSs through user specific downlink precoded pilots. We have included the information on what each network entity knows in the paragraph after (36). We have included the type of duplexing scheme adopted in the model, *i.e*, TDD system, in the system description in the last paragraph of Section II.

10. 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.

Reply:

(a) The heuristic part is that instead of updating the involved variables sequentially, we allow all the variables to be updated at the same time. Due to this idea, the convergence is not always guaranteed as in standard methods, but we have observed significantly improved convergence results in all numerical examples.

(b) It follows the same points as that of the centralized approach if the algorithm is iterated in the same order, *i.e* the dual variables are allowed to converge before the update of the fixed operating point and the receiver. Here instead we perform the update of the transmit precoders, receive beamformers and the dual variables all at once in each iteration. Thus, we sacrifice the formal convergence for the improved speed of convergence.

(c) Since the update of the optimization variables are grouped together, it is difficult to prove the monotonicity of the objective function theoretically to prove the convergence of the algorithm. However, in all the numerical experiments in Section V-B, the proposed KKT conditions based algorithm converges, which can be seen from Fig. 3.

(d) We have provided the discussions on the convergence of the KKT based approach in the Appendix B on the supplementary document.

11. 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.

Reply:

(a) We agree with the reviewer comment and is in line with Q.2. Please refer to the response for the comment on delay as the performance metric discusion.

(b) In accordance with the reviewer comment, we have included Section V-C to discuss the queue deviation over multiple transmission slots. We have presented Fig. 4a by comparing the average number of backlogged packets for different algorithms with various arrival rates and Fig. 4b for the number of backlogged packets at each instant. Note that the question on delay arises when we are considering the resource allocation over certain duration. Since the paper is about the precoder design to minimize the number of backlogged packets at each instant, it may not be a valid performance metric for our objective.

(c) We agree that delay can be controlled by reducing the number of packets on an average, as we can see from Fig. 4, the proposed algorithm performs better in comparison with the existing schemes in minimizing the average number of backlogged packets. Note that we can also prioritize the users by controlling the variable a_k in (6a) to address the QoS constraints for a particular user. In addition to that, we can also change the objective to ℓ_2 and ℓ_∞ norm to address the delay and the fairness implicitly. It is included in Section II-B after (6) and in the enumerated listing following (7).

12. What is SRA in the simulation figures?

Reply: SRA stands for spatial resource allocation (SRA). It is updated in the revised manuscript in Section III-D.

13. 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?

Reply: Since the problem is nonconvex, the JSFRA formulation can find a stationary point upon convergence. The converged point of the JSFRA problem is in fact a stationary point of the original nonconvex problem, which is discussed in Appendix A-E provided on the supplementary document. We have removed the statement mentioning the Pareto-optimal solutions in the discussions on Fig. 1 in the revised manuscript.

Response to Reviewer - 2's Comments

We would like to thank the reviewer for providing valuable comments.

1. 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.

Reply: The reformulation is required, since the SINR expression in (2) cannot be handled directly in the problem defined in (6). Note that the equality constraint imposed by the SINR expression in (2) is handled by the two explicit inequality constraints (16b) and (16c), therefore leading to an approximation for the original problem in (6). We have updated the manuscript to include these details for better clarity in the third paragraph in Section III-B.

2. 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.

Reply:

   (a) Please note that the objective function in the problem formulation (16) is convex, since the $\gamma_{l,k,n}$ is now treated as an optimization variable and not as an expression. Since the problem in (16) is not jointly convex on the variables $\mathbf{m}_{l,k,n}$ and $\mathbf{w}_{l,k,n}$, we use alternating optimization (AO) approach by fixing $\mathbf{w}_{l,k,n}$ and optimize for $\mathbf{m}_{l,k,n}$ as a variable.

   (b) Even after fixing $\mathbf{w}_{l,k,n}$ as a constant, the problem in (16) is nonconvex due to the DC constraint (16b), which is handled by the first order relaxation around some fixed operating point. Once the linear relaxation is performed, the problem in (20) is a convex optimization problem with the variables being $\mathbf{m}_{l,k,n}, \gamma_{l,k,n}, \beta_{l,k,n}$. The manuscript is updated to illustrate this clearly in the lines following (19).

3. 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.

Reply: We thank the reviewer for pointing out the inappropriate text. We have removed the incorrect statement on the block coordinate descent method with the AO approach in the paragraph following (16) in Section III-B. Indeed, due to the coupling of the transmit and the receive precoder variables, we cannot use the proof of the standard block coordinate descent method for the convergence of the proposed algorithm as pointed out by the reviewer. We have provided a completely rewritten convergence proof in Appendix A.

4. 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.

