WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2010; **10**:1493–1508
Published online 21 August 2009 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/wcm.838

Resource allocation in OFDMA networks based on interior point methods

Mehri Mehrjoo[1], Somayeh Moazeni[2] and Xuemin (Sherman) Shen[1*,†]

[1]*Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1*
[2]*School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1*

Summary

This paper studies a joint optimization problem of sub-carrier assignment and power allocation in orthogonal frequency division multiple access (OFDMA) wireless networks. A major challenge in solving the optimization problem is non-convexity caused by the combinatorial nature of sub-carrier assignment problem and/or non-convex objective functions. To address the combinatorial complexity, we formulate the resource allocation problem as an optimization problem with continuous variables. We propose a novel approach based on a penalty function method and an interior point method (PM/IPM) to solve the problem. In specific, using a two-step implementation, the penalty method is applied first to convert the non-convex feasible region to a convex one. Then, the interior point method is deployed to solve the problem which is non-convex only in the objective function. To evaluate the performance of PM/IPM, we apply a genetic algorithm (GA) that achieves near optimal solutions of the problem by iterative searching. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: resource allocation; OFDMA; non-convexity; utility function

1. Introduction

In wireless networks with orthogonal frequency division multiple access (OFDMA), concurrent resource allocation problems, sub-carrier assignment to users and power allocation to sub-carriers, referred as OFDMA resource allocation, affect the network performance dramatically. Since sub-carriers have diverse channel gain on each user's channel, due to fading effect, transmitting a specific amount of data on each sub-carrier demands different amount of power. Therefore, to allocate optimal power to sub-carriers, an OFDMA resource allocation problem is formulated as an optimization problem where the constraints along with the objective function represent the OFDMA restrictions, users' requirements, and the network's objectives.

In practice, OFDMA resource allocation optimization problems are categorized into non-convex problems where either the feasible region, i.e.,the set

*Correspondence to: Xuemin (Sherman) Shen, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.
†E-mail: xshen@uwaterloo.ca

of feasible allocations that satisfy all constraints, or the objective function are non-convex.[‡] The restrictions imposed by OFDMA network specifications and users' requirements determine the feasible region. A major restriction of OFDMA network is exclusive sub-carrier assignment which does not allow a sub-carrier to be allocated to more than one user. The exclusive sub-carrier assignment causes the feasible region to be discrete and consequently non-convex (See Appendix A). The objective function of the optimization problem depends on users' demand and the network service provider's goals. A variety of linear convex to nonlinear non-convex objective functions (of rate), such as maximizing sum of allocated rate to users and maximizing aggregate users' utilities, respectively, can be considered in practice [1,2]. Deploying nonlinear or non-convex objective functions along with the non-convexity of the feasible region contribute to the difficulty and non-convexity of OFDMA optimization problems.

In general, non-convex optimization problems are NP-hard [3], and there is no polynomial time algorithm to find their global optimal solutions. Therefore, OFDMA optimization problems are usually solved for a local optimal solution by either search algorithms or some convex optimization techniques when the feasible region and objective function are approximated with convex ones [1,2,4,5]. Search algorithms span almost the entire feasible region of the problem to find the highest local maximum (or lowest local minimum). As they do not stop searching after finding a local optimum, it is expected that the algorithms achieve near optimal solutions when searching time approaches infinity. However, the long response time of search algorithms limits their usage for real-time applications. On the contrary, using convex optimization techniques usually shortens the execution time to obtain the solution. Nevertheless, convex relaxation of the objective function is not practical for many applications, such as network utility maximization (NUM) problems where utilities represent users' perceived quality of service (QoS) and can be non-convex.

The shortcomings of search algorithms and convex relaxation approaches in solving non-convex problems limit the developing of elaborated OFDMA resource allocation schemes, while OFDMA is emerging in broadband wireless networks, and the OFDMA resource allocation arises in many contexts. In this paper, we investigate new optimization approaches that can treat the non-convexity of the OFDMA optimization problem, specifically when the objective function is non-convex. The rationale is that many problems in network flow control [6], utility fairness [7,8], and resource allocation for heterogeneous traffic types are formulated by an OFDMA optimization problem with a non-convex objective function which cannot be relaxed with a convex one. Although a great success has been achieved in solving the OFDMA optimization problem when the objective function is convex [2,9–18], solving the problem without convex relaxation of the feasible space or the objective function has not been addressed in the literature.

Our approach to treat the non-convexity of the OFDMA optimization problem is based on continuous optimization approaches. It is highly expected that interior point methods will be successful in solving continuous nonlinear problems, particularly with convex feasible regions [19–21]. Accordingly, we focus on formulating the OFDMA resource allocation problem into continuous optimization one [22], and propose an interior point penalty method to solve the problem. More precisely, using a penalty function method, non-convex constraints multiplied by a large coefficient, which penalizes the constraints deviations, are added to the objective function. Then, the new problem with convex feasible region is solved by the interior point penalty method. The penalty function method combined with an interior point method (PM/IPM) is applied to the OFDMA resource allocation in a comprehensive form so that users can have heterogeneous rate requirements and the objective function of the resource allocation scheme can be non-convex. To compare the solutions obtained by PM/IPM with near optimal solutions obtained by an iterative search algorithm, we implement a genetic algorithm (GA). We allow for a large number of search iterations to ensure that the solutions offered by GA are closed to global optimal solutions with a high level of confidence [23] and can be considered as benchmarks to assess the performance of the solutions obtained by PM/IPM.

The contribution of the paper is threefold. First, it presents a continuous optimization problem formulation for the OFDMA sub-carrier assignment and power allocation. Second, it deploys a continuous nonlinear technique based on a combination of a penalty method and an interior point method, PM/IPM, to solve the problem. Third, it implements a GA to compare the performance of the resource allocation obtained by PM/IPM with the ones obtained from the GA.

[‡]A function f is *convex* if the domain of f, D_f, is a convex set, i.e., $(1-t)x + ty \in \mathrm{D}_f$ for every x, $y \in \mathrm{D}_f$ and $t \in [0, 1]$, and $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$.

DOI: 10.1002/wcm

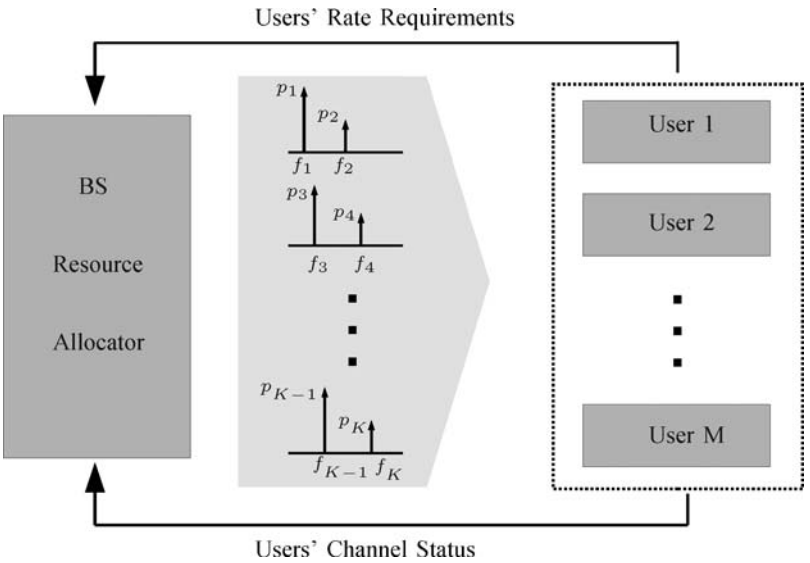

Figure 1. Network platform.

The remainder of the paper is organized as follows. The problem formulation for the resource allocation problem is presented in Section 2, followed by a discussion about the problem complexity and performance. In Section 3 the proposed method, PM/IPM, for solving the resource allocation problem is presented. An iterative algorithm based on GA is described in Section 4, and numerical results are presented in Section 5. The paper is concluded in Section 6.

2. Problem Formulation

We consider a network consisting of a central controller, named base station (BS), and several users located in one hop neighborhood from BS in a point to multi-point (PMP) manner. A resource allocator in the BS assigns sub-carriers and allocates a fraction of the total BS power, P_{BS}, to each user based on the users' requirements and resource constraints, as shown in Figure 1. In Subsection 2.1, we present two optimization programming formulations for the resource allocation scheme in this network: a mixed integer nonlinear programing (MINLP) problem and a nonlinear programming (NLP) problem. A discussion about the computational complexity of the problems is followed in Subsection 2.2.

2.1. MINLP and NLP Problems

Due to discrete nature of sub-carriers and continuous nature of power, first, we formulate the problem as an MINLP problem, i.e., a mathematical programming problem with both integer and continuous variables, whose constraints or objective function are nonlinear. Then, we prove that the set of constraints including the integer variables, in the MINLP problem, can be substituted by a set of nonlinear constraints with continuous variables. Accordingly, we propose an NLP problem that unifies sub-carrier assignment and power allocation in a rate (or power) allocation problem. For more readability of formulas, the network parameters used in the optimization problems are given in Table I.

