Dynamic Bandwidth Allocation and Precoding Design for Highly-Loaded Multiuser MISO in Beyond 5G Networks

Thang X. Vu, *Member, IEEE*, Symeon Chatzinotas, *Senior Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

***Abstract*—Multiuser techniques play a central role in the fifth-generation (5G) and beyond 5G (B5G) wireless networks that exploit spatial diversity to serve multiple users simultaneously in the same frequency resource. It is well known that a multi-antenna base station (BS) can efficiently serve a number of users not exceeding the number of antennas at the BS via precoding design. However, when there are more users than the number of antennas at the BS, conventional precoding design methods perform poorly because inter-user interference cannot be efficiently eliminated. In this paper, we investigate the performance of a highly-loaded multiuser system in which a BS simultaneously serves a number of users that is larger than the number of antennas. We propose a dynamic bandwidth allocation and precoding design framework and apply it to two important problems in multiuser systems: i) User fairness maximization and ii) Transmit power minimization, both subject to predefined quality of service (QoS) requirements. The premise of the proposed framework is to dynamically assign orthogonal frequency channels to different user groups and carefully design the precoding vectors within every user group. Since the formulated problems are non-convex, we propose two iterative algorithms based on successive convex approximations (SCA), whose convergence is theoretically guaranteed. Furthermore, we propose a low-complexity user grouping policy based on the singular value decomposition (SVD) to further improve the system performance. Finally, we demonstrate via numerical results that the proposed framework significantly outperforms existing designs in the literature.**

***Index Terms*—Beyond 5G, multiuser MISO, precoding vector, optimization, successive convex approximation, singular value decomposition.**

I. Introduction

Multiple-input multiple-output (MIMO) communications play an essential role in 5G and beyond 5G (B5G) wireless systems to fulfill the rapidly increasing demand for data-hungry applications. By equipping transceivers with multiple antennas, a multi-antenna base station (BS) is able to serve multiple users simultaneously thanks to spatial multiplexing techniques. It is well known that under mild conditions, a BS equipped with N antennas can send independent data streams to a number of users not exceeding N via precoding methods. In this case, the BS can provide adequate degree of freedoms for the users to mitigate inter-user interference [1], [2]. A proper precoding design can improve a multiuser system in terms of both spectral efficiency (SE) and energy efficiency (EE) [3]. The research on multiuser systems has attracted much attention with main focus on the performance analysis based on fundamental physical layer properties. It is shown in [4] that a good SE is achievable in MIMO networks with a limited number of antennas when the pilot sequences of adjacent cells are carefully designed to eliminate pilot contamination impacts [5], [6]. The authors of [7] propose a power control algorithm for maximizing the EE in massive MIMO uplinks under both perfect and imperfect channel state information (CSI) conditions. In [8], the authors analyze the impacts of the number of antennas at the BS on the system EE and conclude that the optimal solution can be derived via a joint resource allocation design. We note that these MIMO system results are obtained in under-loaded conditions in which the number of users is usually smaller than the number of antennas at the BS. In such cases, the BS can simultaneously serve the users in the same frequency band via efficient spatial multiplexing, e.g., multiuser precoding, because the BS has sufficient antennas to mitigate inter-user interference.

The proliferation of mobile handsets and internet-of-thing (IoT) devices in the B5G era has not only imposed stressful requirements for high data rate and stringent latency but also likely put the communication system into overloaded situations. With the number connected devices larger than the global population [9], it is highly probable that the number of users connected to a BS exceeds the number of antennas of the BS. In these cases, the system resources are scarce and exploiting only the spatial diversity may not be sufficient. Consequently, the conventional multiuser precoding design perform poorly because the inter-user interference cannot be efficiently eliminated. To overcome this issue, the authors of [10] propose a joint user selection and precoding design to a subset of users not exceeding the number of antennas at the BS. The proposed method therein, however, cannot provide the required data rate to all the users simultaneously as it serves only a subgroup of users at a time. This asks for novel resource allocation techniques for highly-loaded multiuser systems in the B5G era. This paper develops a resource allocation framework that exploits the frequency dimension together with the spatial multiplexing gain.

A. Related Works

Joint bandwidth and power allocation design has been studied in multi-homing networks. The authors of [11] propose

Manuscript received Feb. 6, 2021; revised Jun. 11, 2021; accepted Aug. 13, 2021. This work is supported by the Luxembourg National Research Fund via projects DISBuS and FlexSAT. The associate editor coordinating the review of this paper and approving it for publication was M. Payaró.

The authors are with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. L-1855 Luxembourg. E-mail: {thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu.

a joint bandwidth and power allocation to improve the EE of multi-homing uplink networks. Therein, the users can select one out of multiple available air interfaces to send data to the BS. A parameter-free solution based on the Charnes-Cooper's transformation is proposed to solve the EE maximization problem. A similar study is considered in [12] to optimize the EE, in which the users select either LTE or WiFi frequency bands to transmit uplink data. A Dinkelbach-based solution is proposed to overcome the fractional formulation of the objective function. Based on the design proposed in [11], [12], a number of subsequent works have been investigated in different scenarios [13]–[18]. The authors of [13] analyze the tradeoff between the spectral and energy efficiencies of cellular networks via dynamic power and bandwidth allocation. Starting with a single-cell use case, the goal therein is to examine the Pareto points of the EE. The resulting optimization minimizes the transmit power subject to resource constraints, assuming a constant transmission rate. Then for a multi-cell use case, a Poisson point process method is used to analyze the inter-cell interference that is independent from the variables. The spectral and energy efficiencies are investigated in [14] via dynamic bandwidth and power allocation for terrestrial systems. By proving the concavity of the rate function with respect to the bandwidth and power variables, closed-form expressions for the optimal bandwidth and power are derived. The authors of [16] study the impacts of joint bandwidth and power control on a geostationary satellite system. Therein, two problems of sum capacity maximization and power minimization are formulated subject to the total bandwidth and power constraints. The performance of such joint resource allocation is evaluated via package loss and transmit power metrics. A matching algorithm is proposed in [17] to solve the resource allocation in multi-homing environments, in which both the users and base stations are equipped with multiple air interfaces. The impact of imperfect channel state information is taken into account in [18] when optimizing the bandwidth assignment in heterogeneous wireless uplinks.

Joint bandwidth and power allocation has also been considered in multiuser downlink systems in [19], [20]. A dynamic frequency and power allocation is proposed in [19] for heterogeneous small cell networks, in which a macro BS provides wireless backhaul to multiple small-cell BSs that serve their users in the same frequency band. In [20], the authors develop an energy-efficient resource allocation scheme for heterogeneous wireless networks in which the users can receive data from both LTE and WiFi orthogonal frequency-division multiplexing (OFDM) signals. Therein, an iterative joint sub-carrier assignment and power control algorithm is proposed by relaxing the binary selection variables followed by a Lagrange dual method. It is worth noting that the above-mentioned works do not consider inter-user interference as each user (or BS) is assigned to an orthogonal frequency channel.

B. Contributions

In this paper, we investigate the performance of a multiuser multiple-input single-output (MISO) system in highly-loaded scenarios in which the number of users can exceed the number of antennas at the BS. Our contributions are summarized as follows:

- We propose a resource allocation framework that fully exploits the spatial diversity and the frequency dimension via a joint design of the precoding vectors and bandwidth allocation in multiuser systems. The effectiveness of the proposed framework is shown via its superior performance compared with the conventional precoding solution [21] and existing joint bandwidth and power allocation strategies [11], [12] in highly- and over-loaded scenarios.
- Two joint optimization problems are formulated to maximize the user fairness and to minimize the transmit power, the two important problems in multiuser systems, subject to minimum quality of service (QoS) requirements, the total bandwidth and a limited transmit power. Unlike previous bandwidth-power allocation designs [11], [12], [14], [15], [17], [19] which avoid inter-user interference by allocating an orthogonal frequency band to each user, our joint optimizations tolerate inter-user interference among users within one group via precoding vectors design. To overcome the non-concavity of the achievable rate (with respect to the bandwidth and precoding vector variables), we reformulate the original problems into a difference-of-convex (DC) representation and propose two iterative algorithms based on the successive convex optimization (SCA) method. The convergence to at least a local optimum of the proposed iterative algorithms is theoretically guaranteed.
- To further enhance the system performance, we propose a heuristic user grouping policy based on the singular value decomposition (SVD) to determine the best user groups partition to which the joint bandwidth and precoding vectors design will be applied.
- Finally, we demonstrate the advantages of the proposed framework via extensive numerical results. It is shown that our solutions achieve more than 50% performance gain over reference schemes proposed in [11], [12], [21].

C. Organization

The remainder of this paper is organized as follows. Section II describes the system model and relevant variables. Section III presents the min rate maximization problem. Section IV derives the joint optimization for the power minimization problem. Section V proposes a low-complexity user grouping policy. Section VI extends the proposed method to a zero-forcing (ZF)-based design. Section VII presents numerical results and demonstrates the effectiveness of the proposed framework. Finally, Section VIII provides conclusions and discussions.

