Title
Downlink User Selection and Resource Allocation for Semi-Elastic Flows in an OFDM Cell

Permalink
https://escholarship.org/uc/item/7qs07830

Journal
Wireless Networks, 19(6)

Authors
Yang, Chao
Jordan, Scott

Publication Date
2013-08-01

DOI
10.1007/s11276-013-0541-9

Peer reviewed

Wireless Netw
DOI 10.1007/s11276-013-0541-9

Downlink user selection and resource allocation for semi-elastic flows in an OFDM cell

Chao Yang · Scott Jordan

Abstract We are concerned with user selection and resource allocation in wireless networks for semi-elastic applications such as video conferencing. While many packet scheduling algorithms have been proposed for elastic applications, and many user selection algorithms have been proposed for inelastic applications, little is known about optimal user selection and resource allocation for semi-elastic applications in wireless networks. We consider user selection and allocation of downlink transmission power and subcarriers in an orthogonal frequency division multiplexing cellular system. We pose a utility maximization problem, but find that direct solution is computationally intractable. We first propose a method that makes joint decisions about user selection and resource allocation by transforming the utility function into a concave function so that convex optimization techniques can be used, resulting in a complexity polynomial in the number of users with a bounded duality gap. This method can be implemented if the network communicates a shadow price for power to power allocation modules, which in turn communicate shadow prices for rate to individual users. We then propose a method that makes separate decisions about user selection and resource allocation, resulting in a complexity linear in the number of users.

Keywords Communication system traffic control · Cellular networks

The preliminary version was presented at the 2011 IEEE Global Communications Conference (GLOBECOM 2011)

C. Yang · S. Jordan (✉)
University of California Irvine, Irvine, CA, USA
e-mail: sjordan@uci.edu

C. Yang
e-mail: Chao.Yang@uci.edu

1 Introduction

Use of video applications on cellular networks has mushroomed in recent years. It is now estimated that video comprises one third of downstream North American mobile Internet access peak period traffic [1]. Most of this video traffic is streaming encoded using either Adobe Flash or MPEG. Some of this video traffic is video conferencing, e.g. Apple's FaceTime for iPhones and Skype's video conferencing application for smartphones. All of the video conferencing traffic and a sizeable portion of the video streaming traffic is *semi-elastic*, meaning that these applications can cope with significant but limited variation in throughput over short time periods.

Semi-elastic applications pose considerable challenges to resource allocation in cellular networks, which have largely been designed to support inelastic applications, e.g. constant bit rate voice, and elastic applications, e.g. email and web-browsing. Resource allocation in wireless networks is particularly sensitive to variation, since wireless networks not only experience variation in demand but also variation in capacity due to fluctuations in the wireless channel.

Resource allocation for *inelastic* applications has been well studied. Inelastic applications can not withstand much variation in short-term throughput over the duration of the connection, since they typically require that a very high percentage of packets be received within a fraction of a second. The classic approach to wireless resource allocation for inelastic applications focuses on a decision of which users should be active and thus consume wireless resources. The most common approaches are to minimize the usage of wireless resources (often power or channels) given a set of active users. This is often referred as *margin adaption* [2]. Another common approach is to allocate resources based on the impact that these resources have

Published online: 18 January 2013

upon application *performance*. Typically resources such as carriers and power are mapped into rate, which is then mapped into application *utility*. Since inelastic applications can not withstand much short-term variation, utility usually is modeled as rising quickly as performance reaches an acceptable level. For instance, utility is often modeled as a step function of rate, which reflects the requirement to achieve a constant bit rate if this user is active. Resource allocation usually attempts to maximize total user utility. This again focuses on a decision of which users should be active, and commonly the optimization requires some type of bin-packing algorithm, see e.g. [3].