A solution of the resource allocation problem is denoted by a rate allocation vector $\mathbf{r}$ or a power allocation vector $\mathbf{p}$ as below:

$$\mathbf{r} = [r_{11}, r_{12}, \cdots, r_{1K}, \cdots, r_{M1}, \cdots, r_{MK}]^T \quad (2.1)$$

$$\mathbf{p} = [p_{11}, p_{12}, \cdots, p_{1K}, \cdots, p_{M1}, \cdots, p_{MK}]^T \quad (2.2)$$

Table I. Notations descriptions.

Notation	Description
M	Total number of users in the network
K	Total number of sub-carriers in the network
i	User index belongs to $\mathcal{M} := \{1, 2, \cdots, M\}$
j	Sub-carrier index belongs to $\mathcal{K} := \{1, 2, \cdots, K\}$
α_{ij}	Channel gain of user i on sub-carrier j
p_{ij}	Allocated power to user i on sub-carrier j
r_{ij}	Allocated rate to user i on sub-carrier j
r_i	Total allocated rate to user i
$R^i_{\min}$	Minimum service rate requirement of user i
B	Network bandwidth
P_{BS}	BS total power budget

DOI: 10.1002/wcm

Similarly, the sub-carrier assignment vector is denoted by $\mathbf{c}$, where

$$\mathbf{c} = [c_{11}, c_{12}, \cdots, c_{1K}, \cdots, c_{M1}, \cdots, c_{MK}]^T \quad (2.3)$$

and c_{ij} is

$$c_{ij} = \begin{cases} 1 & \text{if sub-carrier } j \text{ is assigned to user } i \\ 0 & \text{otherwise} \end{cases} \quad (2.4)$$

Every user can use several sub-carriers, but each sub-carrier can be assigned to at most one user. Mathematically, this restriction is given by

$$\sum_{i=1}^{M} c_{ij} \leq 1 \quad \forall j \in \mathcal{K} \quad (2.5)$$

If sub-carrier j has not been assigned to user i, then allocated power to user i on sub-carrier j must be zero. Therefore, for every user $i \in \mathcal{M}$ and every sub-carrier $j \in \mathcal{K}$, we must have the following condition:

$$\text{if } c_{ij} = 0 \text{ then } p_{ij} = 0 \quad (2.6)$$

We include this restriction through the following constraint:

$$p_{ij} \leq P_{\text{BS}} c_{ij} \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \quad (2.7)$$

Note that, if $c_{ij} = 0$, (2.7) implies $p_{ij} \leq 0$ that along with the non-negativity constraint $p_{ij} \geq 0$ yields $p_{ij} = 0$ and satisfies (2.6). When $c_{ij} = 1$, (2.7) is reduced to the redundant constraint $p_{ij} \leq P_{\text{BS}}$, because of the existence of the following constraint, which assures total allocated power to the sub-carriers in each time slot is limited to P_{BS}:

$$\sum_{i=1}^{M} \sum_{j=1}^{K} c_{ij} p_{ij} \leq P_{\text{BS}} \quad (2.8)$$

As (2.7) includes (2.6), variables c_{ij}'s can be removed from (2.8) as follows:

$$\sum_{i=1}^{M} \sum_{j=1}^{K} p_{ij} \leq P_{\text{BS}} \quad (2.9)$$

If noise spectral density is equal to one and rate adaptation is assumed to be continuous [24], the approximate transmission rate for user i on sub-carrier j, r_{ij}, is given by:

$$r_{ij} = \frac{B}{K} \log_2 \left(1 + \alpha_{ij} p_{ij}\right). \quad (2.10)$$

Moreover, QoS requirements are projected on the objective function and constraints. $R^i_{\min}$, the minimum service rate requirement of user i with rate r_i is guaranteed through the following constraint:

$$r_i = \sum_{j=1}^{K} r_{ij} \geq R^i_{\min} \quad \forall i \in \mathcal{M} \quad (2.11)$$

Also, QoS requirements of users can be taken into account through users' utilities, which represent users' satisfaction of allocated rate. However, to present a general problem that unifies most of the existing problems for OFDMA resource allocation, a general objective function $\mathcal{F}(\mathbf{r})$, is used in this subsection. $\mathcal{F}(\mathbf{r})$ can be substitute by any function of rate, such as, sum of users' weighted rate, $\sum \omega_i r_i$, or sum of users' utilities, $\sum u_i(r_i)$, where ω_i and u_i are the assigned weight and utility to user i. The optimization problem becomes:

$$P_1: \max_{\mathbf{c},\mathbf{p}} \ \mathcal{F}(\mathbf{r}) \quad (2.12a)$$

$$\text{s.t } r_{ij} = \frac{B}{K} \log_2 \left(1 + \alpha_{ij} p_{ij}\right) \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \quad (2.12b)$$

$$r_i = \sum_{j=1}^{K} r_{ij} \geq R^i_{\min} \quad \forall i \in \mathcal{M} \quad (2.12c)$$

$$\sum_{i=1}^{M} \sum_{j=1}^{K} p_{ij} \leq P_{\text{BS}} \quad (2.12d)$$

$$\sum_{i=1}^{M} c_{ij} \leq 1 \quad \forall j \in \mathcal{K} \quad (2.12e)$$

$$0 \leq p_{ij} \leq P_{\text{BS}} c_{ij} \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \quad (2.12f)$$

$$c_{ij} \in \{0, 1\} \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \quad (2.12g)$$

Problem P_1 is an MINLP problem. We eliminate integer variables c_{ij}s and formulate the problem as a continuous nonlinear one-stage programming problem P_2:

$$P_2: \max_{\mathbf{p}} \ \mathcal{F}(\mathbf{r}) \quad (2.13a)$$

DOI: 10.1002/wcm

$$\text{s.t}\quad r_{ij} = \frac{B}{K}\log_2\left(1+\alpha_{ij}p_{ij}\right) \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \tag{2.13b}$$

$$r_i = \sum_{j=1}^{K} r_{ij} \geq R_{\min}^{i} \quad \forall i \in \mathcal{M} \tag{2.13c}$$

$$\sum_{i=1}^{M}\sum_{j=1}^{K} p_{ij} \leq P_{\text{BS}} \tag{2.13d}$$

$$p_{\hat{i}j}p_{ij} = 0 \quad \forall j \in \mathcal{K},\ \forall i \in \mathcal{M}\setminus\{\hat{i}\} \tag{2.13e}$$

$$0 \leq p_{ij}, \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K} \tag{2.13f}$$

Proposition 2.1. *There is a one-to-one correspondence between the set of feasible solutions of P_1 and the set of feasible solutions of P_2.*

Proof. We prove it by showing that from each feasible solution of P_2, a feasible solution of P_1 is obtained and *vice versa*.

Let $\mathbf{p}^*$ be a feasible solution of P_2. For every $i \in \mathcal{M}$ and $j \in \mathcal{K}$, define c_{ij}^* as follows:

$$c_{ij}^* = \begin{cases} 1 & \text{if } p_{ij}^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

Clearly $\mathbf{p}^*$ and $\mathbf{c}^*$ satisfy (2.12b), (2.12c), (2.12d), (2.12f) and (2.12g). We claim that this solution also satisfies (2.12e). If this is not true, there exists some $j \in \mathcal{K}$ so that $\sum_{i=1}^{M} c_{ij}^* \geq 2$. This implies that there are at least two i_1 and i_2 such that $c_{i_1 j}^* = c_{i_2 j}^* = 1$. However, the derivation of c_{ij}^* from p_{ij}^* in (2.14) yields $p_{i_1 j}^* > 0$ and $p_{i_2 j}^* > 0$. Hence $p_{i_1 j}^* p_{i_2 j}^* > 0$ which is in contradiction to the fact that $\mathbf{p}^*$ satisfies (2.13e). So $\mathbf{p}^*$, $\mathbf{c}^*$ must also satisfy (2.12e).

Next, assume that $(\mathbf{p}^*, \mathbf{c}^*)$ is a feasible solution of P_1. Thus $\mathbf{p}^*$ satisfies (2.13b), (2.13c), (2.13d) and (2.13f). If $\mathbf{p}^*$ does not satisfy (2.13e), then there must be $\bar{i}, \check{i} \in \mathcal{M}$ and $\bar{j} \in \mathcal{K}$ such that $p_{\bar{i}\bar{j}}^* p_{\check{i}\bar{j}}^* > 0$ or equivalently $p_{\bar{i}\bar{j}}^* > 0$ and $p_{\check{i}\bar{j}}^* > 0$ for some $\check{j}$. Constraint (2.12f) implies that $c_{\bar{i}\bar{j}}^* = 1$ and $c_{\check{i}\bar{j}}^* = 1$. Thus $\sum_{i=1}^{M} c_{i\bar{j}}^* \geq c_{\bar{i}\bar{j}}^* + c_{\check{i}\bar{j}}^* \geq 2$, which is in contradiction to the assumption that $(\mathbf{p}^*,\ \mathbf{c}^*)$ satisfies (2.12e). Thus $\mathbf{p}^*$ also satisfies (2.13e) and therefore, is a feasible solution of P_2. ■

For every feasible solution of P_1 and associated feasible solution of P_2, the rate allocation vectors are identical. Thus, Proposition 2.1 implies there is a one-to-one correspondence between the set of optimal solutions of P_1 and P_2; As a result, they have the same optimal value.