II. System Model

We consider a wireless communications system in which a BS serves multiple users via the shared wireless medium. The BS is equipped with N antennas while the K users, where $K \geq N$, are equipped with a single-antenna. In this paper, we

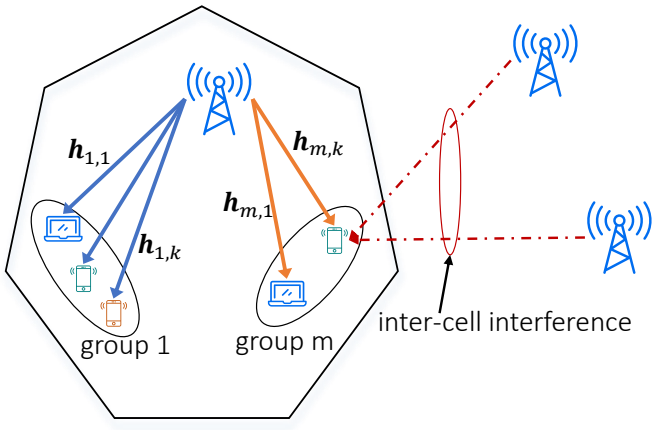


Fig. 1. The multiuser MISO system under highly-loaded conditions. The number of users is greater than the number of antennas at the BS. Interference comes from adjacent cells.

exploit the frequency dimension beside the spatial diversity to optimize the system resources utilization. It is assumed that the frequency channels can be flexibly assigned to the users. We propose a dynamic resource allocation framework that jointly allocates orthogonal frequency channels to different user groups and optimizes the precoding vectors within every user group. In particular, the users are divided into M groups, each having a number of users smaller than N. Let $\mathcal{K}_m$ denote the set of $K_m \leq N$ users in user group m, i.e., $\mathcal{K}_m = \{u_1^m, \ldots, u_{K_m}^m\}$. By definition, $\sum_{m=1}^{M} K_m = K$. The way the user groups are formed (refer to user grouping policy) will be studied in Section V.

Let $\boldsymbol{h}_{m,k} \in \mathbb{C}^{N\times 1}$ denote the channel coefficients from the BS to the k-th user in group m, including the pathloss. In order to mitigate inter-user interference, the BS performs precoding for every group. Denote $\boldsymbol{w}_{m,k} \in \mathbb{C}^{N\times 1}$ as the precoding vector designed for the k-th user in group m. In addition, let b_m be the orthogonal frequency bandwidth allocated for the user group m. The achievable rate of user k in group m is given as follows:

$$r_{m,k} = b_m \log\left(1 + \frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\sum_{k\neq j\in\mathcal{K}_m} |\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m N_0 + I_{out}}\right), \tag{1}$$

where N_0 is the Gaussian noise density and I_{out} is the interference caused by the transmission of neighboring cells which is calculated as $I_{out} = b_m N_I$, where N_I is the average interfering power density [13].

A. Problem Formulation

In this paper, we consider two important problems in multiuser systems: 1) `Problem 1`: Maximizing the fairness and 2) `Problem 2`: Minimizing the transmit power. The generic joint design can be formulated as follows:

$$\underset{\boldsymbol{b},\boldsymbol{w}}{\texttt{Maximize}} \quad f(\boldsymbol{b},\boldsymbol{w}) \tag{2}$$

$$\texttt{s.t.} \quad r_{m,k} \geq \eta_{m,k}, \forall m, \forall k \in \mathcal{K}_m, \tag{2a}$$

$$\sum_{m=1}^{M} b_m \leq B, \tag{2b}$$

$$\sum_{m=1}^{M}\sum_{k=1}^{K_m} \|\boldsymbol{w}_{m,k}\|^2 \leq P_{tot}, \tag{2c}$$

where $r_{m,k}$ is given in (1), $\boldsymbol{b} \triangleq \{b_m\}_{m=1}^{M}$, $\boldsymbol{w} \triangleq \{\boldsymbol{w}_{m,k}\}_{\forall m,k}$ is the short-hand notations for the bandwidth allocation and the precoding vectors, respectively, and $f(\boldsymbol{b},\boldsymbol{w})$ is the generic objective function which is defined as

$$f(\boldsymbol{b},\boldsymbol{w}) = \begin{cases} \min_{m,k}\{r_{m,k}\}, & \text{for } \texttt{Problem 1} \\ -\sum_{m=1}^{M}\sum_{k=1}^{K_m} \|\boldsymbol{w}_{m,k}\|^2, & \text{for } \texttt{Problem 2} \end{cases}.$$

In (2), the first constraint is to satisfy the user QoS requirements, the second constraint states that the aggregated bandwidth allocated to all user groups cannot exceed the total bandwidth B. Orthogonal allocation in frequency domain between different user groups is also ensured by this constraint. Note that although inter-group interference is avoided, there exists inter-user interference among the users with one group. The last constraint limits the transmit power not exceeding P_{tot}. In the next sections, we propose joint bandwidth allocation and optimal precoding vectors designs for these two problems.

III. Maximizing the User Fairness

In this section, we would like to maximize the minimum rate among the users under limited bandwidth and power resources. The optimization problem is formulated as follows:

$$\underset{b_m,\boldsymbol{w}_{m,k}}{\texttt{Maximize}} \quad \min_{m,k} \; b_m \log(1+\gamma_{m,k}) \tag{3}$$

$$\texttt{s.t.} \quad b_m \log(1+\gamma_{m,k}) \geq \eta_{m,k}, \forall m, k, \tag{3a}$$

$$\sum_{m=1}^{M} b_m \leq B_{tot}, \tag{3b}$$

$$\sum_{m=1}^{M}\sum_{k=1}^{K_m} \|\boldsymbol{w}_{m,k}\|^2 \leq P_{tot}, \tag{3c}$$

where $\gamma_{m,k} \triangleq \frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\sum_{k\neq j\in\mathcal{K}_m} |\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I)}$ is the signal to noise plus interference (SINR) of user k in group m. We note that the SINR is also determined by the allocated bandwidth b_m for the user group m.

The problem (3) is non-convex due to the objective function and constraint (3a). Unlike the conventional fairness design [21], in which the max-min rate problem can be equivalently represented via a corresponding max-min SINR that comprises only the precoding vector variables, the considered optimization in (3) jointly optimizes the bandwidth and precoding vectors for all user groups. To add insult to injury, the allocated bandwidth variable b_m also appears in the denominator of the SINR, which makes the problem more difficult.

To overcome this challenge, we will transform problem (3) into a more traceable formulation that can be easier to tackle.

Specifically, by introducing the auxiliary positive variables $x_{m,k}$ and $\gamma_{m,k}$, we can reformulate problem (3) as follows:

$$\underset{b_m,\boldsymbol{w}_{m,k},\gamma_{m,k},x_{m,k}}{\texttt{Maximize}} \quad \min_{m,k} \; x_{m,k} \tag{4}$$

$$\texttt{s.t.} \quad b_m \log(1+\gamma_{m,k}) \geq x_{m,k}, \forall m,k, \tag{4a}$$

$$\frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\sum\limits_{k\neq j\in\mathcal{K}_m} |\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I)} \geq \gamma_{m,k}, \forall m,k, \tag{4b}$$

$$b_m \log(1+\gamma_{m,k}) \geq \eta_{m,k}, \forall m,k, \tag{4c}$$

$$(3b), (3c),$$

where the two new constraints (4a) and (4b) leverage the achievable rate in the objective function of (3), and constraint (4c) is to satisfy the QoS requirements.

The problem (4) is equivalent to problem (3). This can be shown from the fact that the equalities are held in constraints (4a) and (4b) at the optimum.

Because the logarithm function is concave and the function $\frac{1}{x}$ is convex, constraint (4c) can be equivalently transformed into a convex representation by dividing both sides by a positive value b_m. Therefore, the main challenge in solving (4) lies in the first two constraints. To deal with the former, with the aid of the slack variable $y_{m,k}$, we can reformulate constraint (4a) as

$$\log(1+\gamma_{m,k}) \geq y_{m,k}, \tag{5}$$

$$b_m y_{m,k} \geq x_{m,k}. \tag{6}$$

It is observed that constraint (5) is convex since the logarithm function is concave. To deal with constraint (6), we use an equivalent representation:

$$(6) \Leftrightarrow (b_m + y_{m,k})^2 \geq 2x_{m,k} + b_m^2 + y_{m,k}^2, \tag{7}$$

which has a DC representation as both sides are convex functions. This suggests to employ the iterative method which approximates the left-hand-side (LFS) of (6) by its first order approximation. In particular, given feasible solutions $\hat{b}_m$ and $\hat{y}_{m,k}$ in the current iteration, constraint (6) will be approximated in the next iteration by the following convex constraint

$$2(b_m + y_{m,k})(\hat{b}_m + \hat{y}_{m,k}) - (\hat{b}_m + \hat{y}_{m,k})^2 \geq 2x_{m,k} + b_m^2 + y_{m,k}^2. \tag{8}$$

To tackle the non-convexity of constraint (4b), we represent it in an equivalent form as

$$\frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\gamma_{m,k}} \geq \sum_{k\neq j\in\mathcal{K}_m} |\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I). \tag{9}$$

