IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 29, 2019, accepted April 27, 2019, date of publication May 9, 2019, date of current version May 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2915573

Resource Allocation Algorithm for MU-MIMO Systems With Double-Objective Optimization Under the Existence of the Rank Deficient Channel Matrix

SU PAN[1,2], **YAN YAN**[1], **KUSI ANKRAH BONSU**[1], **AND WEIWEI ZHOU**[1,3]

[1]Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2]Jiangsu Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[3]Global Big Data Technologies Centre (GBDTC), University of Technology Sydney, Sydney, NSW 2007, Australia

Corresponding author: Su Pan (supan@njupt.edu.cn)

This work was supported in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX18_0883, in part by the Key Technologies R&D Program of Jiangsu Province under Grant BE2018733, and in part by the Open Research Fund of the Jiangsu Engineering Research Center of Communication and Network Technology, NJUPT.

ABSTRACT This paper proposes a double-objective optimization resource allocation algorithm for the multi-user multiple-input/multiple-output (MU-MIMO) system in the general wireless environment and demonstrates the maximum number of simultaneously supportable users and the achievable bit rates of users in the general wireless environment with full rank and rank-deficient channels. The double-objective joint optimization algorithm proposed in this paper simultaneously optimizes energy efficiency and system throughput by user selection and power allocation. On this basis, the proposed algorithm guarantees the different QoS requirements of various services, including rate requirements and delay requirements.

INDEX TERMS MU-MIMO, rank deficient, double-objective optimization, QoS guarantee, resource allocation.

I. INTRODUCTION

MU-MIMO technology can effectively utilize spatial resources to improve the throughput of wireless communication systems without consuming additional spectral bandwidths [1]. Therefore, MU-MIMO has become one of the key technologies of 5th-Generation (5G) networks [36], [37]. With the increasing demand for wireless communications and the increasing requirement for environmental protection, the optimizations of system throughput and energy efficiency (EE, represented by transmission data rate per unit energy) are two important goals of MU-MIMO resource allocation algorithm. The algorithm optimizes the two goals through user selection and power allocation [2]–[4]. In MU-MIMO systems, user selection and power allocation are related to the rank of user's MIMO channel matrix, because the rank of the user's channel determines the number of space-division channels, thus affecting the user's achievable data rate given a certain power level [5]–[7]. The existing literatures only consider the full-rank channel when designing a resource allocation algorithm, that is, they assume that the channel matrices and the aggregate channel matrix of users are full ranks. However, in a generally practical environment, the full rank channels and the rank deficient channels exist simultaneously in many cases due to the scattering environment and physical antenna spacing [8]–[10]. For example, when line-of-sight (LOS) paths exist between the base station and the user terminal or the distance between the antennas is less than an integral multiple of the wavelength, the channel matrix of the user is rank deficient. The simultaneous existence of full rank and rank deficient channels is more likely to occur in the user-intensive area because small cells are easier to cause LOS paths [11], [12]. When the rank deficient channel exists in MU-MIMO systems, the user's achievable data rate and the system throughput vary with the rank of user's channel matrix, thus the total system power varies accordingly, and the maximum number of simultaneously supportable users will vary with the combinations of selected users.

The associate editor coordinating the review of this manuscript and approving it for publication was Syed Mohammad Zafaruddin.

Therefore, those resource allocation algorithms in existing literatures are not suitable for optimizing the system throughput and energy efficiency in MU-MIMO systems in the practical wireless environment with full rank and rank deficient channels. All the above motivate us to study the resource allocation algorithm with the existence of the rank deficient channel matrix.

Another problem in the existing literatures is that they only minimize total system power based on the given minimum system throughput or maximize the system throughput based on the given maximum system power [13]–[15]. As far as we know, there is no research focusing on the double-objective (throughput and EE) optimization problem. The difficulty lies in the fact that the system throughput and the energy efficiency are coupled to each other; thereby the common-used greedy algorithm cannot optimize the two objectives at the same time. In addition, the exhaustive method cannot be practically implemented because of excessive algorithm complexity. The above observation prompts us to research the double-objective optimization resource allocation algorithm that maximizes both the system throughput and the energy efficiency.

The QoS guarantee, including delay requirements and rate requirements, must be taken into account when designing the resource allocation algorithm. Since different kinds of services have different QoS requirements and too many requirements may lead to excessive constraints or non-convex optimization problem, the existing literatures do not provide accurate QoS guarantee, that is, they only consider the lower bound of the rate requirements without considering the upper bound of the rate requirements in the QoS profile. In fact, most real-time services do not require excessive data rate, for example, 64Kbps can make the quality of the voice service reach its upper limit, i.e., make MOS (Mean Opinion Score) reach 4.4 [16]–[18]. Therefore, for real-time services, the data rate exceeding the upper limit of the data rate requirements is meaningless and invalid.

To solve the problems mentioned above, the paper proposes a resource allocation algorithm which simultaneously optimizes the effective system throughput (effective system throughput only counts users' data rates between their required upper and lower bounds) and the energy efficiency while guaranteeing the QoS requirements in the practical wireless environment with full rank and rank deficient channels. We assume that the MU-MIMO system discussed in the paper has prefect channel state information. Firstly, we derive the maximum number of simultaneously supportable users and the achievable data rates of selected users with full rank and rank deficient channel matrices in MU-MIMO systems. Then, we establish a double-objective optimization model. In this model, the optimization objectives are maximizing the energy efficiency and the effective system throughput; the constraints are the upper bound of antenna power, the lower and upper bounds of user's data rate. Since the proposed optimization problem has two optimization objectives, it cannot be directly solved by convex optimization method. We utilize the Lagrange dual method to solve the double-objective optimization problem. We derive the Lagrange dual convex optimization problem of the original optimization problem and prove the strong duality between the original problem and the dual problem, thus the original problem can be solved by solving its Lagrange dual problem. A convex optimization problem with the same optimal solution as the Lagrange dual problem can be obtained by simplifying the Lagrange dual problem. We solve the simplified convex optimization problem in two steps. Firstly, we obtain the optimal parallel channel power of users by using convex optimization method directly. This power depends on the eigenvalues of user's equivalent channel matrix. Secondly, we rewrite the simplified convex optimization problem as a function of the eigenvalues of selected user's equivalent channel matrix by substituting the optimal power of users which depends on the eigenvalues into the simplified convex optimization problem. Therefore, we can solve the problem by user selection to find the optimal eigenvalues of users' equivalent channel matrices. In order to guarantee the delay requirement, the user whose waiting slots reach the maximum waiting slots is first selected, and then the user selection criterion is to optimize the value of the above function.

The contributions of this paper are summarized as follows,

1. We derive the maximum number of simultaneously supportable users, the achievable data rates of selected users and the total system power in the practical MIMO environment which contains full rank and rank deficient channels. We prove that the maximum number of simultaneously supportable users in the rank deficient environment is larger than that in the full rank environment, which means more users can be selected in the rank deficient environment. However, the achievable data rates and the power of selected users decrease in the rank deficient environment.

2. We propose a double-objective optimization resource allocation algorithm which simultaneously optimizes the energy efficiency and the system throughput in the practical MIMO systems.

3. We provide accurate QoS guarantee, including rate requirements and delay requirements. We consider both the upper and lower bounds of the data rate requirements and limit the real-time user's rate to an effective range, then more power and data rate are allocated to non-real-time users correspondingly, thereby optimizing the effective system throughput.

The rest of the paper is organized as follows. Section II introduces related works. The maximum number of simultaneously supportable users and achievable data rate in the full rank and rank deficient environments are described in Section III. In Section IV, a double-objective optimization model is established, and the solution is presented. Simulation results are provided in Section V, and Section VI concludes the paper.

II. RELATED WORKS

[19]–[23] introduced the user selection algorithms that only optimize the system throughput on the given maximum system power. Reference [19] proposed a user selection algorithm to optimize the system throughput while guaranteeing the sum of all selected users' data rates and the sum of all selected users' delay time are not less than the thresholds. The algorithm in [19] considered neither QoS differentiation among various services nor the upper bound of real-time services' data rate, so the algorithm could not meet the requirement of a specific user and could not achieve reasonable and effective resource allocation. Reference [20] introduced a Proportional Fair (PF) user selection algorithm that computes the priority for each user and selects the user with the highest priority in each scheduling time slot. Moreover, the priority for a user was proportional to the user's rate in the current scheduling time slot and inversely proportional to the user's average rate of the earlier scheduling time slots. However, this algorithm did not consider the QoS requirements of different services, thus it could not be well applied to communication systems with multi-services. The capacity based and the Frobenius norm based user selection algorithms were proposed in [22], and the chordal distance-based user selection algorithm was proposed in [23]. However, these algorithms proposed in [22] and [23] did not consider QoS differentiation among various services so that they could not optimize the effective throughput of the system.