Problem P_2 can be written only in terms of allocated rates r_{ij}, if an equivalent constraint of r_{ij} replaces constraint (2.13e). It can be shown that the following constraints are equivalent to (2.13e):

$$r_{\hat{i}j}r_{ij} = 0 \quad \forall j \in \mathcal{K}, \forall i \in \mathcal{M}\setminus\{\hat{i}\} \tag{2.15a}$$

$$r_{\hat{i}j} + r_{ij} = \max\{r_{\hat{i}j}, r_{ij}\} \quad \forall j \in \mathcal{K}, \forall i \in \mathcal{M}\setminus\{\hat{i}\} \tag{2.15b}$$

$$|r_{\hat{i}j} - r_{ij}| = r_{\hat{i}j} + r_{ij} \quad \forall j \in \mathcal{K}, \forall i \in \mathcal{M}\setminus\{\hat{i}\} \tag{2.15c}$$

$$(r_{\hat{i}j} - r_{ij})^2 = (r_{\hat{i}j} + r_{ij})^2 \quad \forall j \in \mathcal{K}, \forall i \in \mathcal{M}\setminus\{\hat{i}\} \tag{2.15d}$$

We use (a) in the rest of the paper, because they are differentiable and have a simple representation. Thus, P_2 can be restated as follows:

$$P_3:\ \max_{\mathbf{r}}\ \mathcal{F}(\mathbf{r}) \tag{2.16a}$$

$$\text{s.t}\ \sum_{j=1}^{K} r_{ij} \geq R_{\min}^{i} \quad \forall i \in \mathcal{M} \tag{2.16b}$$

$$\sum_{i=1}^{M}\sum_{j=1}^{K} \frac{1}{\alpha_{ij}}\left(2^{\frac{r_{ij}K}{B}} - 1\right) \leq P_{\text{BS}} \quad \forall i \in \mathcal{M}, \forall j \in \mathcal{K} \tag{2.16c}$$

$$r_{\hat{i}j}r_{ij} = 0 \quad \forall i \in \mathcal{M}\setminus\{\hat{i}\}\ \forall j \in \mathcal{K} \tag{2.16d}$$

$$0 \leq r_{ij}, \quad \forall i \in \mathcal{M}, \forall j \in \mathcal{K} \tag{2.16e}$$

The feasible region of P_3 is closed and bounded. Thus, when the objective function is a continuous function of $\mathbf{r}$, *Weierstrass Theorem* [25] implies that P_3 has global optimal solution(s). Although Weierstrass Theorem guarantees that the global optimal solution exists, finding such a global solution for a general continuous objective function is hard, i.e., there is no polynomial time algorithm for obtaining the global optimal solution.

2.2. Related Works and Discussion

In general, objective function $\mathcal{F}$ is a function of users' rates. The choice of $\mathcal{F}$ along with the set of constraints

DOI: 10.1002/wcm

affect both computational complexity of P_3 and the network performance. The following discussion will provide an insight into the problem and review the approaches which have been proposed in the literature for OFDMA problem with different objective functions and constraints.

2.2.1. Linear objective function

Common linear objective functions, used in the OFDMA optimization problems, are $\mathcal{F}(\mathbf{r}) = \sum_i r_i$ and $\mathcal{F}(\mathbf{r}) = \sum_i \omega_i r_i$. The former maximizes total users' data rate and the later maximizes aggregate users' rate multiplied by a vector of weights, ω_i's. Although the objective function is linear, the feasible region of the problem is still non-convex. To simplify the problem, a decomposition method that separates sub-carrier assignment problem from power allocation problem has been suggested in the literature [11–16], and variant schemes have been proposed for each separate problem depending on constraints and problem formulation. For instance, [13] proposes a two-step approach, where in the first step, a sub-carrier is assigned to only one user who has the best channel gain on that sub-carrier. In the second step, the amount of transmit power to be allocated to each sub-carrier is determined by water-filling scheme [26] to maximize overall data rate. To reduce computational complexity of water-filling, equal power allocation scheme may be adopted. It has been shown that water-filling and equal power allocation schemes have only marginal performance difference [27].

2.2.2. Nonlinear objective functions and constraints

To providing QoS and fairness or maximizing resource utilization, some OFDMA resource allocation schemes have been proposed that use nonlinear objective functions or add a set of nonlinear constraints in the optimization problem. Following, we survey some of these schemes.

First, the objective function can be chosen properly to achieve a specific object. For example, an objective function that maximizes minimum users' data rates achieves max–min fair rate allocation [5], and an objective function that maximizes aggregate logarithm of users' data rate [28,29] achieves rate proportional fairness. Similarly, aggregate users' utility, a function that corresponds a user's resource requirement, e.g., rate, to the user's satisfaction of service, may be maximized that obtains maximum resource utilization [1,22,29].

Second, a set of constraints can be added to the problem to force a notion of fairness or QoS. For instance, using a set of nonlinear constraints, [15] maintains a fixed rate ratio among users to achieve fair rate allocation. Similarly, an associated set of constraints to a specific QoS characteristic can be considered to guarantee the required QoS, e.g., [30] provides users' minimum rate requirements, and [16] guarantees tolerable signal to noise ratio of the users' receivers by including corresponding rate and signal to noise ratio constraints to the optimization problem.

3. Penalty Function and Interior Point Methods

We propose a solution based on a combination of penalty function and interior point methods for the NLP problem. Mainly, the approach takes advantage of an interior point method, which can be successfully applied to NLP problems [20]. The success of interior point methods in solving a non-convex nonlinear problem strongly depends on how non-convexity of the problem is treated. We apply a penalty function method to deal with non-convexity of P_3. In other words, by using a penalty function method, we convert P_3 to a new problem with a convex feasible region, and then, we apply an interior point method to solve the problem.

In P_3, all constraints except (2.16d) are convex. We add this set of constraints to the objective function as a penalty term, which is negative when one of the constraints in (2.16d) is violated, and zero otherwise. After adding the penalty term to the objective function, the new objective becomes:

$$P_L: \max_{\mathbf{r}} \quad f(\mathbf{r}) = \mathcal{F}(\mathbf{r}) - \frac{L}{2} \sum_{\hat{i}=1}^{M} \sum_{i=1, i\neq\hat{i}}^{M} \sum_{j=1}^{K} r_{\hat{i}j} r_{ij} \tag{3.1}$$

where positive constant L is the penalty parameter. The new objective function along with the constraints of P_3 form the following problem:

$$P_4: \max_{\mathbf{r}} \quad f(\mathbf{r}) \tag{3.2a}$$

$$\text{s.t} \quad C(\mathbf{r}) \geq 0, \tag{3.2b}$$

where $C(\mathbf{r})$ is the vector of inequality constraints (2.16b), (2.16c) and (2.16e), and is represented as follows:

DOI: 10.1002/wcm

$$C(\mathbf{r}) = \begin{pmatrix} \sum_{j=1}^{K} r_{1j} - R_{\min}^{1} \\ \vdots \\ \sum_{j=1}^{K} r_{Mj} - R_{\min}^{M} \\ -\sum_{i=1}^{M}\sum_{j=1}^{K} \frac{1}{\alpha_{ij}}\left(2^{\frac{Kr_{ij}}{B}} - 1\right) + P_{\mathrm{BS}} \\ r_{11} \\ \vdots \\ r_{MK} \end{pmatrix} \tag{3.3}$$

Instead of solving P_3, we solve P_4 whose feasible region is convex. However, an optimal solution of P_4 with a positive L will not be an optimal solution of P_3, unless the (positive) penalty term is zero. By making L larger, we penalize constraint violations more severely, thereby forcing the minimizer of the penalty function to be smaller. We formally prove this statement in the following proposition:

Proposition 3.1. *The value of penalty term* $\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K} r_{\hat{i}j}r_{ij}$ *at an optimal solution of Problem* P_L *decreases, as* L *increases.*

Proof. Let L_1 and L_2 be two penalty parameters so that $L_1 \leq L_2$. Denote optimal solutions of Problems P_{L_1} and P_{L_2}, with $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively. Since $\mathbf{r}_1$ is an optimal solution associated with parameter L_1, the value of the objective function of P_{L_1} at $\mathbf{r}_1$ is larger than the value of the objective function of P_{L_2} at $\mathbf{r}_2$, so

$$\mathcal{F}(\mathbf{r}_2) - \frac{L_1}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij} \leq \mathcal{F}(\mathbf{r}_1) - \frac{L_1}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} \tag{3.4}$$

and consequently

$$\frac{L_1}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij}\right) \leq \mathcal{F}(\mathbf{r}_1) - \mathcal{F}(\mathbf{r}_2) \tag{3.5}$$