In Appendix A, we show that function $f(x,y) = \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{y}$ is jointly convex in $\boldsymbol{x}$ and y in their supports. Therefore, the LFS of (9) is convex with respect to $\boldsymbol{w}_{m,k}$ and $\gamma_{m,k}$. In addition, the right-hand-side of (9) is also convex as it is a summation of a linear term and square terms. Therefore, iterative method can be used to solve problem (4) in which at each iteration, the constraint (9) is approximated the first-order approximation of the LHS. In particular, given feasible solutions $\hat{\boldsymbol{w}}_{m,k}$ and $\hat{b}_m$ at the t-th iteration, in the $(t+1)$-th iteration, constraint (9) is approximated by a convex constraint below

$$\sum\nolimits_{k\neq j\in\mathcal{K}_m} \boldsymbol{w}_{m,k}^H \boldsymbol{H}_{m,k} \boldsymbol{w}_{m,j} + b_m(N_0+N_I) \leq \frac{2\boldsymbol{w}_{m,k}^H \boldsymbol{H}_{m,k} \hat{\boldsymbol{w}}_{m,k}}{\hat{\gamma}_{m,k}} - \gamma_{m,k} \frac{\hat{\boldsymbol{w}}_{m,k} \boldsymbol{H}_{m,k} \hat{\boldsymbol{w}}_{m,k}}{\hat{\gamma}_{m,k}^2}, \tag{10}$$

where $\boldsymbol{H}_{m,k} \triangleq \boldsymbol{h}_{m,k}\boldsymbol{h}_{mk}^H$.

By using (8) and (10) as the inner approximations of (6) and (9), respectively, the original optimization problem (4) can be approximated by

$$\mathcal{P}_1(\hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{\gamma}}): \quad \underset{b_m,\boldsymbol{w}_{m,k},x_{m,k},y_{m,k},\gamma_{m,k}}{\texttt{Maximize}} \quad \min_{m,k} x_{m,k} \tag{11}$$

$$\texttt{s.t.} \quad (2b), (2c), (5), (8), (10),$$

$$\log(1+\gamma_{m,k}) \geq \frac{\eta_{m,k}}{b_m}, \tag{11a}$$

where (11a) is directly obtained from (4c) by dividing both sides by a positive number b_m.

We observe that problem (11) is convex, hence it can be efficiently solved by standard methods, e.g., the interior point method [22]. Because the solutions of (11) also satisfy all the constraints (3a), (3b) and (3c), it provides a (sub) optimal solution for (3). It is worth noting that the solution of (11) largely depends on the initial values $\hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{\gamma}}$. Therefore, to close the gap to the global optimum of (3), we propose an iterative algorithm that consists of solving a sequence of convex optimization problems. The idea behind the proposed iterative algorithm is to obtain better approximated solutions through iterations. The steps of the proposed algorithm are listed in Algorithm 1.

Algorithm 1 ITERATIVE ALGORITHM TO SOLVE (3)

1: Initialize a feasible solution $\hat{\boldsymbol{w}}_{m,k}, \hat{b}_m, \hat{\gamma}_{m,k}, \hat{y}_{m,k}$, ϵ, $t = 1$, R_{old}, t_{MAX} and $\texttt{error} = 1$.
2: **while** $\texttt{error} > \epsilon$ and $t < t_{MAX}$ **do**
3: Solve problem $\mathcal{P}_1(\hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{\gamma}})$ in (11) to obtain $\boldsymbol{w}_{m,k}^\star, b_m^\star, \gamma_{m,k}^\star, x_{m,k}^\star, y_{m,k}^\star$.
4: Compute $R^{(t)} = \min_{m,k}\{x_{m,k}^\star\}$.
5: Compute $\texttt{error} = |R^{(t)} - R_{old}|$
6: Update $\hat{\boldsymbol{w}}_{m,k} = \boldsymbol{w}_{m,k}^\star; \hat{b}_m = b_m^\star$, $\hat{y}_{m,k} = y_{m,k}^\star$, $\hat{\gamma}_{m,k} = \gamma_{m,k}^\star, R_{old} = R^{(t)}$, $t := t+1$.

A. Convergence of Algorithm 1

Proposition 1: The sequence of the objective values generated by Algorithm 1 in solving the problem $\mathcal{P}_1(\hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{\gamma}})$ is non-decreasing.

The proof of Proposition 1 is shown in Appendix B. Although not guaranteeing the global optimal solution of the problem (3), Proposition 1 justifies the convergence to at least a local optimum of the proposed iterative algorithm. The study of the performance gap to the global optimum is left for future work.

B. Complexity of the Proposed Algorithm

The complexity of the proposed Algorithm 1 is determined by the computation complexity of one iteration and the number

of iterations. Assuming that the interior point method is used to solve the convex problem (11), in the worst case the complexity is equal to the cube of the number of real variables [22]. Since there are $2NK + 3K + M$ real variables in the problem (11), the computational complexity for solving (11) is $(2NK + 3K + M)^3$. As the result, it requires the overall complexity of $t_{MAX}(2NK + 3K + M)^3$ to solve the proposed Algorithm 1, where t_{MAX} is the maximum number of iterations.

IV. Minimizing the Transmit Power

In this section, we aim to serve the users with given QoS requirements for the least energy consumption. In particular, the BS must provide an achievable rate to every user k in group m that is not less than the QoS target $\eta_{m,k}$. The power minimization problem is formulated as follows:

$$\underset{b_m, \boldsymbol{w}_{m,k}}{\texttt{Minimize}} \quad \sum_{m=1}^{M}\sum_{k=1}^{K_m} \|\boldsymbol{w}_{m,k}\|^2 \tag{12}$$

$$\texttt{s.t.} \quad b_m \log\Big(1 + \frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\sum_{k\neq j\in\mathcal{K}_m}|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I)}\Big) \geq \eta_{m,k}, \forall m, \forall k \in \mathcal{K}_m, \tag{12a}$$

$$\text{(2b), (2c)},$$

where we have used (1) for the achievable rate in the QoS constraint.

In problem (12), the first constraint is to satisfy the minimum QoS requirement for every user, the second and the third constraints are to satisfy the total bandwidth and the power budget, respectively.

Although having a convex objective function, solving problem (12) is difficult due to the non-convexity of constraint (12a) that involves both the precoding vectors $\boldsymbol{w}_{m,k}$ and bandwidth allocation b_m. To tackle this challenge, we reformulate constraint (12a) as follows:

$$\log(1+\gamma_{m,k}) \geq \frac{\eta_{m,k}}{b_m}, \tag{13}$$

$$\frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\sum_{k\neq j\in\mathcal{K}_m}|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I)} \geq \gamma_{m,k}, \tag{14}$$

where $\gamma_{m,k}$ is the auxiliary variable representing the SINR at user k in group m.

It is observed that constraint (13) is convex as the logarithm function is concave and the function $\frac{1}{x}$ is convex. Therefore, the main challenge lies in (14). Similar to the previous section, we will transform this constraint into a DC representation, which can be effectively solved via iterative algorithm.

Indeed, because both the variables are positive, (14) is equivalent to the following constraint:

$$\frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}|^2}{\gamma_{m,k}} \geq \sum_{k\neq j\in\mathcal{K}_m}|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}|^2 + b_m(N_0+N_I). \tag{15}$$

Accordingly Appendix A, the LHS of (15) is jointly convex in $\boldsymbol{w}_{m,k}$ and $\gamma_{m,k}$. Because the RHS of (15) is also convex, we can use (16) as the approximation of the original constraint

Algorithm 2 Iterative Algorithm to solve (12)

1: Initialize a feasible solution $\hat{\boldsymbol{w}}_{m,k}, \hat{\gamma}_{m,k}$, ϵ, $t = 1$, E_{old}, t_{MAX} and $\texttt{error} = 1$.
2: **while** $\texttt{error} > \epsilon$ and $t < t_{\text{MAX}}$ **do**
3: Solve problem $\mathcal{P}_2(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\gamma}})$ in (17) to obtain $\boldsymbol{w}_{m,k}^\star, b_m^\star, \gamma_{m,k}^\star$.
4: Compute $E^{(t)} = \sum_{m,k}\|\boldsymbol{w}_{m,k}^\star\|^2$.
5: Compute $\texttt{error} = |E^{(t)} - E_{old}|$
6: Update $\hat{\boldsymbol{w}}_{m,k} = \boldsymbol{w}_{m,k}^\star; \hat{b}_m = b_m^\star$, $E_{old} = E^{(t)}$, $t := t+1$.