References [24]–[28] focused on the energy efficiency optimization in wireless communication systems. References [24] and [25] proposed resource allocation algorithms to optimize the energy efficiency based on frequency selective channel and channel state information respectively, but neither of them considered the QoS requirements of services. The energy efficiency optimization algorithm proposed in [27] only considered the lower bound of rate requirement of QoS, neither considered the delay requirement of QoS nor QoS differentiation among various services. The energy efficiency optimization algorithm proposed in [28] considered the upper and lower bounds of rate requirement, but it did not consider the delay requirement of QoS and the power constraint of the antenna in the base station (BS), which makes the power allocation not significant in practice.

In [38], the authors used the Dinkelbach method to solve the EE maximization problem that was a typical fractional programming problem, and the Dinkelbach method performed well in this scenario. However, the double-objective optimization problem established in our paper is not a fractional programming problem, hence the Dinkelbach method is not suitable in our paper. Reference [39] adopted a different framework from our paper. Reference [39] formulated a double-objective problem by using the weighed Tchebycheff method, which was a Multi-objective Evolutionary Algorithm. The Multi-objective Evolutionary Algorithm has many advantages, but it still has many shortcomings, such as high computational complexity, lack of complete convergence proof etc. The algorithm used in our paper can be essentially classified into the traditional Multi-objective optimization algorithm (main objective method) which has many unique advantages, such as small computational complexity, easy implementation etc. We evaluate the computational complexity of our algorithm in the appendix and find that computational complexity is at the same level as which of single objective problem.

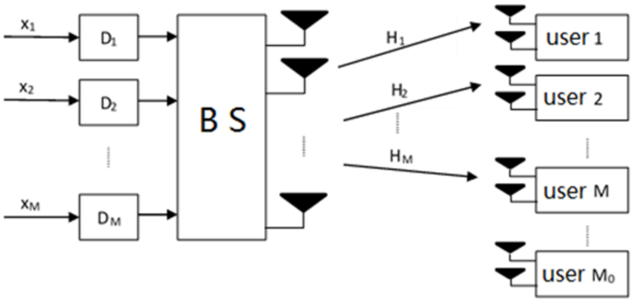


FIGURE 1. System model.

III. SYSTEM MODEL

As shown in Figure 1, we consider the downlink of an MU-MIMO system with a normalized bandwidth carrier, where the BS is equipped with N_T transmit antennas and has K users totally in the cell. The user m $(1 \leq m \leq K)$ has n_m antennas and $N_T \geq n_m$ in general. M denotes the maximum number of simultaneously supportable users in the cell, which varies with the rank of aggregate channel matrix of selected users.

A. *M* IN THE RANK DEFICIENT ENVIRONMENT

In the MU-MIMO system, multiple users can use the same frequency simultaneously to receive the data from the BS, and it is necessary to eliminate inter-user interference by precoding. The commonly used precoding methods include dirty paper coding (DPC) and block diagonalization (BD) [29]. We adopt BD in this paper for its lower complexity.

Denoting the transmitting signal vector of the user m $(1 \leq m \leq K)$ as $\boldsymbol{x}_m \in \mathbb{C}^{n_m \times 1}$, the receiving signal $\boldsymbol{y}_m \in \mathbb{C}^{n_m \times 1}$ can be expressed as:

$$\begin{aligned} \boldsymbol{y}_m &= \boldsymbol{H}_m \sum_{j=1}^{M} \boldsymbol{D}_j \boldsymbol{x}_j + \boldsymbol{k}_m \\ &= \boldsymbol{H}_m \boldsymbol{D}_m \boldsymbol{x}_m + \boldsymbol{H}_m \sum_{j=1, j \neq m}^{M} \boldsymbol{D}_j \boldsymbol{x}_j + \boldsymbol{k}_m \end{aligned} \tag{1}$$

where $\boldsymbol{H}_m \in \mathbb{C}^{n_m \times N_T}$ denotes the complex channel matrix of the user m, $\boldsymbol{D}_m \in \mathbb{C}^{N_T \times n_m}$ denotes the precoding matrix of the user m, $\boldsymbol{k}_m \in \mathbb{C}^{n_m \times 1}$ denotes the noise vector with zero mean and covariance σ^2.

In (1), $\boldsymbol{H}_m \sum_{j=1, j \neq m}^{M} \boldsymbol{D}_j \boldsymbol{x}_j$ represents the interference from other users to user m. To eliminate the interference, the precoding matrix for the user m needs to meet the following equation:

$$\boldsymbol{H}_m\boldsymbol{D}_j = 0, \quad m = 1, 2, \mathrm{L}, \cdots, M; m \neq j \tag{2}$$

Let $\hat{\boldsymbol{H}}_m = [\boldsymbol{H}_1^T, \boldsymbol{H}_2^T, \mathrm{L}, \cdots, \boldsymbol{H}_{m-1}^T, \boldsymbol{H}_{m+1}^T, \mathrm{L}, \cdots, \boldsymbol{H}_M^T]^T$, $\hat{\boldsymbol{H}}_m \in \mathrm{C}^{\sum_{j=1,j\neq m}^{M} n_j \times N_T}$ is the joint matrix of interference users, $\hat{L}_m$ denotes the rank of $\hat{H}_m$. Applying the singular value decomposition (SVD) to $\hat{H}_m$, we have:

$$\hat{\boldsymbol{H}}_m = \begin{bmatrix} \boldsymbol{U}_m^{(1)} & \boldsymbol{U}_m^{(0)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_m^{(1)} & \boldsymbol{V}_m^{(0)} \end{bmatrix}^H \tag{3}$$

where Σ_m denotes a $\hat{L}_m \times \hat{L}_m$ diagonal matrix and $V_m^{(0)}$ denotes a $N_T \times \left(N_T - \hat{L}_m\right)$ unitary matrix. According to the property of SVD, we have:

$$\hat{\boldsymbol{H}}_m \boldsymbol{V}_m^{(0)} = \boldsymbol{0} \tag{4}$$

According to (4), the precoding matrix of the user m can be expressed as:

$$\boldsymbol{D}_m = \boldsymbol{V}_m^{(0)} \boldsymbol{B}_m \tag{5}$$

where $V_m^{(0)}$ is used to eliminate inter-user interference and $\boldsymbol{B}_m$ is used to maximize the data rate [4]. To guarantee $V_m^{(0)}$ in (4) has the nonzero solutions, the number of the equations in (4) must be less than the number of variables, thus:

$$\hat{L}_m \leq N_T, \quad \forall m = 1, 2, \cdots, M \tag{6}$$

where $\hat{L}_m$ denotes the rank of $\hat{\boldsymbol{H}}_m$, which increases with the number of selected users. It is obvious that (6) defines the upper bound of selected users, i.e.,M.

In the full rank environment, $\hat{L}_m = \sum_{j=1,j\neq m}^{M} n_j$, while in the rank deficient environment, $\hat{L}_m < \sum_{j=1,j\neq m}^{M} n_j$. Therefore, the maximum number of simultaneously supportable users M in the rank deficient environment is larger than that in the full rank environment, that is, the system can serve more users simultaneously in the rank deficient environment.