Similarly, since $\mathbf{r}_2$ is an optimal solution of P_{L_2}, the value of the objective function of P_{L_2} at $\mathbf{r}_2$ is greater than its value at $\mathbf{r}_1$. Hence

$$\mathcal{F}(\mathbf{r}_1) - \frac{L_2}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} \leq \mathcal{F}(\mathbf{r}_2) - \frac{L_2}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij} \tag{3.6}$$

and consequently

$$\frac{L_2}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij}\right) \geq \mathcal{F}(\mathbf{r}_1) - \mathcal{F}(\mathbf{r}_2) \tag{3.7}$$

Inequalities (3.5) and (3.7) imply that

$$\begin{aligned}\frac{L_2}{2}&\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij}\right) \geq \mathcal{F}(\mathbf{r}_1) - \mathcal{F}(\mathbf{r}_2) \\ &\geq \frac{L_1}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij}\right)\end{aligned} \tag{3.8}$$

Hence

$$\left(\frac{L_2}{2} - \frac{L_1}{2}\right)\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij}\right) \geq 0 \tag{3.9}$$

DOI: 10.1002/wcm

Using the assumption that $L_1 \leq L_2$, we have

$$\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_1)_{\hat{i}j}(\mathbf{r}_1)_{ij} \geq \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(\mathbf{r}_2)_{\hat{i}j}(\mathbf{r}_2)_{ij} \tag{3.10}$$

which completes the proof. ■

Therefore, the larger L is, the more penalized the constraint violations of penalty term is, and the smaller the penalty term will be. Indeed, for a large enough choice of L, global optimal solution(s) of P_4 is (are) the optimal solution(s) of P_3 (Theorem 17.1 of [20]).[§]

Theorem 3.1 [20]. *Suppose that each $r^{(k)}$ is the exact global maximizer of Problem P_4 corresponding to the penalty parameter L_k. Then every limit point r^* of the sequence $\{r^{(k)}\}$ is a global solution of the problem P_3.*

However, the maximization of $f(\mathbf{r})$ in P_L becomes more difficult as L becomes large [20]. In this paper, we find an appropriate value for L through a simple search method. Even though the objective function of P_4 is a non-concave nonlinear function, but its feasible region is convex. Convexity of the feasible region motivates us to use some interior point methods to solve P_4.

Before applying the interior point method, we first convert the inequality constraints in $C(\mathbf{r})$ to equality constraints by associating a positive slack variable to each constraint. Denote the $(2M+1)K$ vector of slack variables with $\mathbf{s}$. Hence, P_4 is converted to the following minimization problem:

$$P_5: \min_{\mathbf{r}} \; -f(\mathbf{r}) \tag{3.11a}$$

$$\text{s.t} \;\; C(\mathbf{r}) - \mathbf{s} = 0 \tag{3.11b}$$

$$\mathbf{s} \geq 0 \tag{3.11c}$$

A necessary condition for a feasible solution of P_5 to be optimal is to satisfy the following conditions, called Karush–Kuhn–Tucker (KKT) conditions:

$$\nabla f(\mathbf{r}) - A^T(\mathbf{r})\mathbf{z} = 0 \tag{3.12a}$$

$$C(\mathbf{r}) - \mathbf{s} = 0 \tag{3.12b}$$

$$S\mathbf{z} = 0 \tag{3.12c}$$

$$\mathbf{s} \geq 0, \quad \mathbf{z} \geq 0 \tag{3.12d}$$

[§]Notice that the statement of Theorem 17.1 in [20] is slightly different from the one presented here, but the proof is the same.

In the aforementioned KKT conditions, S is a diagonal matrix with diagonal elements given by vector $\mathbf{s}$, and vector $\mathbf{z}$ contains $(2M+1)K$ Lagrange multipliers used in the definition of the Lagrangian function of P_5:

$$\mathcal{L}(\mathbf{r}, \mathbf{s}, \mathbf{z}) = f(\mathbf{r}) - \mathbf{z}^T(C(\mathbf{r}) - \mathbf{s}) \tag{3.13}$$

The matrix A in (3.12a) is the Jacobian matrix of $C(\mathbf{r})$ represented by:

$$A = \begin{pmatrix} & \Theta & \\ \frac{-K\ln(2)2^{\frac{Kr_{11}}{B}}}{B\alpha_{11}} & \cdots & \frac{-K\ln(2)2^{\frac{Kr_{MK}}{B}}}{B\alpha_{MK}} \\ & I & \end{pmatrix} \tag{3.14}$$

where I is an identity matrix of dimension $MK \times MK$, and Θ is the following $M \times MK$ matrix:

$$\Theta = \begin{pmatrix} 1_{(1,K)} & 0_{(1,K)} & \cdots & 0_{(1,K)} \\ 0_{(1,K)} & 1_{(1,K)} & \cdots & 0_{(1,K)} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{(1,K)} & 0_{(1,K)} & \cdots & 1_{(1,K)} \end{pmatrix} \tag{3.15}$$

where $1_{(1,K)}$ and $0_{(1,K)}$ are K vectors of 1 and 0, respectively.

To find an approximation for a local optimum of the nonlinear problem, interior point methods solve a series of perturbed KKT conditions in which only the right-hand-side in Equation (3.12c) is replaced by a vector $\mu\mathbf{e}$:

$$\nabla f(\mathbf{r}) - A^T(\mathbf{r})\mathbf{z} = 0 \tag{3.16a}$$

$$C(\mathbf{r}) - \mathbf{s} = 0 \tag{3.16b}$$

$$S\mathbf{z} = \mu\mathbf{e} \tag{3.16c}$$

$$\mathbf{s} \geq 0, \quad \mathbf{z} \geq 0 \tag{3.16d}$$

with $\mathbf{e} = (1, 1, \cdots, 1)^T$ and $\mu > 0$. Interior point methods start with an initial interior point in the feasible region that satisfies perturbed KKT conditions for some μ and proceeds to find another interior point that satisfies perturbed KKT conditions (3.16a)–(3.16d) for a smaller value of μ. As the method proceeds, μ is decreased, and consequently the solution of the perturbed KKT conditions approaches the solution of the KKT conditions, in which $\mu = 0$. It is expected that after several iterations the solution will converge to a point that satisfies the KKT conditions of the problem [20].

DOI: 10.1002/wcm

In each iteration of interior point method, the directions and lengths of movements are updated based on the first and second order gradients of the objective function and constraints. The vector of movement directions for variables $\mathbf{r}$, $\mathbf{s}$, and $\mathbf{z}$, denoted by $\mathbf{b} = [b_{\mathbf{r}}, b_{\mathbf{s}}, b_{\mathbf{z}}]^T$, is computed by solving the following linear system of equations:

$$\begin{pmatrix} \nabla^2_{\mathbf{rr}}\mathcal{L} & 0 & -A^T(\mathbf{r}) \\ 0 & Z & S \\ A(\mathbf{r}) & -I & 0 \end{pmatrix} \begin{pmatrix} b_{\mathbf{r}} \\ b_{\mathbf{s}} \\ b_{\mathbf{z}} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{r}} f(\mathbf{r}) - A^T(\mathbf{r})\mathbf{z} \\ S\mathbf{z} - \mu\mathbf{e} \\ C(\mathbf{r}) - \mathbf{s} \end{pmatrix} \tag{3.17}$$

where, Z denotes the diagonal matrix whose diagonal elements are given by vector $\mathbf{z}$. As matrices $\nabla^2_{\mathbf{rr}}\mathcal{L}$ and $\nabla_{\mathbf{r}} f(\mathbf{r})$ depend on the objective function chosen for the problem, we provide their descriptions in Appendix B for a chosen objective function.

After obtaining movement directions, the length of movement in each direction, step length, denoted by $\alpha^{\max}_{\mathbf{s}}$ and $\alpha^{\max}_{\mathbf{z}}$, are specified as below:

$$\alpha^{\max}_{\mathbf{s}} = \max\{\alpha \in (0, 1] : \mathbf{s} + \alpha b_{\mathbf{s}} \geq (1 - \tau)\,\mathbf{s}\} \tag{3.18a}$$

$$\alpha^{\max}_{\mathbf{z}} = \max\{\alpha \in (0, 1] : \mathbf{z} + \alpha b_{\mathbf{z}} \geq (1 - \tau)\,\mathbf{z}\} \tag{3.18b}$$

where $\tau \in (0, 1)$. A large value of τ close to one, e.g., $\tau = 0.995$, is usually chosen to avoid $\mathbf{s}$ and $\mathbf{z}$ approaching zero too quickly. The new interior point, slack variables, and Lagrange multipliers, $(\mathbf{r}^+, \mathbf{s}^+, \mathbf{z}^+)$, are determined with the information of movement directions and step lengths accordingly:

$$\mathbf{r}^+ = \mathbf{r} + \alpha^{\max}_{\mathbf{s}} b_{\mathbf{r}} \tag{3.19a}$$

$$\mathbf{s}^+ = \mathbf{s} + \alpha^{\max}_{\mathbf{s}} b_{\mathbf{s}} \tag{3.19b}$$

$$\mathbf{z}^+ = \mathbf{z} + \alpha^{\max}_{\mathbf{z}} b_{\mathbf{z}} \tag{3.19c}$$

For the next iteration, μ is updated to a smaller value, say $\mu^+ < \mu$. There are several strategies to choose μ^+. Among them we use a linear equation to update μ:

$$\mu^+ = \sigma\mu \quad \sigma \in (0, 1) \tag{3.20}$$

Since $\sigma < 1$, μ approaches zero over several iterations. However, choosing a very small σ or a very large σ will cause faster or slower convergence, respectively. Although fast convergence is always desired, it may cause some parameters of the method, such as $\mathbf{s}$ and $\mathbf{z}$, approaching zero too quickly. This may reduce the performance of the method, e.g., the offered solution may be infeasible or far from optimality.