(15), where $\hat{\boldsymbol{w}}_{m,k}$ and $\hat{b}_m$ are feasible solutions of the previous iteration.

$$\sum_{k\neq j\in\mathcal{K}_m} \boldsymbol{w}_{m,k}^H \boldsymbol{H}_{m,k} \boldsymbol{w}_{m,j} + b_m(N_0+N_I) \leq \frac{2\boldsymbol{w}_{m,k}^H \boldsymbol{H}_{m,k}\hat{\boldsymbol{w}}_{m,k}}{\hat{\gamma}_{m,k}} - \gamma_{m,k}\frac{\hat{\boldsymbol{w}}_{m,k}\boldsymbol{H}_{m,k}\hat{\boldsymbol{w}}_{m,k}}{\hat{\gamma}_{m,k}^2}. \tag{16}$$

Therefore, we propose an iterative algorithm that solves problem (12) via a sequence of convex problems. The steps of the proposed algorithm are presented in Algorithm 2.

$$\mathcal{P}_2(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\gamma}}): \quad \underset{b_m, \boldsymbol{w}_{m,k}, \gamma_{m,k}}{\texttt{Minimize}} \quad \sum_{m=1}^{M}\sum_{k=1}^{K_m}\|\boldsymbol{w}_{m,k}\|^2 \tag{17}$$
$$\texttt{s.t.} \quad \text{(2b), (2c), (13), (16)}.$$

Proposition 2: The sequence of the objective values generated by Algorithm 2 in solving the problem $\mathcal{P}_2(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\gamma}})$ is non-decreasing.

The proof of Proposition 2 can be obtained in a similar way as the proof of Proposition 1.

V. User Grouping Optimization

The proposed joint bandwidth and precoding vectors optimization algorithms in previous sections work on a predefined user groups, e.g., $\mathcal{K}_1, \ldots, \mathcal{K}_m, \ldots, \mathcal{K}_M$. As the optimal precoding vectors depend on the channel gains of each user group, it is vital to find appropriate groups partition, which is called as the user grouping policy, to maximize the overall system performance. In this section, we aim at finding the best user groups partition $\{\mathcal{K}_m\}_{\forall m}$ for optimizing the joint bandwidth and precoding vectors design. Denote $\boldsymbol{\mathcal{K}} \triangleq \{\mathcal{K}_1, \ldots, \mathcal{K}_M\}$ as a realization of user grouping. The generic optimal joint user grouping, bandwidth allocation and precoding vectors design can be formulated as follows:

$$\underset{\boldsymbol{\mathcal{K}}, \boldsymbol{b}, \boldsymbol{w}}{\texttt{Maximize}} \quad f(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\mathcal{K}}) \tag{18}$$
$$\texttt{s.t.} \quad r_{m,k} \geq \eta_{m,k}; \quad \sum\nolimits_{m=1}^{M} b_m \leq B;$$
$$\sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K_m}\|\boldsymbol{w}_{m,k}\|^2 \leq P_{tot},$$

where $\boldsymbol{b} = \{b_m\}_{m=1}^M$ and $\boldsymbol{w} = \{\boldsymbol{w}_{m,k}\}_{\forall m,k}$ are the short-hand notations for the bandwidth allocation variables and the precoding vectors, respectively, and $f(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\mathcal{K}})$ is the generic objective function. For the power minimization problem $f(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\mathcal{K}}) = -\sum_{m=1}^{M}\sum_{k=1}^{K_m}\|\boldsymbol{w}_{m,k}\|^2$, while in the

minimum rate maximization problem we have $f(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\mathcal{K}}) = \min_{m,k} r_{m,k}$.

Solving (18) is challenging not only due to the non-concavity of the rate function (with respect to the bandwidth and precoding vector variables) but also the combinatorial property of user grouping in $\boldsymbol{\mathcal{K}}$. More specifically, for every given user grouping $\{\mathcal{K}_1, \ldots, \mathcal{K}_M\}$, the resulting problem of (18) is still non-convex and has to be solved iteratively as in Section III and IV. For example, for a system with $N = 10$ antennas, $K = 18$ users, and $M = 2$ there are more than 136000 groupings, for each of which we need to run the iterative algorithm. This expensive computational complexity may limit the optimal method from using in practical scenarios which usually demand timely decision making actions.

One possible alternative is to deploy machine learning-based solutions [23], [24] to predict the best user groups, whose central idea is to establish underlying relation between the full channel state information and the best user group partition. Supervised learning method is used to train the neural network (NN) parameters over the training set, which consists of labeled input-output pairs. When the number of training samples are sufficiently large, the trained NN is robust and can efficiently generalize the relation between the best user grouping and the channel state information. However, in order to generate the sufficient training samples, it requires considerable computations, as it needs to run the iterative algorithm for all user groups for every training sample.

To tackle the high computation issue of the selection process in (18), we propose a singular value decomposition (SVD)-based user grouping which does not require running any iterative algorithm. Our proposed user grouping policy originates from the observation that the parallelism of a multiuser MISO channel heavily depends on the singular values of the channel matrix. In fact, the larger the singular values are, the higher rate the users can achieve and the vice versus. This observation ignites a grouping policy based on the singular values of the channel matrix of each user group. Before introducing detailed steps of the proposed user grouping policy, we denote $\bar{\boldsymbol{H}}_m = \frac{\boldsymbol{H}_m}{\sigma_\Sigma}$ as the effective channels matrix, where $\boldsymbol{H}_m \in \mathbb{C}^{K_m \times N}$ is the channel matrix from the BS's antennas to the users in group $\mathcal{K}_m$ and $\sigma_\Sigma^2 = B(N_0 + N_I)$ is the total Gaussian noise and inter-cell interference. It is worth noting that the effective channel matrix $\boldsymbol{H}_m$ takes into account the geographical distribution of users as it contains the pathloss and total noise plus inter-cell interference. The SVD of $\bar{\boldsymbol{H}}_m$ is given as:

$$\bar{\boldsymbol{H}}_m = \boldsymbol{U}_m \boldsymbol{D}_m \boldsymbol{V}_m^{'}, \tag{19}$$

where $\boldsymbol{D}_m$ is a diagonal matrix containing the singular values of $\bar{\boldsymbol{H}}_m$, and $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$ are unitary matrices of corresponding dimensions. Let $[\lambda_1^m, \ldots, \lambda_{K_m}^m]$ denote the singular values of $\boldsymbol{H}_m$. Furthermore, we denote

$$\boldsymbol{\lambda} = [\lambda_1^1, \ldots, \lambda_{K_1}^1, \ldots, \lambda_1^M, \ldots, \lambda_{K_M}^M]$$

as the vector containing all singular values of all user groups in the grouping $\boldsymbol{\mathcal{K}} \triangleq \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_M\}$.

We propose to select the user groups based on the singular values of the users in all groups. In particular, for each user groups partition $\boldsymbol{\mathcal{K}}$ we compute the ratio between the mean singular value and the minimum value in $\boldsymbol{\lambda}$ and select the groups partition with the largest score, which is formulated as

$$\mathcal{K}^\star = \arg\max_{\boldsymbol{\mathcal{K}}} \frac{\texttt{mean}(\boldsymbol{\lambda})}{\min(\boldsymbol{\lambda})}, \tag{20}$$

where $\texttt{mean}(\boldsymbol{\lambda}) = \frac{1}{K}\sum_{m=1}^{M}\sum_{k=1}^{K_m}\lambda_k^m$. The insight of the formulation in (20) is to select the grouping that minimizes the divergence among the singular values of all users.

Once the selection in (20) is carried out, the optimal groups $\{\mathcal{K}_1^\star, \ldots, \mathcal{K}_M^\star\}$ are determined. Then, we apply Algorithm 1 and 2 for the min rate maximization and power minimization problems, respectively.

VI. Special Case: Zero-forcing based Joint Design

In particular scenarios, obtaining the optimal precoding vectors is unfavorable due to, e.g., stringent latency requirements. In such cases, a low-complexity design based on the ZF beamforming is preferred. More specifically, the direction of the precoding vectors are defined by the ZF beamformers, therefore only the magnitudes of the precoding vectors need to be optimized jointly with the bandwidth allocation. Denote $\boldsymbol{W}_m = \boldsymbol{H}_m^T(\boldsymbol{H}_m\boldsymbol{H}_m^T)^{-1}$ as the ZF-beamforming matrix of the m-th user group and denote $\bar{\boldsymbol{w}}_{m,k}$ as the k-th column of $\boldsymbol{W}_m$. In the ZF-based precoding design, the precoding vector designed for user k in group m is $\boldsymbol{w}_{m,k}^{ZF} = \sqrt{p_{m,k}}\bar{\boldsymbol{w}}_{m,k}$, where $p_{m,k}$ is the power scaling factor. By definition we have $\boldsymbol{h}_{m,k}^T\bar{\boldsymbol{w}}_{m,j} = \delta_{j,k}, \forall k, j, m$, where $\delta_{j,k}$ equals to 1 if $k = j$ and 0 otherwise. As the result, the achievable rate for user k in group m under the ZF design is given as

$$\begin{aligned} r_{m,k}^{ZF} &= b_m \log\Big(1 + \frac{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,k}^{ZF}|^2}{\sum\limits_{k\neq j\in\mathcal{K}_m} |\boldsymbol{h}_{m,k}^H \boldsymbol{w}_{m,j}^{ZF}|^2 + b_m(N_0 + N_I)}\Big) \\ &= b_m \log\left(1 + \frac{p_{m,k}}{b_m(N_0 + N_I)}\right). \end{aligned} \tag{21}$$

Denote $\alpha_{m,k} \triangleq \|\bar{\boldsymbol{w}}_{m,k}\|^2$, the total transmit power under the ZF-based design is $P_{ZF} = \sum_{m=1}^{M}\sum_{k=1}^{K_m} p_{m,k}\alpha_{m,k}$. Then the min rate maximization problem under the ZF design can be formulated as

$$\underset{b_m, p_{m,k}}{\texttt{Maximize}} \ \min_{m,k} \ b_m \log\left(1 + \frac{p_{m,k}}{b_m(N_0 + N_I)}\right) \tag{22}$$

$$\texttt{s.t.} \ \ b_m \log\left(1 + \frac{p_{m,k}}{b_m(N_0 + N_I)}\right) \geq \eta_{m,k}, \forall m, k, \tag{22a}$$

$$\sum\nolimits_{m=1}^{M} b_m \leq B_{tot}, \tag{22b}$$

$$\sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K_m} \alpha_{m,k}p_{m,k} \leq P_{tot}, \tag{22c}$$

and the power minimization problem is formulated as

$$\begin{aligned} \underset{b_m, p_{m,k}}{\texttt{Minimize}} \ & \sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K_m} \alpha_{m,k}p_{m,k} \\ \texttt{s.t.} \ & \text{(22a), (22b).} \end{aligned} \tag{23}$$

Lemma 1: The rate function under the ZF design in (21) is jointly concave in b_m and $p_{m,k}$.