Resource allocation for *elastic* applications has also been well studied. Elastic applications can withstand a great deal of variation in short-term throughput over the duration of the connection, since they are typically not very interactive. Many of these applications measure performance by completion time not by short-term throughput. The classic approach to wireless resource allocation for elastic applications chooses to make all users active and thus focuses on the wireless resources each consumes. The most common approaches are to maximize capacity, often measured by total user throughput, and/or to minimize the usage of wireless resources (often power or channels) given a set of active users, see e.g. [4–7]. As with inelastic applications, an alternate approach is to represent an application's satisfaction with its performance using a utility function. Again, typically resources such as carriers and power are mapped into rate, which is then mapped into utility. Since elastic applications can withstand much short-term variation, utility usually is modeled as an increasing concave function of rate. Resource allocation again usually attempts to maximize total user utility. As a consequence of the concavity, however, convex optimization techniques can often be used for elastic applications, and these techniques can be used to design packet scheduling algorithms, see e.g. [8–12].

In contrast, resource allocation in cellular networks for semi-elastic applications has not been well studied. A few papers have proposed modeling semi-elastic applications using sigmoid utility function, which are convex at rates less than a threshold and concave at rates above that threshold. This shape is thought to reflect the nature of the compression techniques used in semi-elastic applications, which are designed to adjust to fluctuations provided that short-term throughput remains above a threshold, but which do not fail gracefully when short-term throughput falls below that threshold. The convex portion of a sigmoid utility function implies that resource allocation algorithms for semi-elastic applications must decide which users should be active, similar to those for inelastic applications. However, the concave portion of a sigmoid utility function implies that resource allocation algorithms for semi-elastic applications must also decide which wireless resources should be allocated to each active user, similar to those for elastic applications. Lee et. al. [13] consider semi-elastic applications in Code Division Multiple Access (CDMA) systems. They propose first using pricing to select which users should be active. They then use pricing again to allocate power to these users. Hande et. al. [14] consider semi-elastic applications in wireline systems. They ignore user selection. They propose using pricing to allocate wireline capacity, and give conditions under which the Nash equilibrium using marginal cost pricing maximizes total user utility. Cheung et. al. [15] consider wireless local area networks with mixed elastic and semi-elastic applications. They also use price based algorithms to allocate rate. Abbas et. al. [16] propose a general framework, which is also based on a pricing algorithm, to allocate bandwidth for elastic and semi-elastic applications in the Internet. Jiong et. al. [17] analyze bandwidth allocation problems for inelastic and semi-elastic applications in wireless sensor networks.

We are concerned here with downlink resource allocation for semi-elastic applications in orthogonal frequency division multiplexing (OFDM) cellular systems. In an OFDM system, the broadband wireless channel is divided into a set of orthogonal narrowband subcarriers, each of which can be allocated to an individual user. The wireless resources are thus comprised of both base station transmission power and subcarriers. This makes the resource allocation problem moderately more complex than the CDMA system and other wireless systems considered in [13, 15, 17] or the wireline system considered in [14, 16], which consider allocation of a single type of resource (power, rate, or bandwidth).

A reasonable conjecture is that it may be simple to combine utility maximization based approaches to resource allocation for inelastic applications and elastic applications and arrive at a reasonable approach for semi-elastic applications. However, as we discuss in detail below, such an approach would require the combination of some type of bin-packing algorithm with convex optimization techniques, resulting in a mixed integer program that is computationally prohibitive to solve.

In this paper, we propose two methods with moderate complexity to allocate resources for semi-elastic applications. The first method jointly makes decisions about which users should be active and what downlink power and subcarriers to allocate to each such active user. The basic idea is to solve the dual problem and we show how active user selection, power allocation, and subcarrier allocation can be efficiently accomplished using shadow cost pricing. The computational complexity of the resulting algorithm is polynomial in the number of users. Not surprisingly, the resulting active user selection is suboptimal; however, we present a bound on the corresponding duality gap.

The second proposed method separates the decisions about which users should be active and what downlink power and subcarriers to allocate to each such active user. The first stage focuses on active user selection, and uses a greedy algorithm that attempts to ensure that every active user will obtain a rate at least equal to the rate at the maximum average utility. The second stage takes the active user set as given and uses standard convex optimization techniques to allocate power and subcarriers. Separating the decisions reduces the computational complexity so that it is now linear in the number of users. Of course, this separation comes at the cost of a decrease in total user utility, but we show in numerical examples that this decrease is small except when the base station transmission power is low enough that user blocking is high.