The implementation of the interior point method is terminated when a stopping criterion is achieved. In this paper, the initial value of $\mu_0 = 1$ has been chosen, and when μ approaches to a very small value or the change in allocated rate vector, $\mathbf{r}$, is negligible, the implementation stops. Algorithm 1 presents a summary of the interior point method used in our simulation.

Algorithm 1 The interior point method for P_5

Input: M, K, P_{BS}, B, α, U_i, *initial_r*, $\mathbf{s}_0$, μ_0, τ, σ
Result: $\mathbf{r}$
begin
  Setting up and initialization:
  Choose *initial_r* and compute $\mathbf{s}_0 > 0$
  Choose $\mu_0 > 0$ and compute $\mathbf{z}_0 > 0$ accordingly
  Set parameters $\tau \in (0, 1)$ and $\sigma \in (0, 1)$
  Set $k = 0$ and $Exit\_flag = 0$
  while $Exit\_flag == 0$ **do**
    Solve (3.17) to obtain movement direction $\mathbf{b} = (b_{\mathbf{r}}, b_{\mathbf{s}}, b_{\mathbf{z}})$
    Compute $\alpha^{\max}_{\mathbf{s}}$, and $\alpha^{\max}_{\mathbf{z}}$ using (3.18a) and (3.18b)
    Compute $(\mathbf{r}^{k+1}, \mathbf{s}^{k+1}, \mathbf{z}^{k+1})$ using (3.19a) to (3.19c)
    Set $\mu^{k+1} \leftarrow \mu^k$ and $k \leftarrow k + 1$
    Compute $Exit\_flag$
  end
  return $\mathbf{r}$
end

4. Genetic Algorithm

4.1. Genetic Algorithm Methodology

In our simulation, we use GA as an intelligent search algorithm to find near-optimal solutions. GA is a randomized adaptive search method that processes a large number of search points at each iteration, then generates a new set of feasible points based on characteristics of the old search points. GA deploys a randomization search technique that avoids searching process being stopped when a local optimum is attained and continues searching the feasible region for a better local optimum [31]. Also, adaptive search based on the previous search points limits computational complexity, i.e., the computational burden does not necessarily increase with an increase in dimensions of search region [32].

In GA context, feasible solutions of a problem are represented by a data structure named chromosome,

DOI: 10.1002/wcm

and a fitness function is defined to evaluate feasible solutions. The algorithm begins with forming an initial population (first generation) of random feasible solutions. Then, the initial population is improved toward the optimal solution by generating a new population from the current chromosomes through several iterations. The new population is generated in each iteration through the following operators:

- **Selection:** The operator chooses better chromosomes of current generation to form a population of parent chromosomes.
- **Crossover:** The operator generates new chromosomes (children) by selecting a point on the chromosomes of the two parents and swapping the chromosomes beyond that point.
- **Mutation:** The operator probabilistically changes an arbitrary element of a chromosome to a new value hoping to find new chromosomes which may have a better fitness value.

4.2. Genetic Algorithm Implementation

The specifics of chromosomes and fitness function as well as operators implementation depend on the problem to be solved. A $K \times M$ vector is chosen for the chromosome in our implementation, where K and M are the numbers of sub-carriers and users, respectively. Chromosome y of the population is a vector $[x_1^y \cdots x_j^y \cdots x_K^y]$ of x_j^y, where $j \in \mathcal{K}$ represents a sub-carrier index, as shown in Figure 2. x_j^y is a $1 \times M$ allocation vector of a continuous value x_{ij}^y, where $i \in \mathcal{M}$ is a user's index, that shows allocated power to user i on sub-carrier j, p_{ij}^y. Each x_j^y contains only one non-zero element, x_{ij}^y, due to the constraint of exclusive sub-carrier assignment to a user.

An initial population, $\mathcal{P}_0$, of N chromosomes is formed by allocating a random user to each sub-carrier of each chromosome. The minimum required power, that satisfies user' minimum required rate, is assigned to the users that are allocated to sub-carriers in initial population. Each chromosome is a feasible solution, so it should satisfy all the constraints of the problem. If a chromosome does not satisfy the problem constraints, the procedure of chromosome generation will be repeated. The fitness function is the objective function of the optimization problem. Selection operator is a fitness proportionate selection, also known as roulette-wheel selection, that selects individuals with a probability proportional to their fitness values. This selection operator gives a chance to weak solutions (low fitness values) to be selected, hoping that those weak solutions will result in some good solutions (high fitness value) in crossover operation. Using a uniform distribution, p_{cross}, a point j from $\{M, \cdots, (K-1)M\}$ is chosen for crossover operation. In other words, crossover is performed over sub-carriers. Mutation operation chooses a mutating element from $\{1, \cdots, KM\}$ with a uniform distribution, p_{mut}. Actually, the mutating element indicates a new user i for sub-carrier j, so allocated power to the previous user of sub-carrier j is altered to zero, and a random power is allocated to the mutating element. Crossover and mutation are repeated if new generated chromosomes do not satisfy the problem constraints. Once a new population $\mathcal{P}_n$ is generated through selection and crossover and mutation, it replaces the old one. However, as the chromosome with the best fitness value, referred to as *elite*, may be lost in selection, crossover, and mutation operators, an elitism operation is performed before substituting $\mathcal{P}_{n-1}$ with $\mathcal{P}_n$. Elitism operation substitutes the corresponding chromosome to the least fitness value of $\mathcal{P}_n$ with *elite*. GA stops after $\mathcal{N}_{\text{itr}}$ iterations or when there is no increment in *elite*'s fitness value for $\mathcal{N}_{\text{fit}}$. Numerical parameters of GA are listed in Table II and the pseudo code of the solution is outlined in Algorithm 2.

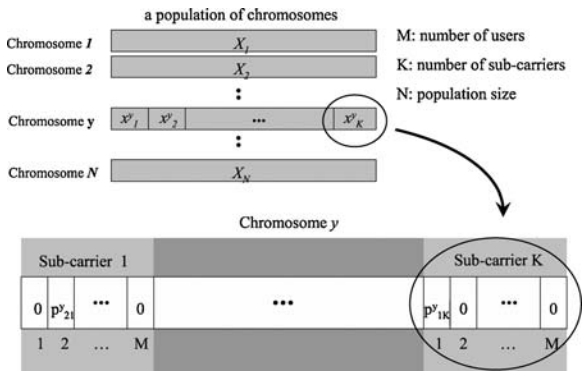

Figure 2. The population and chromosomes representations.

Table II. Simulation parameters.

Parameter	Value
Maximum power budget of the BS	20 Watt
Total bandwidth	2400 Hz
Number of sub-carriers	24
Number of users	4
Minimum required rate of users with convex utility	100 bit/symbol
Minimum required rate of users with concave utility	1 bit/symbol
Number of iterations	30000
Crossover probability	0.75
Mutation probability	0.1
Initial population	200

DOI: 10.1002/wcm

Algorithm 2 GA implementation for the problem

Input: $M, K, \mathcal{N}_{\text{itr}}, \mathcal{N}_{\text{fit}}, p_{\text{cross}}, p_{\text{mut}}, P_{\text{BS}}, B, \alpha, \mathcal{F}$
Result: p_{ij}
begin
  Setting up and initialization
  Generate initial population, $\mathcal{P}_0$
  Find $elite_0$
  $n = 1$ and $Exit\_flag = 0$
  while $Exit\_flag == 0$ **do**
    Perform *selection* using roulette wheel sampling scheme
    for $y = 1 : N$ **do**
      while *constraints* (2.13*c*) *to* (2.13*f*) *are not held* **do**
        crossover with probability p_{cross}
      end
      while *constraints* (2.13*c*) *to* (2.13*f*) *are not held* **do**
        Mutation with probability p_{mut}
      end
      Find $elite_n$
      $\mathcal{P}_{n+1} = \mathcal{P}_n$
      Replace the worst chromosome with $elite_{n-1}$
      $Exit\_flag =$ **Check_termin_conditions**
      $n = n + 1$
    end
  end
  return p_{ij}
end

5. Numerical Results

In this section, the convergence of GA is investigated in Section 5.1, which then will be used as a benchmark to evaluate the performance of PM/IPM in terms of optimality and sensitivity to network parameters in Section 5.2.