B. THE RELATIONSHIP BETWEEN THE TRANSMISSION POWER AND THE USER'S DATA RATE IN THE RANK DEFICIENT ENVIRONMENT

Let $\bar{\boldsymbol{H}}_m = \boldsymbol{H}_m \boldsymbol{V}_m^{(0)}, \bar{\boldsymbol{H}}_m \in \mathrm{C}^{n_m \times \left(N_T - \hat{L}_m\right)}$, the rank of $\bar{\boldsymbol{H}}_m$ is $\bar{L}_m$ and the rank of $\boldsymbol{H}_m$ is L_m. Applying the SVD to $\bar{\boldsymbol{H}}_m$, we have:

$$\bar{\boldsymbol{H}}_m = \boldsymbol{H}_m \boldsymbol{V}_m^{(0)} = \begin{bmatrix} \boldsymbol{U}_m^1 & \boldsymbol{U}_m^0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_m^1 & \boldsymbol{V}_m^0 \end{bmatrix}^H \tag{7}$$

where $\boldsymbol{\Lambda}_m$ denotes a $\bar{L}_m \times \bar{L}_m$ diagonal matrix and $\boldsymbol{V}_m^1$ denotes a $\left(N_T - \hat{L}_m\right) \times \bar{L}_m$ unitary matrix. Since $\boldsymbol{V}_m^{(0)}$ is a unitary matrix with the rank $N_T - \hat{L}_m$, the rank of $\bar{\boldsymbol{H}}_m$ can be expressed as:

$$\bar{L}_m = \min\left(L_m, N_T - \hat{L}_m\right) \tag{8}$$

Let $\boldsymbol{B}_m = \boldsymbol{V}_m^1$ to maximize the data rate, thereby, $\boldsymbol{D}_m = \boldsymbol{V}_m^{(0)} \boldsymbol{V}_m^1$. Substituting $\boldsymbol{D}_m$ into (1), we have:

$$\begin{aligned} \boldsymbol{y}_m &= \boldsymbol{H}_m \boldsymbol{D}_m \boldsymbol{x}_m + \boldsymbol{k}_m = \boldsymbol{H}_m \boldsymbol{V}_m^{(0)} \boldsymbol{V}_m^1 \boldsymbol{x}_m + \boldsymbol{k}_m \\ &= \boldsymbol{U}_m^1 \boldsymbol{\Lambda}_m \boldsymbol{x}_m + \boldsymbol{k}_m \end{aligned} \tag{9}$$

TABLE 1. Types of services and the corresponding requirements.

Service Type z	Rate Request r_z (Kbps)	Delay Requirement d_z (ms)
z=1 Voice	4~64	100
z=2 Streaming	50~85	150
z=3 Interaction	3~385	250
z=4 Background	15~10^5	None

Multiplying $\boldsymbol{y}_m$ by $\boldsymbol{U}_m^{1^H}$ which is a unitary matrix, we obtain:

$$\boldsymbol{y}'_m = \boldsymbol{U}_m^{1^H} \boldsymbol{y}_m = \boldsymbol{\Lambda}_m \boldsymbol{x}_m + \boldsymbol{U}_m^{1^H} \boldsymbol{k}_m \tag{10}$$

(10) indicates that the MIMO channels of each user can be divided into several parallel equivalent channels and the number of these parallel channels is $\bar{L}_m$ which is the rank of $\boldsymbol{\Lambda}_m$ [30], [31]. Therefore, the relationship between the achievable data rate R_m of the user m and the transmission power in the parallel channel k of the user m can be expressed as:

$$\begin{aligned} R_m &= \sum_{k=1}^{\bar{L}_m} \log_2\left(1 + \frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right) \\ &= \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} \log_2\left(1 + \frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right) \end{aligned} \tag{11}$$

where $\lambda_{m,k}$ is the k-th diagonal element of $\boldsymbol{\Lambda}_m$.

IV. DOUBLE-OBJECTIVE OPTIMIZATION MODEL

A. DOUBLE-OBJECTIVE OPTIMIZATION PROBLEM

In this paper, we consider the QoS requirements of four types of services divided by 3GPP as shown in TABLE 1 [32].

According to Table 1, for a real-time service user such as voice user, if the user's achievable data rate exceeds 64kbps, the excess is meaningless. However, the existing literatures only consider the lower bound of data rate requirements and ignore the upper bound of data rate requirements in the QoS profiles. In this paper, we consider both the lower and upper bounds of QoS rate requirements. We assume that each user uses one service.

R_{m1} and R_{m2} denote the upper bound and the lower bound of data rate of the user m respectively, we have:

$$R_{m0} < R_m = \sum_{k=1}^{\bar{L}_m} \log_2\left(1 + \frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right) < R_{m1} \tag{12}$$

M denotes the maximum number of simultaneously supportable users in the MU-MIMO system, so the effective system throughput is the sum of all selected users' rates. Denote the effective system throughput as C, which follows:

$$C = \sum_{m=1}^{M} R_m = \sum_{m=1}^{M} \sum_{k=1}^{\bar{L}_m} \log_2\left(1 + \frac{p_{m,k} \cdot \lambda_{m,k}^2}{\sigma^2}\right) \tag{13}$$

Denote the minimum system throughput as C_0, and we have:

$$C \geq C_0 \tag{14}$$

P_i^{TX} denotes the transmission power of the antenna i, $i = 1, 2, \cdots, N_T$. According to [33], P_i^{TX}can be expressed as:

$$P_i^{TX} = \sum_{m=1}^{M} \sum_{k=1}^{\bar{L}_m} \left| D_m(i,k) \right|^2 \cdot p_{m,k} < P_0 \tag{15}$$

where P_0 denotes the upper bound of antenna power. Then, the total system power E can be expressed as:

$$E = e \cdot \sum_{i=1}^{N_T} P_i^{TX} + P_c \tag{16}$$

where e denotes the reciprocal of drain efficiency of the power amplifier, and P_c denotes the circuit power dissipation [34].

According to (13) (15) (16), the energy efficiency EE can be expressed as:

$$EE = \frac{C}{E} = \frac{\sum\limits_{m=1}^{M} \sum\limits_{k=1}^{\bar{L}_m} \log_2 \left(1 + \frac{p_{m,k} \cdot \lambda_{m,k}^2}{\sigma^2} \right)}{e \cdot \sum\limits_{i=1}^{N_T} \sum\limits_{m=1}^{M} \sum_{k=1}^{\bar{L}_m} \left| D_m(i,k) \right|^2 \cdot p_{m,k} + P_c} \tag{17}$$

According to (12) (13) (14) (15) (17), the double-objective optimization problem that simultaneously optimizes the energy efficiency EE and the effective system throughput C through power allocation $(p_{m,k})$ and user selection $(\lambda_{m,k})$ is established as follows:

$$\begin{aligned} & max\ EE = \frac{C\left(p_{m,k}, \lambda_{m,k}\right)}{E(p_{m,k})} && (18) \\ & max\ C\left(p_{m,k}, \lambda_{m,k}\right) \\ & s.t.\ C\left(p_{m,k}, \lambda_{m,k}\right) \geq C_0 \\ & \qquad P_i^{TX} < P_0, \quad i = 1, 2, \ldots, N_T \\ & \qquad R_{m0} \leq R_m \leq R_{m1}, \quad m = 1, 2, \ldots, M \\ & \qquad p_{m,k} \geq 0, \quad \forall i, m && (19) \end{aligned}$$

Note that (18) and (19) indicate that there are two maximizing goals in the above optimization problem. The double-objective optimization problem can be rewritten as:

$$\begin{aligned} & min\ E(p_{m,k}) && (20) \\ & max\ C\left(p_{m,k}, \lambda_{m,k}\right) \\ & s.t.\ C\left(p_{m,k}, \lambda_{m,k}\right) \geq C_0 \\ & \qquad P_i^{TX} < P_0, \quad i = 1, 2, \ldots, N_T \\ & \qquad R_{m0} \leq R_m \leq R_{m1}, \quad m = 1, 2, \ldots, M \\ & \qquad p_{m,k} \geq 0, \quad \forall i, m && (21) \end{aligned}$$

It is noted that the data delay requirement is not included in the above equations. This requirement is considered in user selection stage which will be explained in the next section.

B. SOLUTION OF THE DOUBLE-OBJECTIVE OPTIMIZATION PROBLEM

Since it takes more energy to increase the throughput, (20) and (21) in the above problem are incompatible. According to [35], a common method for solving multi-objective optimization problems is the main objective method, which retains (20) as the main objective and converts (21) into a constraint. Therefore, the double-objective optimization problem can be rewritten as:

$$\begin{aligned} & min\ E(p_{m,k}) \\ & s.t.\ max\ C\left(p_{m,k}, \lambda_{m,k}\right) \geq C_0 \\ & \qquad P_i^{TX} < P_0, \quad i = 1, 2, \ldots, N_T \\ & \qquad R_{m0} \leq R_m \leq R_{m1}, \quad m = 1, 2, \ldots, M \\ & \qquad p_{m,k} \geq 0, \quad \forall i, m \end{aligned} \tag{22}$$