The rest of this paper is organized as follows. In Sect. 2, we pose the problem and find that direct solution requires solving a large set of fixed point equations. We turn to a dual formulation in Sect. 3, and show that the complexity can be reduced with a small loss in efficiency by distributing the resource allocation process amongst users, the network, and intermediate power allocation modules. The network allocates power and subcarriers, and charges the power allocation module a shadow price for power. The power allocation module translates this price per unit power into a price per unit rate, and resells the system resources to users. Users choose desired rates based on the cost and the resulting utility. We pose an iterative algorithm that determines near-optimal shadow prices. The resulting algorithms are polynomial in the number of users but less complex than direct solution of the primal problem. In Sect. 4, we derive properties of the resulting allocation and a bound on the sub-optimality. We then propose an algorithm in Sect. 5 with a complexity linear in the number of users, by decomposing subcarrier allocation and rate scheduling. Finally in Sect. 6 we show via simulation that the performance of this latter algorithm is very close to that of the dual iterative algorithm, with the principal loss emanating from the simplified subcarrier allocation.

2 System model and problem formulation

We focus on the downlink of a single OFDM cell serving K users, with N subcarriers. The bandwidth of each subcarrier is B which is assumed to be less than the coherence bandwidth of the channel so that the channel response can be considered flat [18] .[1] The rate of user k on subcarrier n in time slot t is:

$$r_{k,n,t}(p_{k,n,t}) = B\log_2\left(1 + p_{k,n,t}\frac{H_{k,n,t}^2}{\delta^2 + I}\right) \tag{1}$$

where $p_{k,n,t}$ is the power allocated, $H_{k,n,t}$ is the composite channel fading which includes both the small scale fading and pathloss, I is the interference power and δ^2 is the noise power. The channel fading is assumed known at the base station. The total rate of user k is:

$$R_{k,t} = \sum_{n=1}^{N} w_{k,n,t} r_{k,n,t} \tag{2}$$

where $w_{k,n,t}$ is an indicator of subcarrier assignment, i.e. $w_{k,n,t} = 1$ if subcarrier n is allocated to user k in time slot t and $w_{k,n,t} = 0$ otherwise.

The natural optimization metric, as used in many previous papers, is total user utility. The utility of user k in time slot t is assumed to be a function $U_k(R_{k,t})$ which maps the rate $R_{k,t}$ to the level of satisfaction perceived by the application. From Eqs. (1) and (2), user k's utility can be represented as:

$$U_k\left(\sum_{n=1}^{N} w_{k,n} B\log_2\left(1 + p_{k,n,t}\cdot\frac{H_{k,n}^2}{\delta^2 + I}\right)\right) \tag{3}$$

In this paper, as with many in the literature, we focus on a snapshot of the system, and thus for the remainder of the paper, we drop the time t subscript on all variables for simplicity of presentation. Consideration of the impact of the variation of resource allocation over time remains a topic for future research. We expect that there the results presented here can be used to guide development of connection access control, e.g. a measurement-based method may admit users on the basis of an admitted user's expected utility, expected rate, and/or in what percentage of slots an admitted user can be expected to be active.

Denote the subcarrier allocation by a vector $\mathbf{w} = \{w_{k,n}\}$, and the power allocations by a vector $\mathbf{p} = \{p_{k,n}\}$. Maximization of total user utility can be represented as:

$$\begin{aligned} U_{tot}^* &= \max_{\mathbf{w},\mathbf{p}} \sum_{k=1}^{K} U_k(R_k) \\ \text{s.t.} &\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} \le P_T; \quad p_{k,n} \ge 0\,\forall k,n \\ &\sum_{k=1}^{K} w_{k,n} \le 1\,\forall n; \quad w_{k,n} = 0 \text{ or } 1\,\forall k,n \end{aligned} \tag{4}$$

where P_T is the total downlink power available to the base station. User k's rate R_k, given its allocated subcarriers and power, can be found using (2). We call user k *inactive* in the current time slot if $R_k = 0$, i.e. if the user has not been allocated power and subcarriers.