In our simulation, we consider a PMP network, where the BS is centered and users are located in different distances around it. Traffic arriving at the BS is first buffered in separate infinite queues dedicated to each user, then, is forwarded to users on the down-link path using assigned sub-carriers and allocated power. We assume the objective function is aggregate utility maximization. In its simplest form, the utility function of user i may be a linear function of its rate, $U_i = r_i$. However, for the worst case, we allow utility functions to be non-concave and nonlinear. There are two sets of users with concave and convex utility functions expressed with Equation (5.1) [33].

$$U_i(r_i) = \begin{cases} 0 & \text{if } r_i \le l_1 \\ \sin^k\left(\dfrac{\pi}{2}\dfrac{r_i - l_1}{l_2 - l_1}\right) & l_1 < r \le l_2 \\ 1 & r_i > l_2 \end{cases} \quad (5.1)$$

r_i denotes allocated rate to user i, $l1$ and $l2$ are thresholds, and k controls the shape of the utility function. The function is concave for $k < 1$ and convex for $k > 1$. $k = 0.7$ and $k = 2$ have been chosen for concave and convex utility functions, respectively. Other simulation parameters are listed in Table II.

5.1. Genetic Algorithm Convergence

To evaluate convergence performance of GA, a scenario consisting of 4 users with concave utility functions is considered. It is assumed that average channel gains are 1 and 0.3 on the first and the second half of the sub-carriers, respectively, for all users. In the first iteration of GA, sub-carriers are assigned to users exclusively and randomly; This assignment of sub-carriers is irrespective of users' channel gain on sub-carriers. Then, the required power to achieve a minimum rate requirement of each user is allocated uniformly to the sub-carriers assigned to each user. It is expected that more power is allocated to the sub-carriers with better average channel gain as iterations proceed, to gain higher rate and utility. Figure 3 depicts the distribution of allocated power to the sub-carriers in the first and the last iteration of GA. A comparison between the two distributions illustrates that GA evolves toward allocating more power to the good status sub-carriers and less power to the bad status (weak) sub-carriers, i.e., evolution of the algorithm toward maximizing the objective function by utilizing the resource efficiently. To show the speed of convergence, the best fitness value, the best users' total utility of a chromosomes, in each iteration is illustrated in Figure 4. The curve is monotonically increasing due to elitism technique, i.e., the best individual of current population is transfered to the next population, so the best fitness value never drops. As expected, there is a noticeable trade off between optimality and short solution time.

5.2. PM/IPM Performance

We evaluate the performance of PM/IPM in terms of optimality, solution time, and sensitivity of solution to users' channel gain variations on sub-carriers. The results achieved by GA is used as a benchmark. To

DOI: 10.1002/wcm

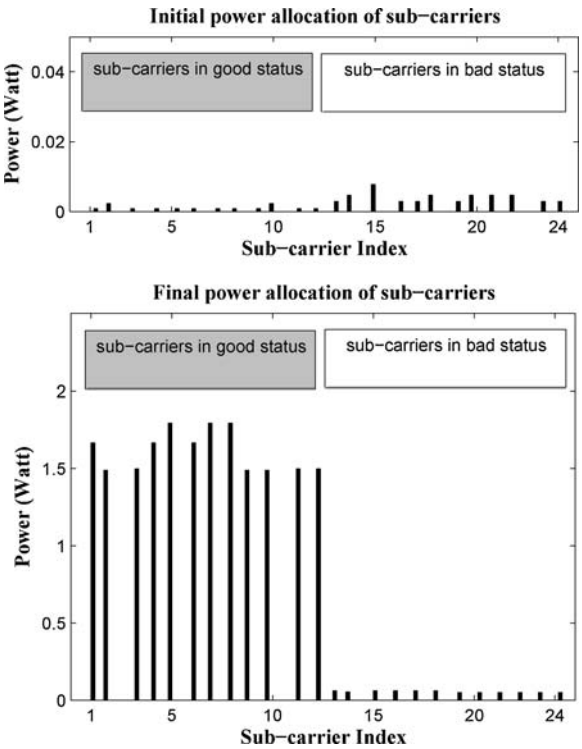

Figure 3. Power allocation distribution on sub-carriers.

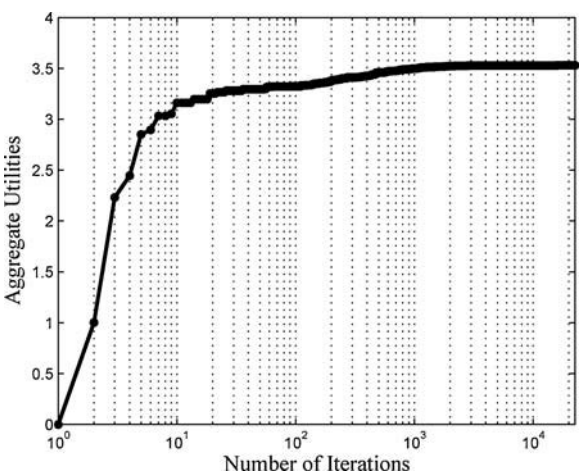

Figure 4. Convergence of fitness value.

increase the convergence time of GA we consider a small number of users. Four users with the same convex utility functions but diverse channel gain on sub-carriers are considered. Average channel gain on sub-carriers is higher for users 1 and 3 than users 2 and 4.

A comparison between the convergence speed of GA and PM/IPM is shown in Figure 5. The iterations of GA

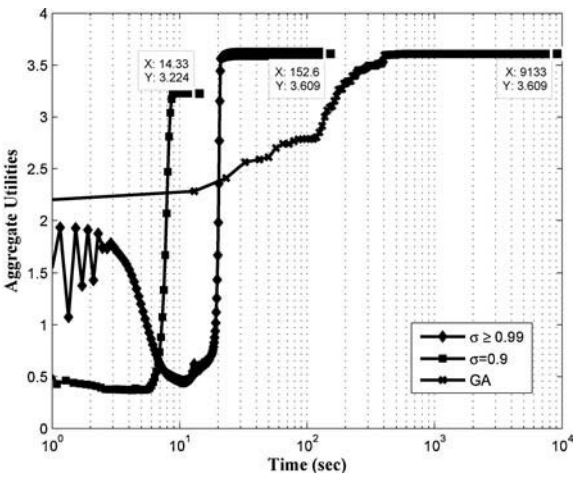

Figure 5. Convergence speed comparison of GA and PM/IPM.

and PM/IPM stop when the improvement in rate allocation vector is less than $1e-13$. GA has a very slow convergence speed, although it starts from an initial allocation with better aggregate utilities than the ones of PM/IPM. In comparison, PM/IPM converges very fast while its maximum achievable aggregate utilities and convergence time depend on the value of σ. The method is faster, but the results are less accurate for smaller values of σ. The data tips on the diagram show the time and aggregate utilities at the data tips x and y, respectively. It can be seen that PM/IPM is much faster than GA, and with $\sigma = 0.99$, PM/IPM obtains the same aggregate utilities as the one that GA obtains in its convergence. When σ increases beyond 0.99, PM/IPM has no more improvement in achievable aggregate utilities or convergence speed.

The convergence of PM/IPM is determined by the aggregate utilities and constraints' violations in the penalty term. For PM/IPM convergence, aggregate utilities should be maximized subject to the fact that constraints' violations are negligible or close to zero. Figure 6 illustrates aggregate constraints' deviations (from zero), for two different values of σ, when PM/IPM iterations proceed over time. The negligible aggregate deviations at convergence points, especially for $\sigma = 0.99$, ensures the rate allocation satisfies the exclusive sub-carrier allocation. Moreover, a comparison between Figures 5 and 6 show aggregate constraints' deviations and aggregate utilities convergence happen simultaneously, which satisfies the convergence requirements of the problem.

Moreover, a comparison between rate allocation of GA and PM/IPM, shown in Figure 7, demonstrates the

DOI: 10.1002/wcm

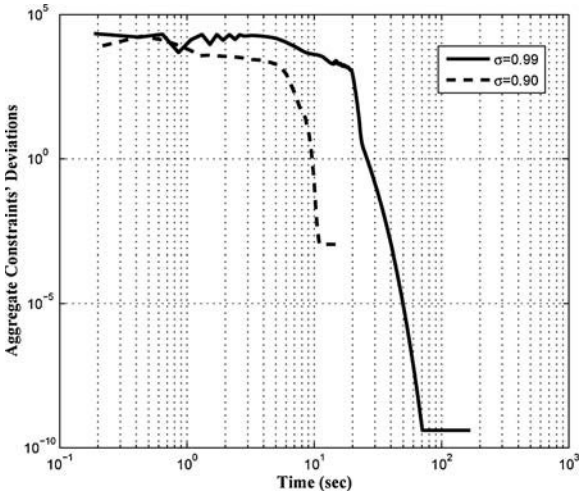

Figure 6. Aggregate penalty term constraints' deviations in PM/IPM.