Reply: We apologize for the inappropriate reference cited in the original manuscript. We have provided a completely rewritten proof on the convergence of the centralized algorithm in Appendix A on the supplementary document.

5. 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.

Reply: We agree with the reviewer's comment. We have provided the updated proof for the convergence of the centralized algorithm on Appendix A on the supplementary document.

6. 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.).

Reply: We have removed the inappropriate citation of the references, and provided appropriate references for SCA in [22,32,34].

7. 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.

Reply: Please note that the distributed algorithm is performed for the **convex subproblem** presented in (20) and (28). Since the problem is convex, the distributed approach convergence can be guaranteed under certain conditions. Those are discussed in Appendix B on the convergence analysis for the distributed algorithms on the supplementary document.

8. 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.

Reply: We have included the sub-channel wise resource allocation or (SRA) scheme in Section III-D for the completeness. It is presented in the manuscript as an alternative suboptimal approach to perform sub-channel wise distributed precoder design by the centralized controller. We have updated the manuscript with more details. We have also stream-lined the manuscript so that some redundant discussion has been removed here and there. We also improved the grammar and continuity of the paper.

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

Reply: We have included the information regarding the norm used for the simulations in the figure captions of Fig. 1 and Fig. 2.

10. 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.

Reply: We thank the reviewer for pointing out the issue. We have updated the manuscript to include the descriptions in the text to refer the legends used in the figures. Table 1 is provided for the purpose of insight, since it deals with the SISO scenario with 3 sub-channels and 3 users.

11. 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.

Reply:

(a) We thank the reviewer for pointing out the issue in the system model description for the figures. Please note that the simulation scenarios are different for Fig. 2 and Fig. 3 in terms of path loss distribution and the sub-channel numbers.

(b) We have provided different scenario with the centralized algorithm as reference for studying the performance under different system models. The user path loss is distributed uniformly between $[0, -6]$ dB for Fig. 2 and between $[0, -3]$ dB for Fig. 3.

(c) We have updated the manuscript to include these details. In addition, the number of backlogged packets in the system for Fig. 3 is $[9, 16, 14, 16, 9, 13, 11, 12]$ bits. These numbers have been updated in the revised manuscript on the third paragraph of Section V-B.

Response to Reviewer - 3's Comments

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.

We thank the reviewer for reading the manuscript and providing valuable comments. The comments are really helps to improve the manuscript.

1. 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.

Reply: 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 and thus t_k is not required as a variable. However, in the MSE formulation, we have explicitly stated the optimization variable t_k, since it is present in the DC constraint (27).

   (a) (6), (16), (20) and (21) does not depend upon the variable t_k

   (b) (26) and (28) depend on the variable t_k and it is defined as an optimization variable.

2. 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.

Reply: We thank the reviewer for the insightful comment. We have updated the manuscript to include the statement that the proposed approximation in (16b) and (16c) together provides an under-estimator for the SINR expression in (2). This information is highlighted in the third paragraph in Section III-B.

For the constraints (16b) and (16c) to be tight, 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 joint space-frequency resource allocation (JSFRA) formulation should be regularized with the transmit power without affecting the solution as

$$\|\tilde{\mathbf{v}}\|_q + \varphi \sum_{k\in\mathcal{U}}\sum_{n=1}^{N}\sum_{l=1}^{L}\mathrm{tr}\left(\mathbf{m}_{l,k,n}\mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \tag{2}$$

where $\varphi \approx 0$. Note that the modified objective will relax the power constraint by making the constraints (16b) and (16c) tight in the final solution. Note that the tightness discussion is valid only for the active spatial streams, *i.e*, $\mathbf{m}_{l,k,n} \neq \mathbf{0}^{\mathrm{T}}$.

3. The solution in (21) is obtained 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.

Reply: We thank the reviewer for the critical comment. 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 indeed depend on the value of the exponent q used in ℓ_q norm. We have modified the text to include this information on the sentence following after (22).

4. 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!

Reply:

   (a) We thank the reviewer for the comment. Rigorous convergence proof for the proposed algorithm is provided in Appendix A on the supplementary document.

(b) The uniqueness is justified implicitly by the MSE relation in (26c) for the MSE reformulation in (28), which makes the transmit precoders to be susceptible to the phase rotation. In general, the uniqueness of the beamformer can be guaranteed by adding a strongly convex term in the objective as discussed in Appendix A-C.

5. (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 decomposition). 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.

Reply:

(a) We agree that the MSE equivalence with the SINR expression is valid only when the receiver is based on the MMSE objective. In our solution based on MSE reformulation, we have used the MMSE receiver irrespective of the ℓ_q norm used in the objective. The formulation assumes the MMSE receiver for all exponents used in the objective, and therefore, the precoders can be designed by using the MSE relation.