Proof: Consider a function $g(x,y) = x\log(1+y/x)$ in $\mathbb{R}_+^2$. The Hessian matrix of $g(x,y)$ is given as follows:

$$\mathcal{H}_g = \begin{bmatrix} \frac{-y^2}{x(x+y)^2} & \frac{y}{(x+y)^2} \\ \frac{y}{(x+y)^2} & -\frac{x}{(x+y)^2} \end{bmatrix}. \quad (24)$$

For arbitrary vector $\boldsymbol{x} = [a\ b]^T$, we can calculate

$$\boldsymbol{x}^T\mathcal{H}_g\boldsymbol{x} = -\frac{(ay-bx)^2}{x^2(x+y)^2}, \quad (25)$$

which is always less than or equal to zero. This implies that the function $x\log(1+y/x)$ is concave in its supports. From (21) we can write $r_{m,k}^{ZF} = \frac{g(b_m(N_0+N_I),p_{m,k})}{(N_0+N_I)}$, which completes the proof of Lemma 1. ∎

Lemma 1 states that the constraint (22a) is convex. Consequently, problem (23) is convex as the objective function and constraint (22b) are linear. Similarly, we can see that all constraints of (22) are convex. Furthermore, as the min function preserves the concavity, problem (22) is also convex. The formulation in (22) has a similar form as in [11], [12], although the work in [11], [12] does not consider the precoding design.

VII. Numerical results

In this section, we present numerical results to validate the proposed joint resource allocation framework and compare to existing solutions. The parameters used in the simulations are summarized in Table I. In order to focus on highly- and overloaded scenarios, the parameters are chosen to satisfy $\frac{K}{N} \geq 1$. In order to efficiently exploit the spatial multiplexing gain, the number of user groups is equal to $\lfloor \frac{K}{N} \rfloor + 1$. The users are randomly located between 100 and 1000 meters from the BS. The simulation results are average over 200 random channel realizations.

We compare the proposed framework with following references:

- *Baseline 1*: joint bandwidth and power allocation assuming no inter-user interference. This scheme was first proposed in [11], [12] and then deployed by other works in [14], [15], [17], [19]. In this scheme, each user is assigned an orthogonal frequency channel and the transmit power for each user is jointly optimized with its frequency channel.
- *Baseline 2*: the joint design in [11], [12] plus spatial diversity via ZF-based power allocation. We note that [11], [12] did not consider the precoding design, but their method can directly be applied to the ZF-based precoding context.
- *Baseline 3*: conventional multiuser precoding method [21], in which all the users share the channel bandwidth.
- *Baseline 4*: The users are time-slotted into groups, and conventional multiuser precoding method [21] is deployed in each group.

Furthermore, in order to evaluate the effectiveness of the optimal precoding design and the proposed user grouping policy, two more schemes are presented:

TABLE I
Simulation parameters

Parameters	Value
Cell radius	1000 m
Number of antennas N	Vary from 5 to 8
Number of users K	Vary and greater than N
User location	Uniformly distributed
QoS $\eta_k = \eta, \forall k$	Vary between 10 Mbps and 40 Mbps
Channel bandwidth B	20 MHz
Thermal noise density N_0	10^{-12} W/Hz
Interfering density N_I	$9N_0$

- *Proposed precoding - random grouping*: The user groups are randomly selected, followed by the proposed joint designs in Algorithm 1 and 2.
- *Proposed grouping - ZF*: The user groups are selected based on the solution in Section V, followed by the joint bandwidth and power allocation based on the ZF design in Section VI. It is noted that in this case, the number of users per group should not exceed the number of antennas at the BS.

A. Initialization

Because the proposed Algorithm 1 and 2 work in an iterative manner, they require initial values used in the inner approximation. Initialization is an important step in any iterative algorithm that affects the feasibility of the optimization. A common method for initialization is to solve the feasible problem, which searches for feasible solutions that satisfy all the constraints. The feasibility problem can be formulated as follows:

$$\underset{b_m,\boldsymbol{w}_{m,k}}{\texttt{Maximize}}\ \ 1 \quad (26)$$

$$\texttt{s.t.}\ \ b_m\log\Big(1+\frac{|\boldsymbol{h}_{m,k}^H\boldsymbol{w}_{m,k}|^2}{\sum_{j\neq k}|\boldsymbol{h}_{m,k}^H\boldsymbol{w}_{m,j}|^2+b_m(N_0+N_I)}\Big) \geq \eta_{m,k}, \forall m,k, \quad (26a)$$

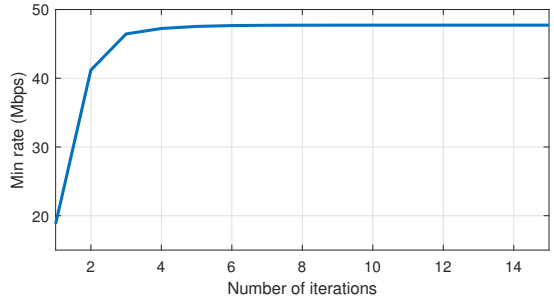
$$\sum_{m=1}^{M} b_m \leq B_{tot},\ \ \sum_{m=1}^{M}\sum_{k=1}^{K_m}\|\boldsymbol{w}_{m,k}\|^2 \leq P_{tot}. \quad (26b)$$

Unfortunately, the feasible problem is still difficult to solve due to constraint (26a). Solving (26) directly requires an inner approximation for (26a), which also results in an iterative solution. Therefore, to speed up the initialization step, we predetermine the direction of the precoding vectors based on the ZF design, and find feasible magnitudes of the precoding vectors and the bandwidth for initialization . This can be done by solving the convex problem (22) without the objective function[1]. From the output of this problem, we can compute the required initial values used in Algorithm 1 and 2.

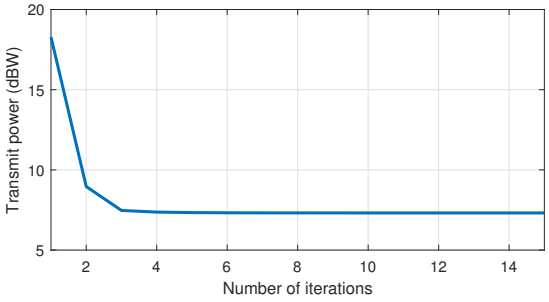
B. Convergence of the Proposed Iterative Algorithms

Fig. 2 shows the objective function values of the proposed iterative algorithms for the min rate maximization and the power

[1]An alternative solution for initialization is to search for valid precoding vectors for a given bandwidth allocation $\{b_m\}_{\forall m}$ and then heuristically adjusting $\{b_m\}_{\forall m}$ until a feasible solution is found. This method, however, converges slower than the ZF-based initialization solution.



(a) Algorithm 1: Min rate maximization



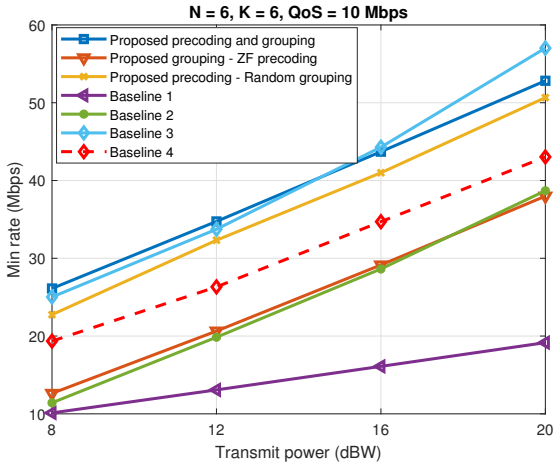
(b) Algorithm 2: Transmit power minimization

Fig. 2. Convergence of the proposed iteration algorithms.