It is noted that the above problem is different from the single-objective optimization problem since the constraint $C\left(p_{m,k}, \lambda_{m,k}\right) \geq C_0$ in single-objective optimization problem does not maximize $C\left(p_{m,k}, \lambda_{m,k}\right)$. The main objective method in the paper, however, is to minimize $E(p_{m,k})$ while maximizing $C\left(p_{m,k}, \lambda_{m,k}\right)$ as much as possible. In the above double-objective optimization problem, the feasible set is a convex set, the constraints include concave functions, (21) is a concave function, (22) is an affine function, so the proposed double-objective optimization problem is not a standard convex optimization problem and cannot be directly solved by convex optimization method. We use Lagrange dual method to solve (22). First, we define the Lagrangian associated with the (22) as:

$$\begin{aligned} & \mathrm{l}(\{p_{m,k}\}, \upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ & \quad = \big(E(p_{m,k}) - \upsilon(max\ C\left(p_{m,k}, \lambda_{m,k}\right) - C_0) \\ & \qquad - \sum_{i=1}^{N_T} \gamma_i \left(P_0 - P_i^{TX}\right) - \sum_{m=1}^{M} \alpha_m \left(R_{m1} - R_m\right) \\ & \qquad - \sum_{m=1}^{M} \beta_m \left(R_m - R_{m0}\right) \Bigg) \\ & \quad = \Bigg(\Bigg(e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} \left| D_m(i,k) \right|^2 \cdot p_{m,k} + P_c \Bigg) \\ & \qquad - \upsilon \Bigg(max \Bigg(\sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} \\ & \qquad\qquad \times \log_2 \left(1 + \frac{p_{m,k} \cdot \lambda_{m,k}^2}{\sigma^2} \right) \Bigg) - C_0 \Bigg) \\ & \qquad - \sum_{i=1}^{N_T} \gamma_i \left(P_0 - P_i^{TX}\right) - \sum_{m=1}^{M} \alpha_m \left(R_{m1} - R_m\right) \\ & \qquad - \sum_{m=1}^{M} \beta_m \left(R_m - R_{m0}\right) \Bigg) \end{aligned} \tag{23}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{N_T})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$, and υ denote the Lagrange multipliers respectively, $\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta} > 0, \forall m, i, k$. We define the Lagrange dual function as:

$$h(\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = min\ \mathrm{l}(\{p_{m,k}\}, \upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \tag{24}$$

According to (23) (24), the Lagrange dual problem of the original problem is:

$$\begin{aligned} &\max h(\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ &subject\ to\ p_{m,k}, \upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta} > 0, \quad \forall m, i, k \end{aligned} \tag{25}$$

According to [35], whether the original problem is a convex optimization problem or not, its dual problem is a convex optimization problem. In the following, we prove that the original problem (22) and the dual problem (25) have the same optimal value. Thus we can solve (22) by solving (25).

Definition 1: E^* denotes the optimal value of the original problem, if q that satisfies $q \leq E^*$ exists, then q is a lower bound of E^*.

Definition 2: Q denotes a set of real numbers, if any $q \in Q$ satisfies $q \leq E^*$, then all elements in Q are the lower bounds of E^* and the maximum element $q_{\max} \in Q$ is the infimum of E^*.

Proposition: For any feasible $\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge$, let $h^\wedge = h\left(\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge\right)$, then the dual function is the lower bound of the optimal value of its original problem, and the optimal value of dual problem is the infimum of the original problem, i.e., For any feasible $\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge$, we have:

$$h^\wedge \leq E^* \tag{26}$$

Proof:

Supposing $\tilde{p}_{m,k}$ is a feasible point of the original problem (22), then:

$$\sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} \log_2\left(1 + \frac{\tilde{p}_{m,k} \cdot \lambda_{m,k}^2}{\sigma^2}\right) - C_0 \geq 0 \tag{27}$$

$$P_0 - \sum_{m=1}^{M} \sum_{k=1}^{\bar{L}_m} |D_m(i,k)|^2 \cdot \tilde{p}_{m,k} \geq 0 \tag{28}$$

$$\sum_{k=1}^{\bar{L}_m} \log_2\left(1 + \frac{\tilde{p}_{m,k} \lambda_{m,k}^2}{\sigma^2}\right) - R_{m0} \geq 0 \tag{29}$$

$$R_{m1} - \sum_{k=1}^{\bar{L}_m} \log_2\left(1 + \frac{\tilde{p}_{m,k} \lambda_{m,k}^2}{\sigma^2}\right) \geq 0 \tag{30}$$

For any feasible $\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge$, and $\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge > 0, \forall m, i, k$, we have:

$$\mathrm{l}(\{\tilde{p}_{m,k}\}, \upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge)$$

$$= \Bigg(\Bigg(e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} |D_m(i,k)|^2 \cdot \tilde{p}_{m,k} + P_c\Bigg)$$

$$- \upsilon^\wedge \Bigg(max \Bigg(\sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)}$$

$$\times \log_2\left(1 + \frac{\tilde{p}_{m,k} \cdot \lambda_{m,k}^2}{\sigma^2}\right)\Bigg) - C_0\Bigg)$$

$$- \sum_{i=1}^{N_T} \gamma_i^\wedge \left(P_0 - P_i^{TX}\right) - \sum_{m=1}^{M} \alpha_m^\wedge (R_{m1} - R_m)$$

$$- \sum_{m=1}^{M} \beta_m^\wedge (R_m - R_{m0})\Bigg)$$

$$\leq e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} |D_m(i,k)|^2 \cdot \tilde{p}_{m,k} + P_c \tag{31}$$

That is:

$$h(\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge) = min\ \ell(\{p_{m,k}\}, \upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge)$$

$$\leq e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)} |D_m(i,k)|^2 \cdot \tilde{p}_{m,k} + P_c \tag{32}$$

For any feasible $\%p_{m,k}, \upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge, h(\upsilon^\wedge, \boldsymbol{\alpha}^\wedge, \boldsymbol{\gamma}^\wedge, \boldsymbol{\beta}^\wedge) \leq E(\%p_{m,k}) \leq E^*$ can be satisfied. Therefore, (26) is proved.

End of proof.

$\upsilon^*, \boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*$ denote the optimal solution of the dual problem, and h^* denotes the optimal value of dual problem, then according to (26), we have:

$$h^* \leq E^* \tag{33}$$

Obviously, the original problem (22) in the paper is an abstract convex optimization problem and satisfies the Slater's condition [35]. Therefore, according to Slater's condition, the strong duality holds and the duality gap $h^* - E^*$ is 0, then:

$$\begin{aligned} h^* &= E^* \\ h^* &= E^* \end{aligned} \tag{34}$$

(34) indicates that the original problem (22) and the dual problem (25) have the same optima, and then the optimal solution of (24) is the optimal solution of the original problem.

The Lagrange dual function (24) can be expanded and written as:

$$h(\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta})$$

$$= \min_{p_{m,k} \geq 0, \forall m,k} \Bigg(\Bigg(e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T - \hat{L}_m\right)}$$

$$\times |D_m(i,k)|^2 \cdot p_{m,k} + P_c\Bigg)$$

$$
\begin{aligned}
&-\upsilon\Bigg(\max_{p_{m,k},\lambda_{m,k}}\Bigg(\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\\
&\times\log_2\left(1+\frac{p_{m,k}\cdot\lambda_{m,k}^2}{\sigma^2}\right)\Bigg)-C_0\Bigg)\\
&-\sum_{i=1}^{N_T}\gamma_i\left(P_0-\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\left|D_m\left(i,k\right)\right|^2\cdot p_{m,k}\right)\\
&-\sum_{m=1}^{M}\alpha_m\Bigg(R_{m1}-\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\\
&\times\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)\Bigg)\\
&-\sum_{m=1}^{M}\beta_m\Bigg(\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\\
&\times\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)-R_{m0}\Bigg)\Bigg)\\
&=h'(\upsilon,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta})+P_c+\upsilon C_0-\sum_{i=1}^{N_T}\gamma_iP_0\\
&-\sum_{m=1}^{M}\alpha_mR_{m1}+\sum_{m=1}^{M}\beta_mR_{m0}
\end{aligned}
\tag{35}
$$

where

$$
\begin{aligned}
&h'(\upsilon,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta})\\
&=\min_{\lambda_{m,k}}\min_{p_{m,k}}\Bigg(\Bigg(\Bigg(e\cdot\sum_{i=1}^{N_T}\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\\
&\times\left|D_m\left(i,k\right)\right|^2\cdot p_{m,k}\Bigg)\\
&-\upsilon\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\log_2\left(1+\frac{p_{m,k}\cdot\lambda_{m,k}^2}{\sigma^2}\right)\\
&+\sum_{i=1}^{N_T}\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\gamma_i\left|D_m\left(i,k\right)\right|^2\cdot p_{m,k}\\
&+\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\alpha_m\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)\\
&-\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\beta_m\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)\Bigg)\Bigg)
\end{aligned}
\tag{36}
$$

(36) is the simplified function of Lagrange dual function (35) by removing the constants from (35). The optimal solution of the Lagrange dual problem (35) can be obtained by solving its simplified function (36). In the paper, we solve (36) by two steps: power allocation and user selection.