(b) The receivers are based on the MMSE objective, and therefore the equivalence is valid between the MSE and the SINR expression. Since the transmit precoders are updated in accordance with the objective, the MSE reformulation yields the same solution as the SCA approach based JSFRA scheme presented in Section III-B. It can also be verified from Fig. 1 with the ℓ_1 objective. We have updated the manuscript to include these comments in the lines following (25) in Section III-C.

6. 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.

Reply:

(a) We agree with the reviewer's comment. We have included reference [11] that discusses on the valid choices of the step size parameter for ADMM. We have included the statement that the parameter ρ depends on the system model under consideration. This information has been updated on the manuscript in the lines following after (40).

(b) We considered the JSFRA scheme without the MSE reformulation as a representative example to discuss the distributed schemes. Since the rate variables are not included in the optimization problem, we do not include the rate variables t_k as an optimization variables in (37). On the contrary, for the MSE reformulation approach, as suggested by the reviewer, we need to have the rate t_k's as the optimization variables.

Response to Reviewer - 4's Comments

This paper proposes a practical method for minimizing the number of currently backlogged packets in a wireless multi-cell MIMO-OFDM network. Resource allocation is performed over space (beamformers) and frequency (sub-carriers), and a norm of the queue deviations is minimized. The problem studied is very relevant, but still most work in the literature has so far focused on the infinite queue model which clearly does not reflect reality particularly well. The proposed methods seem practical, due to their possibility to be implemented as distributed methods coupled with distributed CSI acquisition. The paper has a multitude of approaches to the problem, but there are several areas which must be improved before a possible publication.

First, we thank the reviewer for providing valuable and insightful comments. The comments are really helpful in improving the manuscript.

1. 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.

Reply: We agree that the original manuscript lacks the convergence proof for the JSFRA approach. We have presented a rigorous convergence proof for the proposed algorithm in Appendix A on the supplementary document. We are grateful to the reviewer for bringing our attention to reference [A] which has been cited as [32] in the revised manuscript. Some of the arguments in [32] are indeed useful to prove the convergence of the proposed algorithms in our paper.

2. 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 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.

Reply:

(a) We thank the reviewer for the suggestion. In this paper, we proposed the precoder design to minimize the number of backlogged packets in the system at a given time. Since the objective is minimized in each time instant, it somehow minimizes the objective on an average as well.

(b) We focused on the design of the precoders based on the current knowledge of the channel state information in the system model.

(c) As suggested by the reviewer, in order to justify the gains over multiple time instants, we have now included Section V-C with Fig. 4 to compare the performance of the JSFRA scheme for different ℓ_q norm over multiple time slots.

(d) Fig. 4(a) demonstrate the performance of the proposed schemes and the existing algorithms using the average of the number of backlogged packets after each transmission instant for different arrival rates. The y-axis denotes the expectation taken over the number of queued packets in the system after each transmission slot.

(e) Fig. 4(b) shows the instantaneous number of backlogged packets of various schemes at each transmission instant. In Fig. 4(b), y-axis shows the total number of backlogged packets in the system after each transmission instant for an average arrival rate of 6 bits for all users. The x-axis represents the transmission slot at which the backlogged packets are evaluated.

3. Third, the distributed methods (at least the primal decomposition and ADMM) seem to be fairly straight-forward 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?

Reply:

(a) In accordance with the reviewer, we have shortened the discussions on the distributed approaches and discussed the convergence proof of the decentralized algorithm in Appendix B.

(b) For the distributed precoder design, we assume that each BS b knows the equivalent downlink channel $\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b,k,n}$ of all users in the system by using precoded uplink pilots, where the precoders are the MMSE receiver evaluated at each user terminal. Note that it includes the equivalent downlink cross channels of the neighbor BS users as well. To update the MMSE receiver, the equivalent channel for the k^{th} user $\mathbf{H}_{b,k,n}\mathbf{m}_{l,k',n}, \forall k' \in \mathcal{U}_b, \forall b \in \mathcal{B}$ is obtained from the BSs through user specific downlink precoded pilots. We have included this discussion in the paragraph after (36).

(c) We have included the type of duplexing scheme adopted in the model, *i.e*, TDD system, in the system description in Section II, last paragraph.

4. 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 sub-carrier. This could easily be clarified.

Reply: We thank the reviewer for the comment. We have clarified the space-frequency terminology in the lines before the last paragraph in Section I.

5. Please see the itemized questions

(a) - p. 1, col. 1, row 42: "userss"

Reply: Corrected in the revised manuscript in Section I first paragraph.