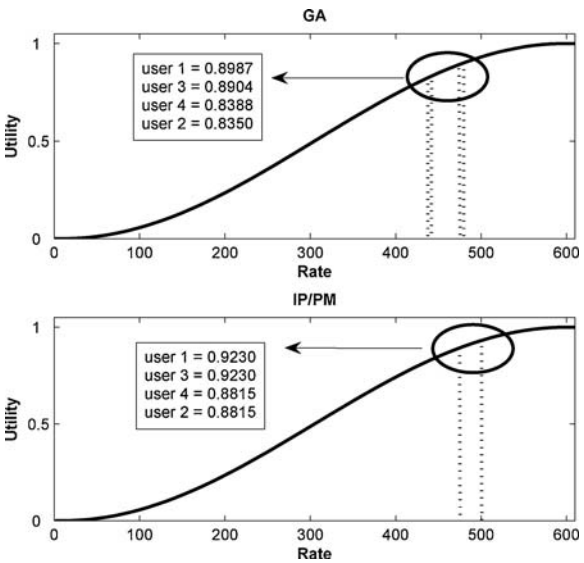

Figure 7. Utility allocation comparison of GA and PM/IPM.

performance of PM/IPM in recognizing diverse channel status and its capability in allocating resource. Let all users have the same channel status, except that average channel gain on sub-carriers is higher for users 1 and 3 than those for users 2 and 4. Therefore, more resource should be allocate to the users with better average channel quality to gain user diversity and maximize aggregate utilities. The numeric tables (data tips) in Figure 7 indicate that both GA and PM/IPM allocate more rate to users 1 and 3 than users 2 and 4. Also, it can be seen that PM/IPM allocates equal rate to the users with the same average channel quality on sub-carriers.

Table III. Users's allocated rates on each sub-carrier.

n	α_1	r_1	α_2	r_2	α_3	r_3	α_4	r_4
1	0.50	37	0.10	0	0.02	0	0.30	0
2	1.30	0	1.04	0	0.59	0	0.40	0
3	0.11	0	1.04	0	6.13	470	0.75	0
4	0.11	0	0.41	0	0.27	0	2.13	221
5	0.29	0	1.97	0	3.48	0	0.98	0
6	3.34	0	1.97	0	1.04	0	1.52	0
7	0.49	0	0.79	0	0.44	0	4.20	0
8	0.52	0	2.25	309	0.83	0	0.02	0
9	1.94	0	1.99	290	0.43	0	0.06	0
10	1.03	0	0.25	0	0.99	0	0.21	0
11	1.33	0	0.20	0	2.22	0	0.95	0
12	1.27	0	0.44	0	1.10	0	0.82	0
13	0.01	0	0.10	0	0.34	0	0.85	43
14	0.62	0	2.89	0	1.58	0	0.30	0
15	1.16	0	3.69	0	1.66	0	2.51	0
16	0.28	0	0.93	0	7.21	128	1.41	0
17	0.47	0	0.42	0	0.37	0	3.16	0
18	3.79	0	0.59	0	1.03	0	1.75	0
19	3.24	0	0.37	0	0.06	0	4.34	334
20	2.37	260	0.06	0	0.42	0	0.65	0
21	1.98	0	2.68	0	2.11	0	0.40	0
22	0.31	0	0.33	0	0.34	0	1.34	0
23	0.83	0	1.18	0	0.33	0	0.46	0
24	3.63	301	0.15	0	1.27	0	2.21	0

Table III presents rate allocation and exclusive sub-carrier assignment by PM/IPM, the vectors of allocated rate to sub-carriers, $n = 1, \cdots, 24$, for users 1 to 4, r_1 to r_4, along with the corresponding channel gains of the users on the sub-carriers, α_1 to α_4. The gray rows of the table represent the assigned sub-carriers to the users, and the sub-carriers on white rows are unassigned. The result confirms the success of PM/IPM in exclusive sub-carrier assignment since no sub-carrier has been assigned to two users. In addition, a sub-carrier is assigned to a user that has the best channel gain on that sub-carrier, which results in a solution closer to the optimum. In numerical results given in Table III, all users achieve a utility equal to one, so some sub-carriers are not needed to be assigned to any user.

6. Conclusions

The non-convexity of OFDMA resource allocation optimization problem has been studied in this paper. A framework for the resource allocation has been developed and a novel approach based on a penalty function method and an interior point method (PM/IPM) has been applied to solve the optimization problem. Numerical results have demonstrated that the proposed approach performs well in achieving near optimal

DOI: 10.1002/wcm

solutions while satisfies the non-convex (sub-carrier assignment) constraints.

In our future work, we aim to extend the proposed approach for resource allocation in networks with alternative topologies, traffic requirements, and channel models.

Appendixes

Appendix A

In the following, we show how an optimization problem with a discrete feasible region is non-convex. Consider a network consisting of two users and three sub-carriers. If allocated power to user i on sub-carrier j, p_{ij} is nonzero, then allocated power to the other users on sub-carrier j should be zero. For example, two power allocation vectors $\mathbf{p} = [1, 0, 0, 0, 1, 1]$ and $\hat{\mathbf{p}} = [0, 1, 1, 1, 0, 0]$ are feasible power allocation vectors. The first three elements of the vectors represent sub-carrier allocation to the first user, and the last three elements indicate sub-carrier allocation to the second user. For $\alpha \in (0, 1)$, the convex combination of $\mathbf{p}$ and $\hat{\mathbf{p}}$:

$$\alpha\mathbf{p} + (1-\alpha)\hat{\mathbf{p}} = [\alpha, (1-\alpha), (1-\alpha), (1-\alpha), \alpha, \alpha] \tag{A.1}$$

does not belong to the feasible region, and the definition of convex feasible region is not held.

Appendix B

The mathematical representations of $\nabla^2_{\mathbf{rr}}\mathcal{L}$ and $\nabla_{\mathbf{r}} f(\mathbf{r})$, required by the interior point method are presented. The objective function of P_5, based on utility functions (5.1), is represented by:

$$f(\mathbf{r}) = -(U_1(r_1) + \cdots + U_M(r_M)) + \frac{L}{2}\left(\sum_i \sum_{\hat{i}} (r_{\hat{i}1} r_{i1} + \cdots + r_{\hat{i}K} r_{iK})\right) \tag{B.1}$$

Accordingly, $\nabla_r f(\mathbf{r}) = (\frac{\partial f}{\partial r_{11}}, \cdots, \frac{\partial f}{\partial r_{MK}})^T$ is computed as follows:

$$\nabla_r f(\mathbf{r}) = -\begin{pmatrix} \frac{\partial U_1(r_1)}{\partial r_{11}} \\ \vdots \\ \frac{\partial U_1(r_1)}{\partial r_{MK}} \\ \vdots \\ \frac{\partial U_M(r_M)}{\partial r_{M1}} \\ \vdots \\ \frac{\partial U_M(r_M)}{\partial r_{MK}} \end{pmatrix} + L \begin{pmatrix} \sum_i r_{i1} - r_{11} \\ \vdots \\ \sum_i r_{iK} - r_{MK} \\ \vdots \\ \sum_i r_{i1} - r_{M1} \\ \vdots \\ \sum_i r_{iK} - r_{MK} \end{pmatrix} \tag{B.2}$$

where, for $j = 1, \cdots, K$, and $\theta = \frac{\pi}{2}\frac{\mathbf{r}-l_1}{l_2-l_1}$:

$$\frac{\partial U_{\check{i}}}{\partial r_{ij}} = \begin{cases} \dfrac{k\pi}{2(l_2-l_1)} \sin^{(k-1)}(\theta)\cos(\theta) & \text{if } i = \check{i} \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

To obtain $\nabla^2_{\mathbf{rr}}\mathcal{L}$, $\nabla^2_{\mathbf{rr}} f(\mathbf{r})$ and $\nabla^2_{\mathbf{rr}} C(\mathbf{r})$ are computed first:

$$\nabla^2_{\mathbf{rr}} f(\mathbf{r}) = -\begin{pmatrix} G(r_1) & 0_{(K,K)} & \cdots & 0_{(K,K)} \\ 0_{(K,K)} & G(r_2) & \cdots & 0_{(K,K)} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{(K,K)} & 0_{(K,K)} & \cdots & G(r_M) \end{pmatrix} + L\begin{pmatrix} 0_{(K,K)} & I_{(K,K)} & \cdots & I_{(K,K)} \\ I_{(K,K)} & 0_{(K,K)} & \cdots & I_{(K,K)} \\ \vdots & \vdots & \ddots & \vdots \\ I_{(K,K)} & I_{(K,K)} & \cdots & 0_{(K,K)} \end{pmatrix} \tag{B.4}$$

where

$$G(r_i) = \begin{pmatrix} \frac{\partial^2 U_i}{\partial r_{i1}\partial r_{i1}} & \cdots & \frac{\partial^2 U_i}{\partial r_{i1}\partial r_{iK}} \\ \frac{\partial^2 U_i}{\partial r_{i2}\partial r_{i1}} & \cdots & \frac{\partial^2 U_i}{\partial r_{i2}\partial r_{iK}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 U_i}{\partial r_{iK}\partial r_{i1}} & \cdots & \frac{\partial^2 U_i}{\partial r_{iK}\partial r_{iK}} \end{pmatrix} \tag{B.5}$$