[1] A major advantage of OFDM is that each subcarrier can be considered as flat fading. Aspects of frequency selective fading are typically addressed at the physical layer.

For elastic applications, it is often assumed that U_k is an increasing concave function of the rate R_k. If the maximization was purely over power allocation, e.g. as in CDMA systems, then the constraint set is convex, and thus the optimization is a convex problem. Packet scheduling algorithms have been proposed for elastic applications in CDMA systems based on standard convex optimization techniques, see e.g. [9, 11].

In OFDM systems, however, the integer constraints on $\mathbf{w}$ are required to ensure that each subcarrier can be allocated to at most one user. These constraints make this optimization problem a mixed-integer program, which is computationally intensive to solve. Some papers, see e.g. [6], have proposed eliminating the integer constraints by allowing subcarriers to be split between users, i.e. by replacing $w_{k,n} = 0$ or $1\ \forall\ k, n$ with $0 \preceq \mathbf{w} \preceq 1$. For elastic applications, the resulting optimization is then a convex problem, and packet scheduling algorithms have been proposed for elastic applications in OFDM systems based on standard convex optimization techniques, see e.g. [10, 12].

For inelastic applications, in contrast, it is often assumed that U_k is a step function of the rate R_k, i.e. utility is zero if the rate is below a threshold and a positive constant if the rate is above the threshold. As a result, convex optimization techniques can not be directly applied in either CDMA or OFDM systems. In CDMA systems, the step function implies only one efficient choice for power allocation if a user is active. In this case, the base station need only identify the active user set (and corresponding transmission powers). The general approach to such cases is bin-packing. However, the complexity of bin-packing algorithms is high, and thus some papers have proposed identifying users that should be active using pricing of power. The resulting algorithms often select users that have high surplus, defined as utility minus the cost of the allocated resources, see e.g. [3, 19]. It is likely that a similar approach could be used in OFDM systems.

For semi-elastic applications, however, utility is assumed to be a sigmoid function as pictured in Fig. 1, namely there exists an inflection point R_k^f such that U_k is convex for $R_k < R_k^f$ and concave for $R_k > R_k^f$. We denote the tangent point rate at the maximum average utility by $R_k^{'}$, namely $R_k^{\prime} = \arg\max_{R_k} U_k/R_k$ (Our notation is summarized in Table 1.)

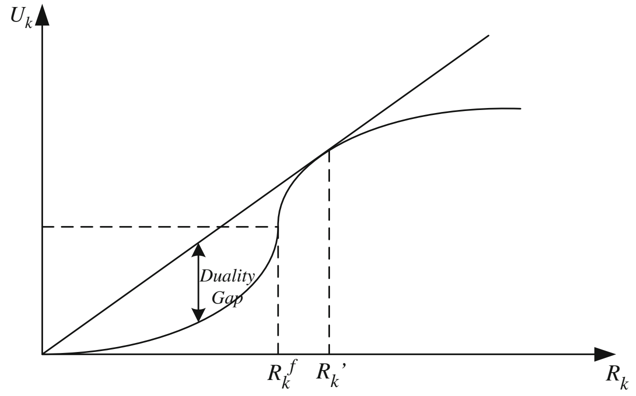


Fig. 1 Sigmoid utility function

Table 1 Notation

Notation	Description
K	number of users
N	number of subcarriers
$H_{k,n}$	composite channel fading of user k on subcarrier n
$p_{k,n}$	allocated power to user k on subcarrier n
$r_{k,n}$	rate of user k on subcarrier n
$w_{k,n}$	indicator of allocation of subcarrier n to user k
R_k	rate of user k
B	subcarrier bandwidth
P_T	downlink power of base station
I and δ^2	interference power and noise power
U_k	utility function of user k
$R_k^{'}$	rate of user k at maximum average utility

Semi-elastic applications in OFDM systems thus present two difficulties: non-concavity of the utility function and integer constraints on subcarrier allocation. To eliminate the integer constraints, one could follow the lead of other papers and allow subcarriers to be split between users: replace $w_{k,n} = 0$ or $1\ \forall\ \ k, n$ by $0 \preceq \mathbf{w} \preceq 1$, define $s_{k,n} = w_{k,n}p_{k,n}$ as the power allocated to user k on subcarrier n, and change the power constraint to $\sum_{k=1}^{K}\sum_{n=1}^{N} s_{k,n} \leq P_T$. The constraint set thus becomes convex. The new objective function is

$$\sum_{k=1}^{K} U_k\left(\sum_{n=1}^{N} w_{k,n} B \log_2\left(1 + \frac{s_{k,n}}{w_{k,n}} \cdot \frac{H_{k,n}^2}{\delta^2 + I}\right)\right) \tag{5}$$

which is neither convex nor concave.