(b) - 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.

Reply: We thank the reviewer for highlighting the insightful point. We have included the statement in the revised manuscript in Section I, second paragraph.

(c) - 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?

Reply: We apologize for the inappropriate reference. We have updated the manuscript to include the proper reference on SCA with [22,32,34].

(d) - p. 2, col. 2, row 36: Write rank(.) and min instead

Reply: We have updated the revised manuscript with the reviewer suggestions on the second paragraph in Section II.

(e) - 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

Reply: We have removed the matrix representation of the transmit precoders and the receive beamformers from the manuscript to avoid the confusion as pointed out by the reviewer.

(f) - 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.

Reply: We thank the reviewer for pointing out the confusion. We have removed the word "combinatorial" from the text on the initial paragraph in Section III.

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

Reply: We have rephrased the sentence to clarify the meaning. The manuscript has been updated with the revised text on the first paragraph in Section III.B.

(h) - (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).

Reply: We have updated the manuscript with the explicit power constraint in (16d) as

$$\sum_{n=1}^{N} \sum_{k \in \mathcal{U}_b} \sum_{l=1}^{L} \mathrm{tr}\left(\mathbf{m}_{l,k,n} \mathbf{m}_{l,k,n}^{\mathrm{H}}\right) \leq P_{\max}, \forall b \tag{3}$$

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

Reply: We thank the reviewer for the comment. We have included the text on proving the NP-hardness of the problem (16). Since the formulation in (16) can be reformulated as a weighted sum rate maximization (WSRM) problem, which is known to be NP-hard [23], therefore, it also belongs to the class of NP-hard problems. This information is included in the third paragraph in Section III-B.

(j) - 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.

Reply: We have removed the sentence mentioning the SCA method. We have updated the manuscript to discuss this as SCA approach in the sentences after (17).

(k) - 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??

Reply: We thank the reviewer for pointing out the mistake. We have included all the constraints except the linearization constraint in (20), which was (19) in the original manuscript. Expression (19d) and (18) are the same. We have removed this double reference from (20).

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

Reply: We have updated the manuscript to avoid the ambiguity in representing the optimal and the MMSE receiver. The optimal receiver is denoted by $\mathbf{w}_{l,k,n}^{o}$ in (22) and the MMSE receiver by $\mathbf{w}_{l,k,n}$ in (23).

(m) - 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?

Reply: We thank the reviewer for the comment. We have updated the manuscript to include this discussion in the sentences following (23). The performance remains the same and the convergence behavior is not affected by this scaling, which can be seen from Fig. 1, which compares the queue deviation objective convergence using the optimal and the MMSE receiver for the JSFRA scheme with ℓ_1 norm.

(n) - 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?

Reply: To solve the nonconvex problem (16), we linearize the DC constraint (16b) around a fixed operating point. Since the operating point is a solution from the earlier iteration, it is also included in the current feasible set. Therefore, at each iteration, the algorithm finds a better solution or the same as compared to the previous solution, which leads to a monotonically decreasing objective. Please refer to Appendix A on the supplementary document for more details on the convergence analysis of the centralized algorithm.

(o) - 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?

Reply:

i. We thank the reviewer for the important comment and the suggestion. It is true that the convex problem need not require a unique solution in general.

ii. The uniqueness of the beamformers for the MSE reformulation in (28) is guaranteed due to the MSE constraint (26c), which is susceptible to the transmit beamformer phase rotations.

iii. On the other hand, the uniqueness of the beamformers for the SCA approach in (20) cannot be guaranteed unless the imaginary part of $q_{l,k,n} = \Im\{\mathbf{w}_{l,k,n}^{\mathrm{H}}\mathbf{H}_{b_k,k,n}\mathbf{m}_{l,k,n}\} = 0$. It can be achieved by using the MMSE receiver for a randomly chosen initial transmit beamformer.

iv. The solution is not unique when the objective is zero even for a single BS. In this case, we can regularize the objective with a strictly convex term as

$$\|\tilde{\mathbf{v}}\|_q + c\,\|\mathbf{x} - \mathbf{x}^{(i)}\|^2 \tag{4}$$

where $\mathbf{x}$ denotes the stacked vector of optimization variables and $\mathbf{x}^{(i)}$ denotes the value of $\mathbf{x}$ in the i^{th} iteration. It has been discussed in the Appendix A-C. It makes the objective strictly convex, therefore, the resulting subproblem has a unique minimizer.

(p) - Table 1: "backpreassure"

Reply: It is updated in the manuscript in Table I.

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

Reply: We thank the reviewer for pointing out the gramatical errors. We have checked the gramatical validity and kept the plural form in some cases.