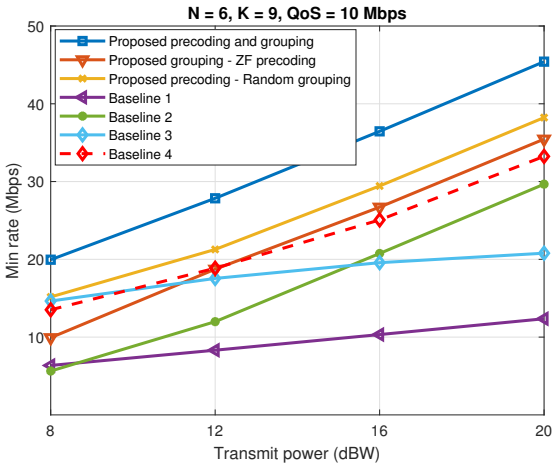
minimization as a function of the iterations when $K = 7$, $N = 6$ and a QoS of 20 Mbps from all users. In Fig. 2a, the total transmit power of the BS is 20 dBW. We observe that both the iterative algorithms quickly converge to the (sub) optimal values after less than 10 iterations. It is shown that Algorithm 1 can double the min rate after 8 iterations, while Algorithm 2 reduces 50% the transmit power, which justifies the effectiveness of the proposed iterative algorithms.

C. Min Rate Maximization Performance

Fig. 3 plots the minimum achievable rate as a function of the total transmit power. The minimum rate requirement of every user is 10 Mbps. For each channel realization, we first apply the proposed SVD-based grouping policy in Section V to determine the user groups. Then we execute the proposed Algorithm 1 to find the optimal bandwidth allocation and precoding vectors. Fig. 3a shows the min rate performance when the system is highly-loaded, i.e., $K = N$. In this case, the BS has a sufficient number of antennas to eliminate inter-user interference. As a result, the conventional precoding design [21] performs efficiently, which is revealed in a good performance of the baseline 3. Interestingly, the proposed design outperforms the baseline 3 when the transmit power is tight. It is also shown that the proposed design significantly outperforms the baselines 1, 2 and 4 and outperforms the baseline 3 when the transmit power of BS is small. Compared to the ZF-based method, the proposed precoding design improves the min rate by 12 Mbps for the same user grouping. This gain results from the fact that the optimal precoding design eliminates inter-user interference more efficiently than the ZF-based method. Another important observation is that the precoding vectors have more influence on the overall



(a) Highly-loaded scenario



(b) Over-loaded scenario

Fig. 3. Min rate versus the total transmit power: (a) - the system is highly-loaded with $K = N$ and (b) - the system is over-loaded with $K > N$.

performance than the user grouping selection, which is shown via a larger gap of the proposed design to the *ZF precoding* curve than the gap to the *Random grouping* curve.

The effectiveness of the proposed design is clearly demonstrated in Fig. 3b when the system is over-loaded, i.e., $K > N$. In this case, exploiting only the spatial diversity is not sufficient because the BS does not have adequate antennas to mitigate inter-user interference, which is shown via a poor performance of the conventional precoding solution (baseline 3). On the other hand, the proposed design achieves the highest rate which is proportional to the transmit power. More importantly, the performance gain of the proposed design compared with other schemes is preserved for all the transmit power values.

Fig. 4 presents the min rate versus different numbers of antennas at the BS, with $K = 9$ users. We note that the total transmit power is fixed at 15 dBW. As expected, the proposed design significantly outperforms other schemes. More importantly, the proposed design successfully exploits the spatial diversity as it achieves a min rate which increases along with the number of antennas. On the other hand, the ZF-based

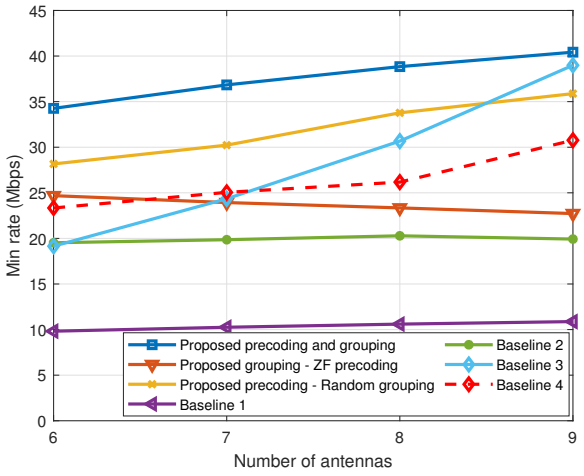


Fig. 4. Min rate versus the number of antennas. There are $K = 9$ users. $P_{tot} = 15$ dBW.

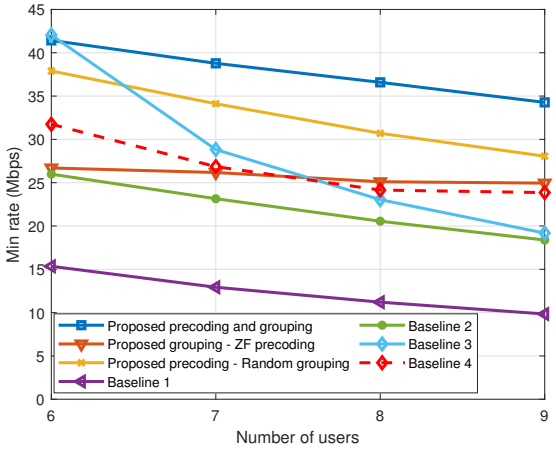


Fig. 5. Min rate versus the number of users. The BS is equipped with $N = 6$ antennas and transmits at $P_{tot} = 15$ dBW.

design, baseline 1 and baseline 2 fail to grasp the benefit of having more antennas. This is because the ZF design spends the energy inefficiently to completely suppress the inter-user interference.

Fig. 5 shows the impact of the number of users on the achievable min rate. The BS is equipped with 6 antennas and has a power budget of 16 dBW. Having more users results in a lower achievable min rate, which is because the power has to be shared among more users. However, the proposed design maintains relative gains compared with the reference schemes, which demonstrates the effectiveness of our joint design.

Fig. 6 compares the achievable min rate of the proposed design with the references for different number of users while keeping the same level of overload, i.e., $\frac{K}{N} = 1.5$. It is shown that the performance gains of the proposed design are preserved, which justifies the advantage of the proposed algorithm, although the gaps between curves may slightly vary. It is also observed that the performance of baseline 2 and the proposed grouping with ZF precoding degrades more quickly than other schemes as K increases. This is because these two

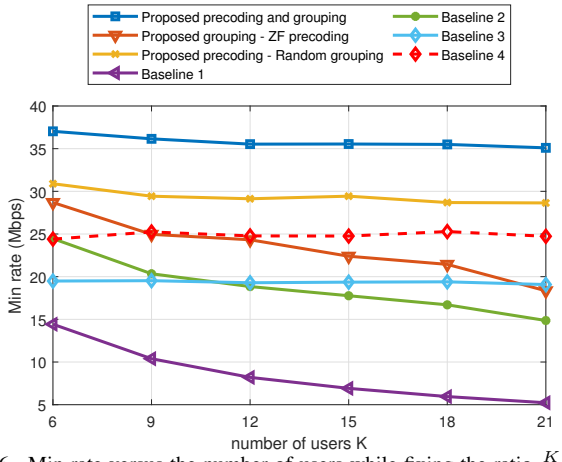


Fig. 6. Min rate versus the number of users while fixing the ratio $\frac{K}{N} = 1.5$, $P_{tot} = 16$ dBW.

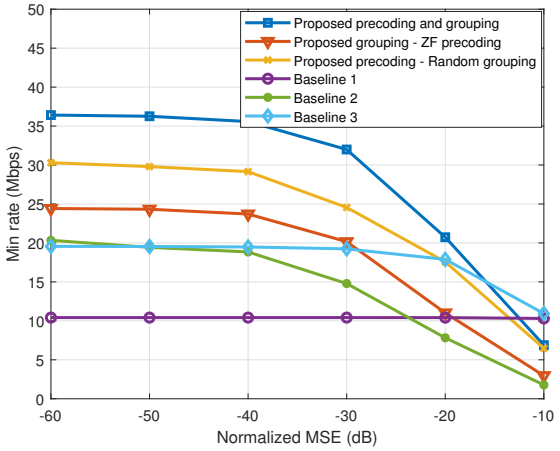


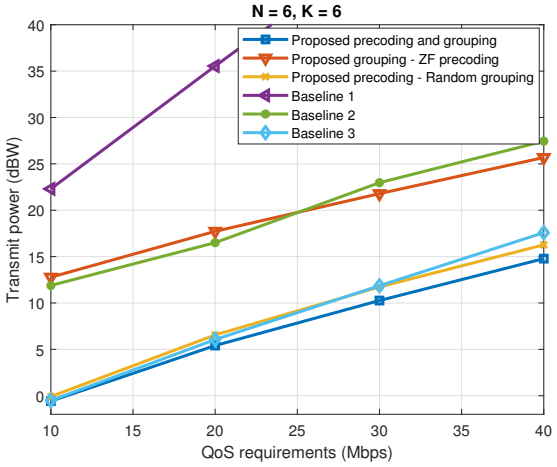
Fig. 7. Min rate versus the channel estimation errors. The BS is equipped with $N = 6$ antennas and transmits at $P_{tot} = 16$ dBW.

schemes completely mitigate inter-user interference via ZF-based power allocation, while the total transmit power is fixed.