1) POWER ALLOCATION

In the power allocation step, we derive the optimal $p_{m,k}$ that depends on $\lambda_{m,k}$, where $\lambda_{m,k}$ is the k-th $\left(k=1,\mathrm{L},\min\left(L_m,N_T-\hat{L}_m\right)\right)$ diagonal element of $\bar{\boldsymbol{H}}_m$ and $\bar{\boldsymbol{H}}_m$ is the equivalent channel matrix of the user m. It means $\lambda_{m,k}$ is depended on which user is selected, so the value of $\lambda_{m,k}$ can be determined in next user selection step. In this step, we assume that $\lambda_{m,k}$ is a constant, so (36) can be rewritten as the following standard convex optimization problem (37):

$$
\begin{aligned}
&\min g(p_{m,k})\\
&=\left(e\cdot\sum_{i=1}^{N_T}\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\left|D_m\left(i,k\right)\right|^2\cdot p_{m,k}\right)\\
&-\upsilon\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\log_2\left(1+\frac{p_{m,k}\cdot\lambda_{m,k}^2}{\sigma^2}\right)\\
&+\sum_{i=1}^{N_T}\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\gamma_i\left|D_m\left(i,k\right)\right|^2\cdot p_{m,k}\\
&+\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\alpha_m\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)\\
&-\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\beta_m\log_2\left(1+\frac{p_{m,k}\lambda_{m,k}^2}{\sigma^2}\right)\\
&\textit{subject to } p_{m,k}\geq 0,\quad\forall m,k
\end{aligned}
\tag{37}
$$

The standard convex optimization problem (37) can be directly solved by the convex optimization method. The Lagrangian of (37) is:

$$
g(p_{m,k})+\sum_{m=1}^{M}\sum_{k=1}^{\min\left(L_m,N_T-\hat{L}_m\right)}\theta_{m,k}p_{m,k}
\tag{38}
$$

where $\boldsymbol{\theta}\left(\theta_{m,k}\right)$ denotes the Lagrange multiplier and $\theta_{m,k}>0,\forall m,k$. According to [35], the point satisfying the Karush-Kuhn-Tucker (KKT) conditions is the optimal solution when the optimization problem is a standard convex optimization problem. Then we have:

$$\nabla g(p_{m,k}) + \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \theta_{m,k} \nabla p_{m,k} = 0 \quad (39)$$

$$\theta_{m,k} p_{m,k} = 0 \quad (40)$$

where ∇ represents taking the derivative of $p_{m,k}$. The condition that makes (40) always true is:

$$\theta_{m,k} = 0 \quad (41)$$

Substituting (41) into (39), then the optimal value $p_{m,k}$ denoted as $p^*_{m,k}$ can be expressed as:

$$\begin{aligned} p^*_{m,k} &= \max\left(\frac{v\beta_m}{\ln 2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} - \frac{\sigma^2}{\lambda^2_{m,k}}, 0 \right) \\ &= \left(\frac{v\beta_m}{\ln 2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} - \frac{\sigma^2}{\lambda^2_{m,k}} \right)^+ \end{aligned} \quad (42)$$

where $(a)^+$ represents the maximum one between a and 0. Substituting (41) into (39), then we obtain the achievable data rate R^*_m with the optimal value$p^*_{m,k}$:

$$\begin{aligned} R^*_m = & \sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \log_2\left(1 + \frac{p_{m,k}\lambda^2_{m,k}}{\sigma^2}\right) \\ = \max & \left(\sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \right. \\ & \left. \times \log_2\left(\frac{v\lambda^2_{m,k}\beta_m}{\ln 2 \cdot \sigma^2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} \right), 0 \right) \\ = & \left(\sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \right. \\ & \left. \times \log_2\left(\frac{v\lambda^2_{m,k}\beta_m}{\ln 2 \cdot \sigma^2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} \right) \right)^+ \end{aligned} \quad (43)$$

Substituting (42) and (43) into (36), we have:

$$\begin{aligned} & h'(\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ & = \min_{\lambda_{m,k}} \Bigg(e \cdot \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \left|D_m(i,k)\right|^2 \cdot \\ & \left(\frac{v\beta_m}{\ln 2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} - \frac{\sigma^2}{\lambda^2_{m,k}} \right)^+ \\ & - \upsilon \sum_{m=1}^{M} \left(\sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \right. \\ & \left. \times \log_2\left(\frac{v\lambda^2_{m,k}\beta_m}{\ln 2 \cdot \sigma^2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} \right) \right)^+ \\ & + \sum_{i=1}^{N_T} \sum_{m=1}^{M} \sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \gamma_i \left|D_m(i,k)\right|^2 \cdot \\ & \left(\frac{v\beta_m}{\ln 2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} - \frac{\sigma^2}{\lambda^2_{m,k}} \right)^+ \\ & + \sum_{m=1}^{M} \alpha_m \left(\sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \right. \\ & \left. \times \log_2\left(\frac{v\lambda^2_{m,k}\beta_m}{\ln 2 \cdot \sigma^2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} \right) \right)^+ \\ & - \sum_{m=1}^{M} \beta_m \left(\sum_{k=1}^{\min\left(L_m, N_T-\hat{L}_m\right)} \right. \\ & \left. \times \log_2\left(\frac{v\lambda^2_{m,k}\beta_m}{\ln 2 \cdot \sigma^2\alpha_m \sum_{i=1}^{N_T} (e+\gamma_i) \left|D_m(i,k)\right|^2} \right) \right)^+ \Bigg) \end{aligned} \quad (44)$$

Obviously, (44) is a function of $\lambda_{m,k}$ and $\lambda_{m,k}$ is determined by user selection.

2) USER SELECTION

In the user selection step, we not only solve (44) but also guarantee the data delay requirements. In Table 1, we assume that the service type of user m is z and we denote r_z and d_z as data rate request and delay requirement respectively. According to the QoS delay requirements in Table 1, the delay requirement d_z of service z can be transformed into the maximum number of waiting slots n_z, that is, $n_z = d_z \big/ tti$ with tti being the length of the time slot. $W_{m,z}$ denotes the number of waiting slots when user m using service z. In user selection, the user whose waiting slots reach the maximum waiting slots is first selected to guarantee the delay requirement, and the other users are selected to minimize $h^{'}(\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta})$. Therefore, the proposed resource allocation algorithm that simultaneously optimizes the energy efficiency and the effective system throughput is as follows:

Step 1: Initialize $\boldsymbol{\upsilon}_{\min} = \mathbf{0}$, $\boldsymbol{\alpha}_{\min} = \boldsymbol{\gamma}_{\min} = \boldsymbol{\beta}_{\min} = \mathbf{0}$, $\upsilon_{\max} \gg 0$, $\boldsymbol{\alpha}_{\max}, \boldsymbol{\gamma}_{\max}, \boldsymbol{\beta}_{\max} \gg \mathbf{0}$.

Step 2: Initialize $\upsilon = \frac{\upsilon_{\min}+\upsilon_{\max}}{2}$;

$$\gamma_i = \frac{\gamma_{i\min} + \gamma_{i\max}}{2}, \quad \forall i = 1, 2, \ldots, N_T;$$
$$\beta_m = \frac{\beta_{m\min} + \beta_{m\max}}{2}, \quad \forall m = 1, 2, \ldots, M;$$
$$\alpha_m = \frac{\alpha_{m\min} + \alpha_{m\max}}{2}, \quad \forall m = 1, 2, \ldots, M.$$

Step 3: Initialize the set of unselected users as $\Omega = \{1, 2, \cdots, K\}$ and the set of selected users as $\Psi = \phi$.

Step 4: Compute $W_{m,z}$ of every user in set Ω. If $W_{m,z} = n_z$, select user m. Then, update the set of selected users $\Psi = \{m : W_{m,z} = n_z\}$ and the set of unselected users $\Omega = \Omega - \Psi$. If $\Psi = \phi$ go to step 5, otherwise go to step 6.

Step 5: Compute $\hat{L}_m$ of every user in set Ψ, and judge whether it satisfies that $\hat{L}_m \leq N_T$ for any user m. If the result is true, compute $h^{'}(\lambda_{m,k})$ of user m in set Ω, select a user who can minimize $h^{'}(\lambda_{m,k})$ as m_1, update $\Psi = \Psi + \{m_1\}$, $\Omega = \Omega - \{m_1\}$ and $h^{'}(\lambda_{m,k})$; otherwise go to step 8.