The Lagrange function is:

$$\begin{aligned} J(\mathbf{w}, \mathbf{s}, \mu, \boldsymbol{v}) &= \sum_{k=1}^{K} U_k\left(\sum_{n=1}^{N} w_{k,n} B \log_2\left(1 + \frac{s_{k,n}}{w_{k,n}} \cdot \frac{H_{k,n}^2}{\delta^2 + I}\right)\right) \\ &+ \mu\left(P_T - \sum_{k=1}^{K}\sum_{n=1}^{N} s_{k,n}\right) + \sum_{n=1}^{N} v_n\left(1 - \sum_{k=1}^{K} w_{k,n}\right) \end{aligned}$$

where μ and $\boldsymbol{v}$ are Lagrange multipliers for the power and subcarrier constraints respectively.

If utility was concave, then convex optimization techniques could be directly applied. Since utility is sigmoid, they can not. It can be easily shown that the corresponding Karush-Kuhn-Tucker necessary conditions lead to:

$$p_{k,n} = \frac{s_{k,n}}{w_{k,n}} = \left(\frac{BU_k'(R_k)}{\mu \ln 2} - \frac{\delta^2 + I}{H_{k,n}^2}\right)^+ \tag{6}$$

$$\frac{\partial J}{\partial w_{k,n}} = BU_k'(R_k)\left[\log_2\left(1 + \frac{s_{k,n}}{w_{k,n}}\frac{H_{k,n}^2}{\delta^2 + I}\right) - \frac{1}{\ln 2}\frac{s_{k,n}/w_{k,n}}{\frac{s_{k,n}}{w_{k,n}} + \frac{\delta^2+I}{H_{k,n}^2}}\right] - v_n \tag{7}$$

where $(x)^+ \triangleq \max(0, x)$ and $U_k'(R_k) = U_k'\left(\sum_{n=1}^{N} w_{k,n} B \log_2\left(1 + \frac{s_{k,n}}{w_{k,n}} \cdot \frac{H_{k,n}^2}{\delta^2+I}\right)\right)$.

Subcarrier allocation is based on (7), which in turn requires solving (6); however, direct solution of (6) requires solving KN nonlinear fixed point equations. This is simpler than solving the mixed integer programming problem but nevertheless computationally difficult. Moreover, since these equations are not sufficient conditions for optimality, there may be multiple solutions which must be compared to identify the optimal solution.

Alternately, one may try to apply the same techniques used for inelastic applications. However, for semi-elastic applications there are an infinite number of efficient choices for power allocation if a user is active. Thus bin-packing algorithms can not be used.

We are thus motivated to look for a less computationally complex but possibly sub-optimal solution.

3 Solution and algorithm based on dual decomposition

3.1 Dual formulation

Due to the computational complexity of solving even the relaxed version of problem (4), in the remainder of the paper we propose two methods with moderate complexity to allocate resources for semi-elastic applications. In this section, we present a method that jointly makes decisions about which users should be active and what downlink power and subcarriers to allocate to each such active user. We propose to use a dual formulation, and identification of an active user set, and allocation of downlink transmission power and subcarriers to active users, can be accomplished with a small loss in efficiency using an iterative search for optimal shadow prices.