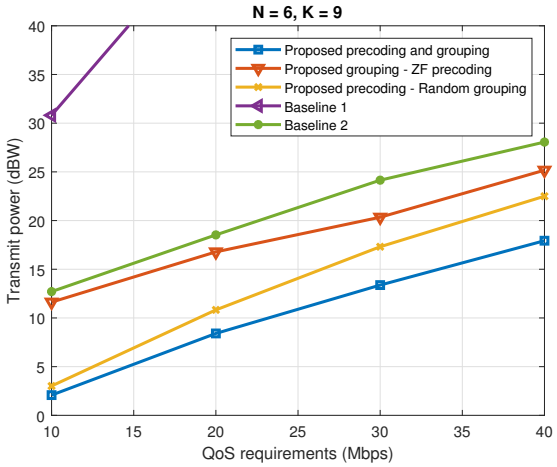
Fig. 7 shows the max-min rate performance of all schemes in the presence of channel estimation errors, measured in normalized mean square error (MSE). In this case, the proposed design treats the estimated CSI as the true one and uses it in the proposed algorithm. It is observed that the effectiveness of the proposed design is retained even in the presence of CSI imperfection. When the CSI is poorly estimated, the performances of all schemes significantly degrade. However, the baselines 1 and 3 are less sensitive to CSI errors than other schemes. This is because the other schemes rely on the estimated CSI for both user group selection and precoding vectors design, whereas the baseline 3 only designs the precoding vectors and the baseline 1 does not implement either precoding vector or user grouping.

D. Power Minimization Performance

Fig. 8 shows the minimum transmit power as a function of the QoS requirements. We note that as the performance of baseline 1 is far from other schemes, it is partially presented in the figure to better show the comparison. In general, it



(a) Highly-loaded scenario



(b) Over-loaded scenario

Fig. 8. Transmit power versus the QoS requirements. In (a), the system is highly-loaded with $K = N$. In (b), the system is over-loaded with $K > N$. Baseline 3 is unable to serve the users in the over-loaded scenario.

requires more power to satisfy the users with higher QoS requirements for all schemes. When the system is highly-loaded, i.e., $K = N$ in Fig. 8a, the proposed design significantly outperforms the baseline 1 and baseline 2 for all QoS values. More interestingly, the proposed design also outperforms the conventional precoding solution [21] in the baseline 3, especially in the high QoS regime, which is in line with the observation of Fig. 3. This is because although having adequate antennas to provide the multiplexing gains ($K = N$), the BS in baseline 3 has to use more energy when the channel is in weak conditions. On the other hand, the proposed design overcomes the weak condition by dividing the users into two groups with the number of users in each group smaller than N. As a result, the proposed design spends less energy to provide the same QoS to all users.

When the system is over-loaded, i.e., $K > N$, the baseline 3 is unable to serve the users' QoS because the BS cannot provide adequate degree of freedom to all users. Thus, the performance curve of the baseline 3 is absent in Fig. 8b. We can observe from Fig. 8b that the precoding design

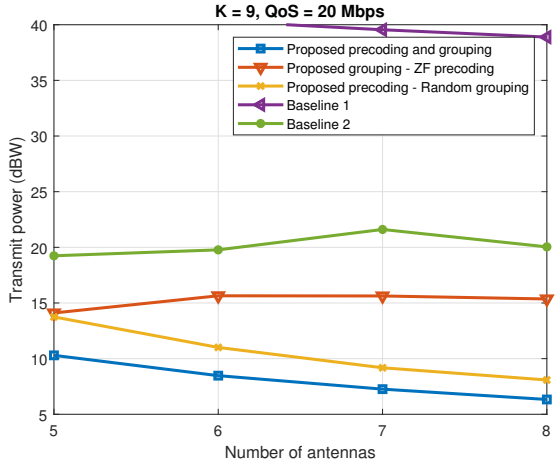


Fig. 9. Transmit power needs to serve $K = 9$ users for different number of antennas. Baseline 3 is unable to serve the users with the given QoS.

has more impacts than the user grouping policy, which is shown via a larger gap between the proposed design to the *ZF precoding* curve than to the *Random grouping* curve. In particular, combining the proposed precoding design with the user grouping saves 7 dB and 3 dB compared to the scheme using the ZF-based precoding and random user grouping, respectively. It is also shown that the proposed design transmits at a power which is much smaller than the power used by the baseline 1 and baseline 2.

In Fig. 9 presents the transmit power as a function of the antenna number. The BS serves $K = 9$ users, each of which requires a QoS of 20 Mbps. It is shown that having more antennas reduces the transmit power needed to serve the users with the same QoS targets. In particular, having 3 more antennas, the proposed joint design cuts 40% the transmit power, reducing from 10 dBW to 6 dBW when the number of antennas increases from 5 to 8. This is because having more antennas, the BS can exploit the spatial diversity more efficiently, resulting in a smaller transmit power. An interesting observation is that only the proposed precoding can take advantage of having more antennas, while the baseline and ZF-based schemes consume almost the same power.

Fig 10 plots the transmit power versus the number of users when the BS is equipped with 6 antennas and the users require a QoS of 20 Mbps. It is common that the BS spends more power to serve more users. The proposed joint design always consumes the least power. More specifically, the proposed framework saves more than 50% the transmit power compared with the baseline 2. Moreover, the proposed user grouping reduces 20% the transmit power compared with the random grouping scheme, which is shown via a constant gap between the two curves.

VIII. Conclusions and Discussions

A. Conclusions

In this paper, we have investigated the performance of a multiuser MISO system under highly-loaded conditions in which conventional multiplexing techniques perform poorly. We proposed a joint bandwidth allocation and precoding

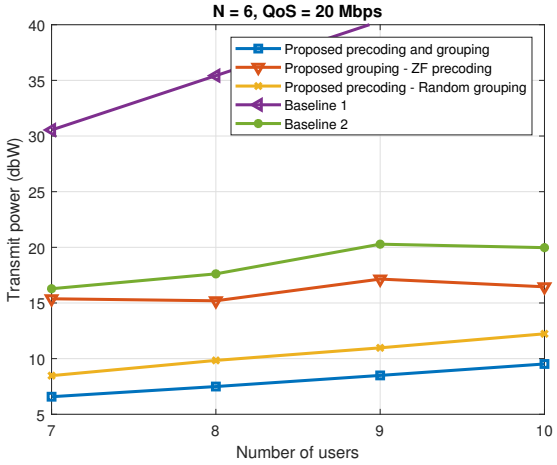


Fig. 10. Transmit power versus the number of users. Baseline 3 is unable to serve the users with the given QoS.

vectors design that fully exploits the spatial multiplexing gain and the frequency dimension. Two iterative algorithms based on successive convex approximation were proposed to solve the min rate maximization and power minimization problems, two important problems in multiuser systems. Furthermore, a low-complexity user grouping policy based on the singular value decomposition was proposed to further enhance the system performances. We showed via numerical results that the proposed design significantly outperforms existing schemes in terms of both the rate improvement and transmit power reduction.

B. Discussions

The advantages of the proposed joint design relies on an assumption that the frequency channel in the B5G networks can be arbitrarily assigned to the users. In the current LTE systems, the frequency channels are divided into sub-carriers whose bandwidth are usually predetermined. In this case, the output of the proposed design can be rounded up to the closest resources available in the system.

The outcomes of this work suggest several research directions. One promising topic is to consider location-dependent inter-cell interference (ICI) and cooperation among neighboring cells to better mitigate the ICI. In order to precisely model the geographical distribution of ICI and to efficiently enable such cooperation, however, it imposes extra signaling overhead among the BSs to exchange relevant information. In addition, it usually requires a distributed optimization method as the BS only obtains the CSI from the users connected to itself. The second topic is to study the problem in the presence of channel estimation errors. Such imperfect CSI scenario will affect both the user grouping selection and joint bandwidth and precoding vectors optimization. The third topic is to consider the users with multiple receive antennas. In this case, the formulated optimization may be different since the precoding design should be adapted to exploit multiple antennas at the users. Another promising problem is consider heterogeneous users with diverse requirements for quality of service and quality of experience, e.g., latency. In such cases, the users' service requirements should be exploited in addition to the effective channel gains to determine the best grouping partition.