Step 6: Compute $\hat{L}_m$ of every user in set Ψ, and judge whether it satisfies that $\hat{L}_m \leq N_T$ for any user m. If the result is true, for each user m in set Ω, define $\Psi_m = \Psi + \{m\}$ and compute $h^{'}(\lambda_{m,k})$ of Ψ_m, select a user who can minimize $h^{'}(\lambda_{m,k})$ as m^*, update $\Psi = \Psi + \{m^*\}$, $\Omega = \Omega - \{m^*\}$ and $h^{'}(\lambda_{m,k})$; otherwise go to step 8.

Step 7: Repeat Step 4.

Step 8: Compute $p_{m,k}$ by substituting $\upsilon, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}$ and $\lambda_{m,k}$ into (42).

Step 9: Substitute $\lambda_{m,k}$ and $p_{m,k}$ into (13) to compute $\sum_{m=1}^{M} R_m$. If $\sum_{m=1}^{M} R_m \geq C_0$, set $\upsilon_{\max} = \upsilon$, $\boldsymbol{\alpha}_{\min} = \boldsymbol{\alpha}$, $\boldsymbol{\beta}_{\max} = \boldsymbol{\beta}$, $\boldsymbol{\gamma}_{\min} = \boldsymbol{\gamma}$, otherwise set $\upsilon_{\min} = \upsilon$, $\boldsymbol{\alpha}_{\max} = \boldsymbol{\alpha}$, $\boldsymbol{\beta}_{\min} = \boldsymbol{\beta}$, $\boldsymbol{\gamma}_{\max} = \boldsymbol{\gamma}$.

Step 10: Repeat step 2 to step 9 until $\upsilon_{\max} - \upsilon_{\min} \leq \delta$, $\boldsymbol{\alpha}_{\max} - \boldsymbol{\alpha}_{\min} \leq \delta$, $\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min} \leq \delta$, $\boldsymbol{\beta}_{\max} - \boldsymbol{\beta}_{\min} \leq \delta$, where δ denotes the constant that we set to control the accuracy of the algorithm.

We evaluate the computational complexity of the proposed algorithm, which is $O\left(KMN_T^3 \log \omega / \delta\right)$. The computational complexity is at the same level as the throughput-based algorithm which only optimizes the throughput of the system [40]. The computational complexity of the throughput-based algorithm is $O\left(KM^2N_T^3\right)$. The complexity derivation of the proposed algorithm is shown in Appendix.

V. SIMULATION RESULTS AND DISCUSSIONS

We consider an MU-MIMO system with a single BS. The upper limit of the transmission power of each antenna P_0 is 10W. The drain efficiency e and the circuit power dissipation of base station power amplifier P_c are 1/0.38 and 10W, respectively. The number of antennas N_T is 20, and the total number of users K is 30. We assume that the number of receiving antennas for each user is 3, that is, $n_m = 3, \forall m$. We denote the transmit signal-to-noise ratio as $\mathrm{SNR} = P_{m,k} / \sigma^2$, the value of SNR is changed by changing the value of σ^2 in the simulation. We assume that the number of real-time users and non-real time users are equal and the achievable data rate is calculated on the base of bandwidth normalization.

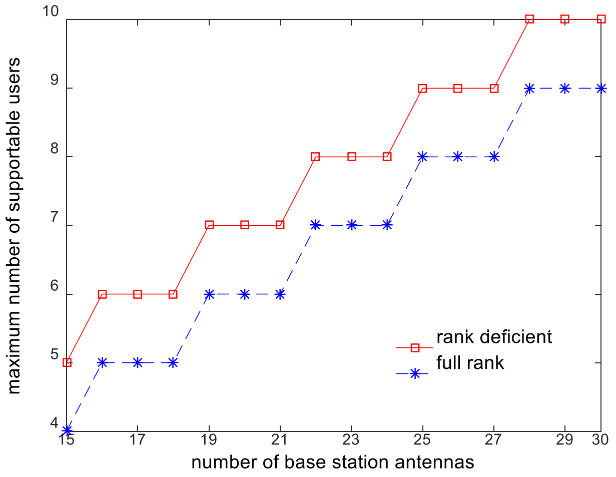


FIGURE 2. **Maximum number of supportable users versus the number of base station antennas.**

Fig. 2 shows that the maximum number of simultaneously supportable users versus the number of base station antennas in the full rank and rank deficient environments. As we can see from the figure, the maximum number of simultaneously supportable users in the system increases with the number of base station antennas when the total number of users is a constant. Meanwhile, the maximum number of supportable users in the rank deficient environment is larger than that in the full rank environment with the same number of base station antennas. The reason can be found in the constraint of the maximum number of simultaneously supportable users (eq. (6)).

Fig. 3. shows the energy efficiency versus the number of transmitting antennas. As it can be seen from Fig. 3, the energy efficiency of the system increases with the number of base station antennas when the total number of users is a constant. The reason is that the number of users that can be selected simultaneously in the system is proportional to the number of transmitting antennas. However, when the increasing number of antennas reaches a certain value, the energy efficiency will not increase with the number of antennas due to the limitation of the total number of users in the system. In Fig. 3, the energy efficiency is fluctuant because the proposed algorithm in the user selection step is essentially an improved greedy algorithm which can only reach the local optimal result instead of the global optimal result in each user selection process. Therefore, the result obtained in each

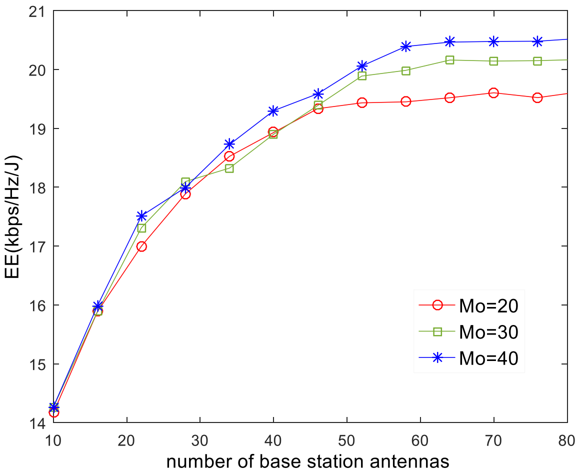


FIGURE 3. Energy efficiency versus the number of base station antennas.

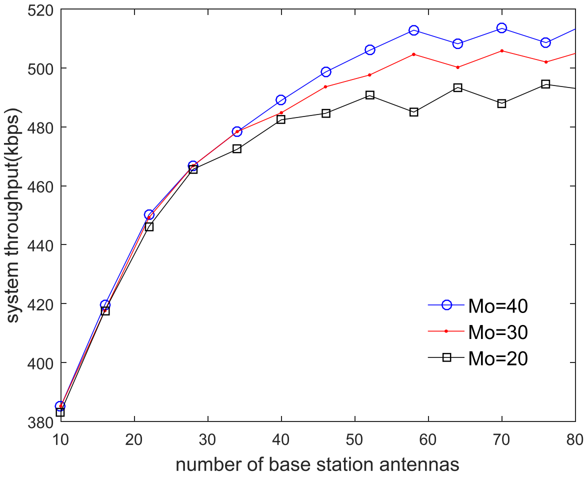


FIGURE 4. System throughput versus the number of base station antennas.

user selection process is uncertain, which has an uncertain gap with the global optimal result. In addition, we can also see from Fig. 3 that when the number of antennas exceeds a certain number, more total users in the system will lead to higher energy efficiency. That is because "better" users that make the local optimal result closer to the global optimal result can be selected.

Fig. 4. shows the system throughput versus the number of antennas in BS. As it can be seen from Fig. 4, the system throughput increases with the number of base station antennas when the total number of users in the system is a constant. The reason is that the maximum number of selected users increases with the number of antennas, thus the system throughput increases.

Fig. 5 shows the comparison of energy efficiency among different algorithms under different SNR. In Fig. 5, the difference between the algorithm A and the proposed algorithm is that the algorithm A is a single objective algorithm which only optimizes the energy efficiency. The algorithm B is obtained from the algorithm A without considering the upper bound of the rate requirements; the algorithm C presented in [21] only maximized the system throughput in the full rank channel environment without considering the upper bound of the rate requirements; the algorithm D is same as the proposed algorithm except that D does not consider the existence of the rank deficient channel matrix. The details of these five algorithms are shown in Table 2.