The idea, used previously for strictly concave utility functions [20], is to decompose the allocation of power and subcarriers and the determination of user rate R_k using an intermediate variable d_k as a lower bound on the achieved rate R_k. Using this decomposition, problem (4) can be rewritten as:

$$\max_{\mathbf{w},\mathbf{p},\mathbf{d}} \sum_{k=1}^{K} U_k(d_k) \tag{8}$$

$$\text{s.t.} R_k \geq d_k;\ \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} \leq P_T;\ p_{k,n} \geq 0\, \forall k, n$$

$$\sum_{k=1}^{K} w_{k,n} \leq 1\, \forall n;\ w_{k,n} = 0 \text{ or } 1\, \forall k, n$$

where $\mathbf{d}$ represents $\{d_k\}$ and R_k is calculated using (2). User k is thus active if and only if $d_k > 0$. The new problem (8) must have the same solution as the original problem (4), since U_k is an increasing function of d_k, and thus at the optimum $d_k = R_k$.

A standard approach to reduce computational complexity is to search for the optimal Lagrange multipliers and to let them determine the optimal resource allocation rather than to directly search for the optimal power and subcarrier allocations. This can be done by posing a dual problem, see e.g. [21]. Each subcarrier can be allocated to at most one user; this requirement can be simply expressed by defining a set $\mathbf{A} = \{\mathbf{p}$ s.t. $\forall\, n,\ p_{k,n} > 0$ for at most one user k $\}$, and thus $w_{k,n} = 1$ if and only if $p_{k,n} > 0$. Within this set, there is always a feasible solution to problem (8), since d_k can be set to an arbitrarily small value. This allows us to incorporate the subcarrier assignments $\mathbf{w}$ into the power allocations $\mathbf{p}$. The Lagrange function of (8) is given by:

$$J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \mu) = \sum_{k=1}^{K} U_k(d_k) + \mu\left(P_T - \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}\right) + \sum_{k=1}^{K} \lambda_k (R_k - d_k) \tag{9}$$

where R_k is now given by $R_k = \sum_{n=1}^{N} r_{k,n}$ with $r_{k,n}$ determined by (1), $\boldsymbol{\lambda}$ are the Lagrange multipliers for the rate constraints, and μ is the Lagrange multiplier for the power constraint. The dual function is then given by:

$$\bar{J}(\boldsymbol{\lambda}, \mu) = \max_{\mathbf{p} \in \mathbf{A}, \mathbf{d}} J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \mu)$$

and the dual problem is to optimally choose the Lagrange multipliers:

$$\bar{J}^* = \min_{\boldsymbol{\lambda}, \mu} \bar{J}(\boldsymbol{\lambda}, \mu) \text{ s.t. } \boldsymbol{\lambda} \succeq 0, \mu \geq 0 \tag{10}$$

The dual function can be decomposed into two pieces, $\bar{J}(\boldsymbol{\lambda}, \mu) = f_1(\boldsymbol{\lambda}) + f_2(\boldsymbol{\lambda}, \mu)$, where:

$$f_1(\boldsymbol{\lambda}) = \max_{\mathbf{d}} \sum_{k=1}^{K} (U_k(d_k) - \lambda_k d_k) \tag{11}$$

$$f_2(\boldsymbol{\lambda}, \mu) = \max_{\mathbf{p} \in \mathbf{A}} \left[\sum_{k=1}^{K} \lambda_k R_k + \mu\left(P_T - \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}\right)\right] \tag{12}$$

The first piece of the dual function, $f_1(\boldsymbol{\lambda})$, can be used to determine the set of active users. If utility were strictly concave, demand for rate would be a continuous and decreasing function of the Lagrange multiplier λ_k. For sigmoid utility, however, the solution for (11) is not continuous. Define $\overline{\lambda}_k = dU_k(R_k)/dR_k|(R_k = R_k^{'})$ as the slope of the utility curve at the tangent point rate. When $\lambda_k < \overline{\lambda}_k$, user k should be active and should be allocated a rate $d_k > R_k^{'}$ that is a continuous and decreasing function of λ_k. When $\lambda_k = \overline{\lambda}_k$, however, (11) produces a tie between $d_k = 0$ and $d_k = R_k^{'}$; we break the tie using $d_k = 0$ if $R_k = 0$ and $d_k = R_k^{'}$ otherwise. When $\lambda_k > \overline{\lambda}_k$, user k should be inactive.