DOI: 10.1002/wcm

$0_{(K,K)}$ is a $K \times K$ matrix with all zero entries, and $I_{(K,K)}$ is a $K \times K$ identity matrix. The second partial derivatives of the utility functions required for calculating $G(r_i)$ functions are:

$$\frac{\partial^2 U_i}{\partial r_{i\breve{j}} \partial r_{ij}} = \frac{K\pi^2}{4(l_2 - l_1)^2} \left((k-1)\sin^{(k-2)}(\theta)\cos^2(\theta) - \sin^k(\theta) \right) \tag{B.6}$$

for $\breve{j}$ and $j \in \{1, \cdots, K\}$. Finally, $\nabla^2_{\mathbf{rr}} C(\mathbf{r})$ for calculating $\nabla^2_{\mathbf{rr}} \mathcal{L}$ is obtained by:

$$\nabla^2_{\mathbf{rr}} C(\mathbf{r}) = \left(\frac{Kln(2)}{B} \right)^2 \times \begin{pmatrix} \frac{2^{\frac{Kr_{11}}{B}}}{\alpha_{11}} & 0 & \cdots & 0 \\ 0 & \frac{2^{\frac{Kr_{12}}{B}}}{\alpha_{12}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2^{\frac{Kr_{MK}}{B}}}{\alpha_{MK}} \end{pmatrix} \tag{B.7}$$

References

1. Song G, Li Y. Cross-layer optimization for OFDM wireless networks-part I: theoretical framework. *IEEE Transactions Wireless Communications* 2005; **4**(2): 614–624.
2. Song G, Li Y. Cross-layer optimization for OFDM wireless networks-part II: algorithm development. *IEEE Transactions Wireless Communications* 2005; **4**(2): 625–634.
3. Vavasis SA. *Nonlinear Optimization: Complexity Issues*. Oxford University Press, Inc.: New York, NY, USA, 1991.
4. Agrawal R, Berry R, Huang J, Subramanian V. Optimal scheduling for OFDMA systems. In *Proceedings of the 40th Annual Asilomar Conference on Signals, Systems, and Computers (invited paper)*, 2006; 1347–1351.
5. Rhee W, Cioffi J. Increase in capacity of multiuser OFDM system using dynamic sub-channel allocation. In *Proceedings IEEE, the 51st Vehicular Technology Conference (VTC2000-Spring)*, 2000; **2**: 1085–1089.
6. Wang W-H, Palaniswami M, Low SH. Application-oriented flow control: fundamentals, algorithms and fairness. *IEEE/ACM Transactions Networking* 2006; **14**(6): 1282–1291.
7. Lu N, Bigham J. On utility-fair bandwidth adaptation for multiclass traffic QoS provisioning in wireless networks. *Computer Network* 2007; **51**(10): 2554–2564.
8. Harks T. Utility proportional fair bandwidth allocation: an optimization oriented approach. In *Proceedings of 3rd International Workshop on QoS in Multiservice IP Networks (QoS-IP)*, 2005; **3375**: 61–74.
9. Mohanram C, Bhashyam S. Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM. *IEEE Transactions Wireless Communications* 2007; **6**(9): 3208–3213.
10. Chen Y, Chen J. A Fast subcarrier, bit, and power allocation algorithm for multiuser OFDM-based systems. *IEEE Transactions on Vehicular Technology* 2008; **57**(2): 873–881.
11. Letaief K, Zhang Y. Dynamic multiuser resource allocation and adaptation for wireless systems. *IEEE Wireless Communications Magazine*, 2006; **13**(4): 38–47.
12. Ergen M, Coleri S, Varaiya P. QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems. *IEEE Transactions Broadcast.* 2003; **49**(4): 362–370.
13. Jang J, Lee K. Transmit power adaptation for multiuser OFDM systems. *IEEE Journal Selected Areas Communications* 2003; **21**(2): 171–178.
14. Zhang Y, Letaief K. Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems. *IEEE Transactions Wireless Communications* 2004; **3**(5): 1566–1575.
15. Shen Z. Multiuser resource allocation in multichannel wireless communication systems. *Ph.D. Dissertation*. University of Texas: Austin, 2006.
16. Cai J, Shen X, Mark J. Downlink resource management for packet transmission in OFDM wireless communication systems. *IEEE Transactions Wireless Communications* 2005; **4**(4): 1688–1703.
17. Seong K, Mohseni M, Cioffi JM. Optimal resource allocation for OFDMA downlink systems. In *Proceedings IEEE International Symposium on Information Theory (ISIT)*, 2006; 1394–1398.
18. Yu W, Lui R. Dual Methods for nonconvex spectrum optimization of multicarrier systems. *IEEE Transactions Communications* 2006; **54**(7): 1310–1322.
19. Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. *Society for Industrial and Applied Mathematics (SIAM)* 2002; **44**(4): 525—597.
20. Nocedal J, Wright S. *Numerical Optimization*. Springer: New York, USA, 2006.
21. Shanno DF, Vanderbei RJ. Interior-point methods for non-convex nonlinear programming: orderings and higher-order methods. *Mathmatical Programming* 2000; **87**: 303–316.
22. Mehrjoo M, Moazeni S, Shen X. A new modeling approach for utility-based resource allocation in OFDM networks. In *Proceedings IEEE International Conference on Communications (ICC)*, 2008; 337–342.
23. Greenhalgh D, Marshall S. Convergence Criteria for Genetic Algorithms. *Society for Industrial and Applied Mathematics (SIAM), Journal on Computing* 2000; **30**(1): 269–282.
24. Mark JW, Zhuang W. *Wireless Communications and Networking*. Pearson Education International: Upper Saddle River, New Jersey, USA, 2003.
25. Rudin W. *Principles of Mathematical Analysis*. McGraw-Hill Publishing Co.: New York, 1976.
26. Cover TM, Thomas JA. *Elements of Information Theory*. John Wiley & Sons: New York, 1991.
27. Biglieri E, Proakis J, Shamai S. Fading Channels: information-theoretic and communications aspects. *IEEE Transactions on Information Theory* 1998; **44**(6): 2619–2692.
28. Mehrjoo M, Dianati M, Shen X, Naik K. Opportunistic fair scheduling for the downlink of IEEE 802.16 wireless metropolitan area networks. In *Proceedings of 3rd International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine)*, 2006; **191**: 52–60.
29. Dianati M, Shen X, Naik K. Cooperative fair scheduling for the downlink of CDMA cellular networks. *IEEE Transactions on Vehicular Technology* 2007; **56**(4): 1749–1760.

DOI: 10.1002/wcm

30. Awad MK, Shen X. OFDMA Based two-hop cooperative relay network resources allocation. In *Proceedings IEEE International Conference on Communications (ICC)*, Vol. 2, 2008.
31. Whitley D. A genetic algorithm tutorial. *Statistics and Computing*. 1994; **4**(2): 65–85.
32. Marrison C, Stengel R. The use of random search and genetic algorithms to optimize stochastic robustness functions. In *Proceedings of American Control Conference*, Vol. 2, 1994.
33. Chiu DM, Tam AS-W. Fairness of traffic controls for inelastic flows in the internet. *Computer Network*. 2007; **51**(11): 2938–2957.

Authors' Biographies

Mehri Mehrjoo (S'06) received the B.A.Sc. and the M.A.Sc. degrees from Ferdowsi University, Mashad, Iran, and Ph.D. from the University of Waterloo, Waterloo, Canada in 1993, 1996 and 2008, respectively. She is presently a post-doctoral fellow at the University of Waterloo. From 1996 to 2003, she was a lecturer with the Department of Electrical Engineering, University of Sistan and Baloochestan, Zahedan, Iran. Her research interests are in the areas of resource allocation, performance analysis of wireless protocols, vehicular communication.

Somayeh Moazeni received the B.Sc. and the M.Sc. degree in Mathematics from Amirkabir University of Technology, Tehran, Iran in February 2005 and March 2005, respectively, and the MMATH degree in Combinatorial Optimization from the Department of Combinatorics and Optimization in the University of Waterloo, Ontario, Canada in 2006. She is currently a Ph.D. candidate at the School of Computer Science in the University of Waterloo. Her research interests include mathematical programming, financial optimization and optimization under uncertainty. She is a member of the Canadian Operational Research Society.

Xuemin (Sherman) Shen received the B.Sc. (1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. Dr Shen's research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications networks, wireless network security, wireless body area networks and vehicular ad hoc and sensor networks. He is a co-author of three books, and has published more than 400 papers and book chapters in wireless communications and networks, control and filtering. Dr Shen served as the Tutorial Chair for IEEE ICC 2008, the Technical Program Committee Chair for IEEE Globecom 2007, the General Co-Chair for Chinacom 2007 and QShine 2006, the Founding Chair for IEEE Communications Society Technical Committee on P2P Communications and Networking. He also serves as a Founding Area Editor for IEEE Transactions on Wireless Communications; Editor-in-Chief for Peer-to-Peer Networking and Application; Associate Editor for IEEE Transactions on Vehicular Technology; KICS/IEEE Journal of Communications and Networks, *Computer Networks*; ACM/Wireless Networks; and Wireless Communications and Mobile Computing (Wiley), etc. He has also served as Guest Editor for IEEE JSAC, IEEE Wireless Communications, IEEE Communications Magazine, and ACM Mobile Networks and Applications, etc. Dr Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004 and 2008 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. Dr Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, and a Distinguished Lecturer of IEEE Communications Society.

DOI: 10.1002/wcm