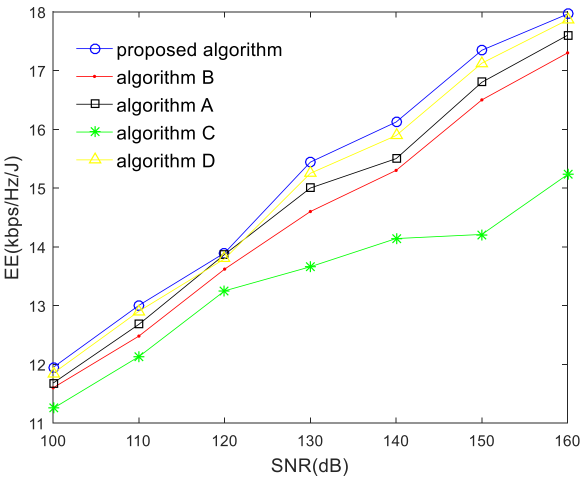


FIGURE 5. Comparison of energy efficiency among different algorithms.

TABLE 2. Algorithms comparison table.

Algorithm	General Environment	Double-Objective	The Upper Bound f QoS Rate Requirements
Algorithm A	√	×	√
Algorithm B	√	×	×
Algorithm C	×	×	×
Algorithm D	×	√	√
Proposed Algorithm	√	√	√

As it can be seen from Fig. 5, the energy efficiency increases with SNR when the total number of users in the system is a constant. The reason is that the increase of signal-to-noise ratio means the reduction of noise power, and hence the higher achievable bit rate can be obtained with a certain transmission power. In addition, the energy efficiency of the proposed algorithm and algorithm D is larger than that of algorithm A, B and C, which indicates that the double-objective optimization algorithm is superior to the single-objective optimization algorithm in optimizing energy efficiency. Meanwhile, the energy efficiency of algorithm A is larger than that of algorithm B, because algorithm B does not consider the upper bound of the rate requirements and rate exceeding the upper bound is ineffective. Algorithm C is designed only in the full rank channel matrix environment, thereby its energy efficiency significantly lower than that of other algorithms in the real environment (which contains full rank and rank deficient channels). In the rank deficient environment, i.e., in the simulation, the maximum number of simultaneously supportable users of algorithm D is smaller than that of the proposed algorithm, thereby the effective

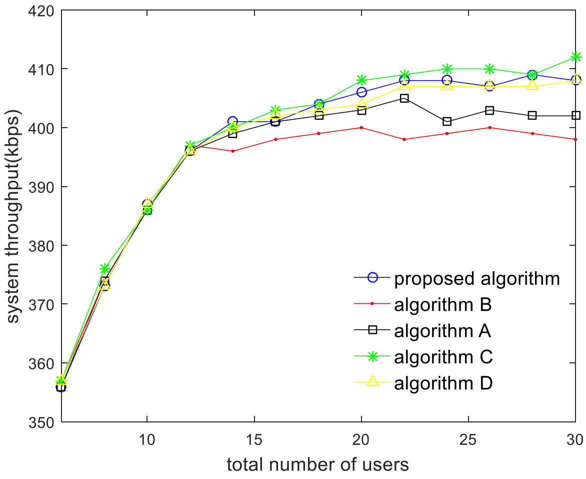


FIGURE 6. Comparison of system throughput among different algorithms.

system throughput and the energy efficiency of algorithm D is smaller than the proposed algorithm.

Fig. 6 shows the comparison of system throughput among different algorithms. It can be seen from Fig. 6 that the system throughput increases with the total number of users in the system when the number of base station antennas is a constant. In addition, the system throughput will not increase when the total number of users in the system reaches a certain value, where the number of simultaneously supportable users reaches its upper bound. It also can be seen from Fig. 6 that the system throughput of the proposed algorithm and algorithm D is larger than that of algorithm A and B, but slightly smaller than that of algorithm C. The reason is that algorithm C only maximizes the system throughput without considering the optimization of system power at all while the proposed resource allocation algorithm optimizes both.

VI. CONCLUSION

In this paper, we propose a resource allocation algorithm for MU-MIMO systems in the general wireless environment with full rank and rank deficient channels. The resource allocation algorithm simultaneously maximizes the energy efficiency and the effective system throughput through power allocation and user selection. The proposed resource allocation algorithm guarantees the different QoS requirements of heterogeneous services, and it considers the upper and the lower bounds of the QoS rate requirements. Due to the excessive constraints, the optimization problem established in this paper is not a standard convex optimization problem and cannot be directly solved by convex optimization method. We construct the Lagrange dual problem of the original problem and then prove the strong duality between the original problem and its dual problem. When solving the dual problem, we firstly simplify and decompose the problem, such that power allocation and user selection can be applied separately to solve the problem. Simulation results show that the proposed double-objective optimization algorithm is superior to the existing single-objective optimization algorithms in terms of energy efficiency and effective system throughput. In addition, compared with the existing algorithms that only consider the full rank channels, the proposed algorithm can select more users in the rank deficient environment, thereby further enlarging the energy efficiency.

APPENDIX

The computational complexity of the proposed algorithm can be counted as the number of flops, where a flop is equal to a real floating point operation. A real operation is counted as one flop, and complex addition and multiplication operations are considered as two flops and six flops, respectively.

For a complex matrix $\boldsymbol{H} \in C^{m\times n}$ $(m \le n)$, the complexity of typical matrix operations is summarized as follows [41]:

1. The flop count for SVD is $48m^2n + 24mn^2 + 54m^3$.

2. Matrix multiplication of a $m \times n$ complex matrix and a $n \times p$ complex matrix has $8mnp$ flops.

In each iteration, the derivation of the proposed algorithm's complexity is as follows.

1. The initialization of the Lagrange multipliers takes $2 + 2N_T + 4M$ flops.

2. For $i = 1$: It takes K to compute the delay of all users.

3. For $i \ge 2$: It is noted that the complexity derivation of the proposed algorithm in the following is in full rank environment. And the flops needed in the rank deficient environment are less than which in full rank environment. Therefore, computational complexity we derive in the following is the upper bound of computational complexity of the algorithm.

1) SVD is used to get $\lambda_{m,k}$ for each user in Ω.

SVD of $\hat{\boldsymbol{H}}_m$ takes

$48\,(i-1)^2\, n_m^2 N_T + 24\,(i-1)\, n_m N_T^2 + 54\,(i-1)^3\, n_m^3$ flops.

To compute $\bar{\boldsymbol{H}}_m$ takes $8\,(i-1)\, n_m\,[N_T - (i-1)\, n_m]$ flops. For a simple analysis, $N_T - (i-1)\, n_m$ is replaced to N_T. Then, the flop count to compute $\bar{\boldsymbol{H}}_m$ can be rewritten as $8\,(i-1)\, n_m N_T$ flops.

SVD of $\hat{\boldsymbol{H}}_m$ takes

$48n_m^2\,[N_T - (i-1)\, n_m] + 24n_m\,[N_T - (i-1)\, n_m]^2 + 54n_m^3$ flops. Similarly, the flop count for SVD of $\hat{\boldsymbol{H}}_m$ can be rewritten as $48n_m^2 N_T + 24n_m N_T^2 + 54n_m^3$ flops.

2) To compute $h^{'}(\lambda_{m,k})$ takes

$2iN_T n_m\,(3N_T + 10) + 9in_m\,(N_T + 3)$ flops.

3) To compute the optimal value of $p_{m,k}$ and $\sum_{m=1}^{M} R_m$ takes $3N_T + 7 + 4n_m$ flops.

Therefore, in the rank deficient environment, the upper bound of total complexity of an iteration is as follow:

$$\begin{aligned}\Psi &= 2 + 2N_T + 4M + K + \sum_{i=2}^{M} [K - (i-1)]\{\\ &48\,(i-1)^2\, n_m^2 N_T + 24\,(i-1)\, n_m N_T^2 + 54\,(i-1)^3\, n_m^3 +\\ &8\,(i-1)\, n_m N_T + 48\,(i-1)\, n_m^2 N_T + 24 n_m N_T^2 + 54 n_m^3 +\\ &2iN_T n_m\,(3N_T + 10) + 9in_m\,(N_T + 3) + 3N_T + 7 + 4n_m\}\\ &\approx O\left(KMN_T^2 n_m\right) \approx O\left(KMN_T^3\right)\end{aligned}$$

In addition, the computational complexity of the number of iterations is the maximum value of $O\left(\log \upsilon_{\max}/\delta\right)$,

$O\left(\log \boldsymbol{\alpha}_{\max}/\delta\right)$, $O\left(\log \boldsymbol{\gamma}_{\max}/\delta\right)$, and $O\left(\log \boldsymbol{\beta}_{\max}/\delta\right)$, which is defined as $O\left(\log \omega/\delta\right)$, ω is a constant. Therefore, the upper bound of total complexity of the proposed algorithm is $O\left(KMN_T^3 \log \omega/\delta\right)$.