The second piece of the dual function, $f_2(\boldsymbol{\lambda}, \mu)$, can be used to determine the power allocation. It can be further decomposed into N independent problems:

$$f_2(\boldsymbol{\lambda}, \mu) = \sum_{n=1}^{N} f_{2,n}(\boldsymbol{\lambda}, \mu) + \mu P_T$$

where

$$f_{2,n}(\boldsymbol{\lambda}, \mu) = \max_{\mathbf{p} \in \mathbf{A}} \left(\sum_{k=1}^{K} \lambda_k r_{k,n} - \mu \sum_{k=1}^{K} p_{k,n} \right) \tag{13}$$

Thus the dual problem (10) can be represented as:

$$\overline{J}^* = \min_{\lambda,\mu} \left[f_1(\boldsymbol{\lambda}) + \sum_{n=1}^{N} f_{2,n}(\boldsymbol{\lambda}, \mu) + \mu P_T \right] \tag{14}$$

According to the first order condition $\partial f_{2,n}/\partial p_{k,n} = 0$, the solution to the maximization in (13) is

$$p_{k,n} = \left(\frac{B\lambda_k}{\mu \ln 2} - \frac{\delta^2 + I}{H_{k,n}^2} \right)^+ \tag{15}$$

Substituting (15) into (13) and simplifying we obtain

$$f_{2,n}(\boldsymbol{\lambda}, \mu) = \max_k \Phi_{k,n} \tag{16}$$

where

$$\Phi_{k,n} = \lambda_k B \left[\log_2 \left(\frac{B\lambda_k}{\mu \ln 2} \frac{H_{k,n}^2}{\delta^2 + I} \right) \right]^+ - \mu \left(\frac{B\lambda_k}{\mu \ln 2} - \frac{\delta^2 + I}{H_{k,n}^2} \right)^+$$

Use of the dual problem removes the need for a mixed integer program by more elegantly satisfying the integer constraints. Rather than treating subcarriers as a separate allocation from power, subcarrier assignments are incorporated into power allocations using the set **A**. The restriction that a subcarrier can not be allocated to more than one user is elegantly implemented in (16).

Use of the dual problem allows standard convex optimization techniques to be used to create a resource allocation. However, the solution to the dual problem may not be the same as the solution to the primal problem. Comparing the conditions for the solution to the dual problem, Eqs. (15) and (16), with the conditions for optimality of the primal problem, Eqs. (6) and (7), we find that the only difference is that $U_k^{'}(R_k)$ in the primal conditions has been replaced by λ_k in the dual conditions.

If utility was concave, $U_k^{'}(R_k) = \lambda_k$, and thus the primal and dual solutions are identical. However, for sigmoid utility, $U_k^{'}(R_k) = \lambda_k$ only when $R_k > R_k^{'}$, and thus the primal and dual solutions are not identical if the primal solution allocates a rate less than $R_k^{'}$ to any user, as illustrated in Fig. 1. In this situation, the dual problem generates a solution in which, following (11), such users will choose a rate d_k equal either to zero or $R_k^{'}$. In the literature on non-convex optimization, this discrepancy is referred to as a *duality gap* [21]. In Sect. 4, we will show that the duality gap is bounded.

3.2 Algorithm development

The decomposition also indicates a method of distributing the optimization. The benefits of distributed optimization are that the base station doesn't need to know the exact utility function of each user and this will reduce signalling between the base station and users. The determination of the desired rates **d** in (11) indicates a role for each user, while the determination of Lagrange multiplier μ in (14) indicates a role for the network. These two roles must be done in coordination. The decomposition suggests to us that there should be an intermediate power allocation module which determines the Lagrange multipliers $\boldsymbol{\lambda}$ in (14) and determines the powers **p** in (15). The communication between the users, power allocation module, and network is illustrated in Fig. 2.

Each of these three roles has a local optimization to accomplish. These can be done iteratively as follows, where the iteration number is denoted by a superscript i.