APPENDIX A
CONVEXITY OF FUNCTION $F(\boldsymbol{x}, y) = \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{y}$ [22]

For any positive semidefinite metric $\boldsymbol{A}$, we will show that the Hessian metric of $F(\boldsymbol{x}, y)$ is positive semidefinite. Indeed, the Hessian matrix of $F(\boldsymbol{x}, y)$ is

$$\mathcal{H}_F = \begin{bmatrix} \frac{2\boldsymbol{A}}{y} & -\frac{(2\boldsymbol{A})\boldsymbol{x}}{y^2} \\ -\frac{\boldsymbol{x}^T(2\boldsymbol{A})}{y^2} & \frac{2\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{y^3} \end{bmatrix}.$$

For arbitrary vector $\boldsymbol{c} = [\boldsymbol{a}^T b]^T$, where $\boldsymbol{a} \in \mathbb{R}^{N\times 1}$, consider a function

$$\begin{aligned} \boldsymbol{c}^T \mathcal{H}_F \boldsymbol{c} =& \frac{\boldsymbol{a}^T(2\boldsymbol{A})\boldsymbol{a}}{y} - \frac{\boldsymbol{a}^T(2\boldsymbol{A})\boldsymbol{x}b}{y^2} - \frac{\boldsymbol{x}^T(2\boldsymbol{A})\boldsymbol{a}b}{y^2} + \frac{2\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}b^2}{y^3} \\ \overset{(*)}{=}& \frac{2\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{y} - 4\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{x}b}{y^2} + \frac{2\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}b^2}{y^3} \\ =& 2\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} - 2\boldsymbol{a}^T\boldsymbol{A}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{x}}^T\boldsymbol{A}\tilde{\boldsymbol{x}}}{y}, \end{aligned} \tag{27}$$

where $\tilde{\boldsymbol{x}} \triangleq \boldsymbol{x}b/y$ and $(*)$ results from the fact that $\boldsymbol{A}$ is symmetric and $\boldsymbol{a}^T\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{x}^T\boldsymbol{A}\tilde{\boldsymbol{a}}$. It is obvious that the RHS of (27) is always non-negative for $y > 0$ and positive semi-definite matrix $\boldsymbol{A}$, which concludes the positive semi-definite of the Hessian metric of $F(\boldsymbol{x}, y)$.

APPENDIX B
PROOF OF PROPOSITION 1

We recall that the approximate constraints of the approximate problem (11) used in Algorithm 1 satisfy properties of the SCA algorithm [25]. Denote $\boldsymbol{\Phi} = (\boldsymbol{w}^\star, \boldsymbol{b}^\star, \gamma^\star_{m,k}, \boldsymbol{x}^\star, \boldsymbol{y}^\star)$ as the set of optimal solutions of problem (11), and let $\boldsymbol{\Phi}^t$ be the optimal solution of the t-th iteration. Furthermore, let $\mathcal{F}_{\text{original}}(\boldsymbol{\Phi})$ and $\mathcal{F}(\boldsymbol{\Phi})$ denote the objective function of the original problem (3) and the approximated problem (11), respectively. Because the feasible region of the approximated problem is a subset of the feasible region of the original problem, we have $\mathcal{F}_{\text{original}}(\boldsymbol{\Phi}) \geq \mathcal{F}(\boldsymbol{\Phi}), \forall \boldsymbol{\Phi}$. Consider a sequence of the objective function $\mathcal{F}(\boldsymbol{\Phi}^t), t = 1, 2, \ldots, ...$ According to [25, Lemma 2.2], $\boldsymbol{\Phi}^{t+1}$ is a better solution of problem (11) than $\boldsymbol{\Phi}^t$, thus we have $\mathcal{F}(\boldsymbol{\Phi}^{t+1}) \leq \mathcal{F}(\boldsymbol{\Phi}^t)$. Furthermore, due to the constraints on the total bandwidth and total transmit power, the sequence of the objective function $\mathcal{F}(\boldsymbol{\Phi}^t), t = 1, 2, \ldots$ is bounded and will converge, which completes the proof.

REFERENCES

[1] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[2] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[4] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3226–3239, September 2012.

[5] T. L. Marzetta, "How much training is required for multiuser MIMO?" in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, Oct 2006, pp. 359–363.
[6] T. X. Vu, T. A. Vu, and T. Q. S. Quek, "Successive pilot contamination elimination in multiantenna multicell networks," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 617–620, 2014.
[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, April 2013.
[8] E. Bjornson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, Feb 2016.
[9] Cisco, "Cisco annual internet report (2018 - 2023)," Tech. Rep.
[10] A. Bandi, M. R. B. Shankar, S. Chatzinotas, and B. Ottersten, "Joint user grouping, scheduling, and precoding for multicast energy efficiency in multigroup multicast systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8195–8210, 2020.
[11] Q. Vu, L. Tran, M. Juntti, and E. Hong, "Energy-efficient bandwidth and power allocation for multi-homing networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1684–1699, 2015.
[12] M. Ismail, A. T. Gamage, W. Zhuang, X. Shen, E. Serpedin, and K. Qaraqe, "Uplink decentralized joint bandwidth and power allocation for energy-efficient operation in a heterogeneous wireless medium," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1483–1495, 2015.
[13] D. Tsilimantos, J. Gorce, K. Jaffrès-Runser, and H. Vincent Poor, "Spectral and energy efficiency trade-offs in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 54–66, 2016.
[14] L. Dong, "Spectral- and energy-efficient transmission with joint bandwidth assignment and transmit power allocation," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 703–707.
[15] X. Zhang and F. Yang, "Joint bandwidth and power allocation for energy efficiency optimization over heterogeneous LTE/WiFi multi-homing networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.
[16] I. Bisio, S. Delucchi, F. Lavagetto, and M. Marchese, "Comparison among resource allocation methods with packet loss and power metrics in geostationary satellite scenarios," in *2013 IEEE International Conference on Communications (ICC)*, 2013, pp. 4271–4275.
[17] H. Zhang and S. Liu, "A resource matching algorithm for dynamic multi-homing service environment," in *2018 2nd IEEE Advanced Information Management,Communicates,Electronic and Automation Control Conference (IMCEC)*, 2018, pp. 97–102.
[18] L. Xu and W. Zhuang, "Energy-efficient cross-layer resource allocation for heterogeneous wireless access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4819–4829, 2018.
[19] H. Zhang, H. Liu, J. Cheng, and V. C. M. Leung, "Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1705–1716, 2018.
[20] R. Liu, M. Sheng, and W. Wu, "Energy-efficient resource allocation for heterogeneous wireless network with multi-homed user equipments," *IEEE Access*, vol. 6, pp. 14 591–14 601, 2018.
[21] D. W. H. Cai, T. Q. S. Quek, and C. W. Tan, "A unified analysis of max-min weighted SINR for MIMO downlink system," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3850–3862, 2011.
[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
[23] T. X. Vu, S. Chatzinotas, V. D. Nguyen, D. T. Hoang, D. N. Nguyen, M. Di Renzo, and B. Ottersteng, "Machine learning-enabled joint antenna selection and precoding design: From offline complexity to online performance," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3710-3722, June 2021.
[24] L. Lei, T. X. Vu, L. You, S. Fowler, and D. Yuan, "Efficient minimum-energy scheduling with machine-learning based predictions for multiuser MISO systems," in *Proc. IEEE Int. Conf. Commun.*, May 2018, pp. 1–6.
[25] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, May 2010.

Thang X. Vu (S'11–M'15) received the B.S. and the M.Sc., both in Electronics and Telecommunications Engineering, from the VNU University of Engineering and Technology, Vietnam, in 2007 and 2009, respectively, and the Ph.D. in Electrical Engineering from the University Paris-Sud, France, in 2014. In 2010, he received the Allocation de Recherche fellowship to study Ph.D. in France. From July 2014 to January 2016, he was a postdoctoral researcher with the Singapore University of Technology and Design (SUTD), Singapore. Currently, he is a research scientist at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. His research interests are in the field of wireless communications, with particular interests of wireless edge caching, cloud radio access networks, machine learning for communications and cross-layer resources optimization. He was a recipient of the SigTelCom 2019 best paper award.

Symeon Chatzinotas (S'06—M'09—SM'13) is currently Full Professor / Chief Scientist I and CoHead of the SIGCOM Research Group at SnT, University of Luxembourg. In the past, he has been a Visiting Professor at the University of Parma, Italy and he was involved in numerous Research and Development projects for the National Center for Scientific Research Demokritos, the Center of Research and Technology Hellas and the Center of Communication Systems Research, University of Surrey. He received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He was a co-recipient of the 2014 IEEE Distinguished Contributions to Satellite Communications Award, the CROWNCOM 2015 Best Paper Award and the 2018 EURASIC JWCN Best Paper Award. He has (co-)authored more than 400 technical papers in refereed international journals, conferences and scientific books. He is currently in the editorial board of the IEEE Open Journal of Vehicular Technology and the International Journal of Satellite Communications and Networking.

Bjorn Ottersten (S'87–M'89–SM'99–F'04) received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linköping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, Belgium, and the University of Luxembourg, Luxembourg. From 1996 to 1997, he was the Director of Research with ArrayComm, Inc., a start-up in San Jose, CA, USA, based on his patented technology. In 1991, he was appointed Professor of signal processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr. Ottersten has been Head of the Department for Signals, Sensors, and Systems, KTH, and Dean of the School of Electrical Engineering, KTH. He is currently the Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg.

He is a recipient of the IEEE Signal Processing Society Technical Achievement Award, the EURASIP Group Technical Achievement Award, and the European Research Council advanced research grant twice. He has co-authored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, 2013, and 2019, and 8 IEEE conference papers best paper awards. He has been a board member of IEEE Signal Processing Society, the Swedish Research Council and currently serves of the boards of EURASIP and the Swedish Foundation for Strategic Research. Dr. Ottersten has served as Editor in Chief of EURASIP Signal Processing, and acted on the editorial boards of IEEE Transactions on Signal Processing, IEEE Signal Processing Magazine, IEEE Open Journal for Signal Processing, EURASIP Journal of Advances in Signal Processing and Foundations and Trends in Signal Processing. He is a fellow of EURASIP.