REFERENCES

[1] H. T. Dao and S. Kim, "Pilot power allocation for maximising the sum rate in massive MIMO systems," *IET Commun.*, vol. 12, no. 11, pp. 1367–1372, Jul. 2018.

[2] A. M. Benaya, A. A. Rosas, and M. Shokair, "Maximization of minimal throughput using genetic algorithm in MIMO underlay cognitive radio networks," in *Proc. NRSC*, Aswan, Egypt, Feb. 2016, pp. 141–148.

[3] X. Xiao, X. Tao, and J. Lu, "Energy-efficient resource allocation in LTE-based MIMO-OFDMA systems with user rate constraints," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 185–197, Jan. 2015.

[4] B. Zhang, Y. Wang, W. Wang, and Y. Tian, "On the downlink throughput capacity of hybrid wireless networks with MIMO," *IEEE Access*, vol. 5, pp. 26086–26091, 2017.

[5] S. E. El-Khamy, K. H. Moussa, and A. A. El-Sherif, "Capacity-maximized transmitter antenna selection for multi-user massive MIMO systems with linear beamforming," in *Proc. NRSC*, Aswan, Egypt, Feb. 2016, pp. 333–339.

[6] T. Qi, Y. Wang, and X. Feng, "Performance analysis of downlink user selection in multiuser MIMO system," in *Proc. ITNEC*, Chengdu, China, Dec. 2017, pp. 122–126.

[7] L.-N. Tran, M. Bengtsson, and B. Ottersten, "Iterative precoder design and user scheduling for block-diagonalized systems," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3726–3739, Jul. 2012.

[8] H. Huang, C. B. Papadias, and S. Venkatesan, *MIMO Communication for Cellular Networks*. New York, NY, USA: Springer, 2012.

[9] D. Gesbert, H. Bolcskei, D. A. Gore, and A. J. Paulraj, "Outdoor MIMO wireless channels: Models and performance prediction," *IEEE Trans. Commun.*, vol. 50, no. 12, pp. 1926–1934, Dec. 2002.

[10] A. Pollok, W. G. Cowley, and I. D. Holland, "Multiple-input multiple-output options for 60 GHz line-of-sight channels," in *Proc. AusCTW*, Christchurch, New Zealand, Jan./Feb. 2008, pp. 101–106.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[12] S. Pan, W. Zhou, Q. Gu, and Q. Ye, "Network selection algorithm based on spectral bandwidth mapping and an economic model in WLAN & LTE heterogeneous networks," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 1, pp. 68–86, Jan. 2015.

[13] G. Femenias, F. Riera-Palou, and J. Pastor, "Resource allocation in block diagonalization-based multiuser MIMO-OFDMA networks," in *Proc. ISWCS*, Barcelona, Spain, Aug. 2014, pp. 411–417.

[14] R. Yang, E. Sun, J. Zhang, and Y. Zhang, "Resource allocation and limited feedback for multimedia traffics in multiuser MIMO-OFDM systems," in *Proc. MMIT*, Kaifeng, China, Apr. 2010, pp. 242–245.

[15] C.-Y. Hsu, M. Wang, P. L. Yeoh, and B. S. Krongold, "Continuous and discrete sum-rate maximization for multiuser MIMO-OFDM systems with CoMP," in *Proc. IEEE ICC*, Austin, TX, USA, Dec. 2014, pp. 3903–3909.

[16] M. R. Tabany and C. G. Guy, "An end-to-end QoS performance evaluation of VoLTE in 4G E-UTRAN-based wireless networks," in *Proc. ICWMC*, Seville, Spain, 2014, pp. 90–97.

[17] A. Al-Shaikhli, A. Esmailpour, and N. Nasser, "Quality of service interworking over heterogeneous networks in 5G," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[18] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Beijing, China: House of Electronics Industry, 2002.

[19] J. Kaleva, A. Tölli, and M. Juntti, "Decentralized sum rate maximization with QoS constraints for interfering broadcast channel via successive convex approximation," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2788–2802, Jun. 2016.

[20] V. Valls and D. J. Leith, "Proportional fair MU-MIMO in 802.11 WLANs," *IEEE Wireless Commun. Lett.*, vol. 3, no. 2, pp. 221–224, Apr. 2014.

[21] P. He and L. Zhao, "Optimal power allocation for maximum throughput of general MU-MIMO multiple access channels with mixed constraints," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 1042–1054, Mar. 2016.

[22] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[23] I. Humar, X. Ge, L. Xiang, M. Jo, M. Chen, and J. Zhang, "Rethinking energy efficiency models of cellular networks with embodied energy," *IEEE Netw.*, vol. 25, no. 2, pp. 40–49, Mar./Apr. 2011.

[24] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 545–554, Feb. 2010.

[25] C. Isheden and G. P. Fettweis, "Energy-efficient link adaptation with transmitter CSI," in *Proc. IEEE WCNC*, Cancun, Mexico, Mar. 2011, pp. 1381–1386.

[26] C. Isheden, Z. Chong, E. Jorswieck, and G. P. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.

[27] X. Xiao, X. Tao, and J. Lu, "QoS-guaranteed energy-efficient power allocation in downlink multi-user MIMO-OFDM systems," in *Proc. IEEE ICC*, Sydney, NSW, Australia, Jun. 2014, pp. 3945–3950.

[28] K. M. S. Huq, S. Mumtaz, J. Rodriguez, and R. L. Aguiar, "Energy efficiency optimization in MU-MIMO system with spectral efficiency constraint," in *Proc. IEEE ISCC*, Funchal, Portugal, Jun. 2014, pp. 1–5.

[29] S. Nam, J. Kim, and Y. Han, "A user selection algorithm using angle between subspaces for downlink MU-MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 616–624, Feb. 2014.

[30] L. Sun and M. Lei, "Adaptive joint nonlinear transmit-receive processing for multi-cell MIMO networks," in *Proc. IEEE GLOBECOM*, Anaheim, CA, USA, Dec. 2012, pp. 3766–3771.

[31] Y. Zeng, X. Xu, Y. L. Guan, E. Gunawan, and C. Wang, "Degrees of freedom of the three-user rank-deficient MIMO interference channel," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4179–4192, Aug. 2014.

[32] *Digital Cellular Telecommunications System (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Policy and Charging Control Architecture*, document 3GPP TS 23.203, V8.10.0, 2010.

[33] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

[34] Z. Xu, C. Yang, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient power allocation between pilots and data symbols in downlink OFDMA systems," in *Proc. IEEE GLOBECOM*, Kathmandu, Nepal, Dec. 2011, pp. 1–6.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[36] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2014.

[37] V. W. S. Wong *et al.*, *Key Technologies for 5G Wireless Systems*, Cambridge, U.K.: Cambridge Univ. Press, 2017.

[38] Y. Sun, D. W. K. Ng, J. Zhu, and R. Schober, "Multi-objective optimization for robust power efficient and secure full-duplex wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5511–5526, Aug. 2016.

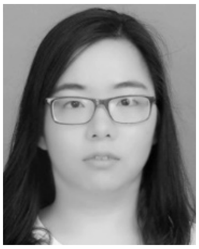
[39] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

[40] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, Jr., and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.

[41] X. Zhang and J. Lee, "Low complexity MIMO scheduling with channel decomposition using capacity upperbound," *IEEE Trans. Commun.*, vol. 56, no. 6, pp. 871–876, Jun. 2008.

SU PAN received the B.S. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1992, the M.S. degree in information and communications engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1995, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2004.

He is currently a Professor with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. His research interest includes radio resource management in broadband wireless networks.



YAN YAN received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016, where she is currently pursuing the M.S. degree in communication and information system. Her research interests include resource management and wireless resource allocation.

KUSI ANKRAH BONSU received the B.Sc. degree in physics and the M.Sc. degree in telecommunications engineering from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 2003 and 2011, respectively. He is currently pursuing the Ph.D. degree in information and communication engineering with the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include wireless resource allocation, MIMO, and optimization.

WEIWEI ZHOU received the B.Sc. degree in communication engineering from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2014. She is currently pursuing the Ph.D. degree with NJUPT and the University of Technology Sydney (UTS). Her current interests include admission control and resource allocation in the wireless networks.