(1) **User** k **Algorithm**: Given λ_k^i,

$$d_k^{i+1} = \arg\max_{d_k} [U_k(d_k) - \lambda_k^i d_k]$$

(2) **Network Algorithm**: Given tentative power and subcarrier allocations $\mathbf{p} \in \mathbf{A}$,

$$\mu^{i+1} = [\mu^i + s_B^i z_B^i]^+$$

where s_B^i is a positive scalar stepsize, $[\cdot]^+$ is the projection on $\mathbb{R}_+$, and $z_B^i = \mathrm{sgn}(\sum_{k=1}^{K} \sum_{n=1}^{N} p_{k,n}^i - P_T)$.

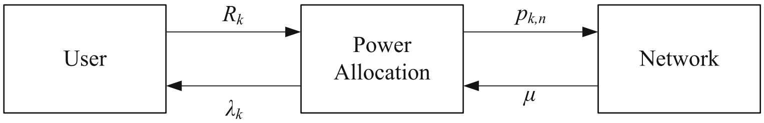


Fig. 2 Communication between user, power allocation module, and network

(3) **Power Allocation Algorithm**: Given target rates $\mathbf{d}^i$ and Lagrange multiplier μ^i, allocate $\mathbf{p}$ using (15) and (16) and update $\boldsymbol{\lambda}^{i+1}$ as follows

$$\boldsymbol{\lambda}^{i+1} = [\boldsymbol{\lambda}^i + s_P^i \mathbf{z}_P^i]^+$$

where s_P^i is a positive scalar stepsize and $\mathbf{z}_P^i$ is any feasible direction that satisfies $\mathrm{sgn}(\lambda_k^{i+1} - \lambda_k^i) = \mathrm{sgn}(d_k^i - R_k^i)\ \forall\ k$.

This set of algorithms has an economic interpretation. The Lagrange multipliers $\boldsymbol{\lambda}$ can be interpreted as shadow costs for rate. If user k is charged a price λ_k per unit rate, then (11) states that the system should allocate rate so as to maximize total user surplus, defined as total user utility minus total user cost. The user k algorithm implements this local optimization for user k. Similarly the Lagrange multiplier μ can be interpreted as a shadow cost for power. The network algorithm iteratively adjusts each μ by raising it if the demand exceeds the supply, and lowering it if the supply exceeds the demand.[2]

The job of the power allocation module is to purchase power at a price μ per unit power, and to resell it in the form of rate to individual users. The power allocation algorithm purchases power using (15) and (16), and iteratively adjusts each price λ_k by lowering it if the resulting average rate exceeds the user's purchased rate, and raising it if the purchased rate exceeds the average rate.[3] These two actions can be interpreted as an attempt by the power allocation module to maximize profit, defined as revenue from users minus cost for power, as illustrated in (13).

3.3 Algorithm convergence and optimality

The solution to a dual formulation is not in general guaranteed to converge, and if it does converge it is not guaranteed to converge to the allocation that is optimal for the primary problem. We address these issues in this subsection, first focussing on the optimality question.

In non-cooperative settings, a Nash equilibrium can be defined as follows. Let $x = (x_1, \ldots, x_n)$ with $x_i \in S_i$ denote the set of strategies of all players, called a strategy profile vector, and let $S = S_1 \times S_2 \ldots \times S_n$ denote the possible set of such strategies. Let $f = (f_1(x), \ldots, f_n(x))$ denote the resulting payoff to each player. Denote by x_{-i} a strategy profile of all players except for player i.

Definition 1 [22] A strategy profile $x^* \in S$ is a Nash equilibrium of a game (S, f) if no unilateral deviation in strategy by any single player is profitable for that player, that is

$$\forall i, x_i \in S_i, x_i \neq x_i^* : f_i(x_i^*, x_{-i}^*) > f_i(x_i, x_{-i}^*)$$

Our setting forces users to cooperate, in an attempt to maximize total user utility, by charging them prices. Nevertheless, we can view each user, plus the network and the power allocation, as a player. The Nash equilibrium of these players formally defines the equilibrium of the user, network, and power allocation algorithms. If the user, network, and power allocation algorithms converge, then they converge to the Nash equilibrium.

In general, the Nash equilibrium only guarantees that no user can increase utility by unilaterally changing its target rate. It does not necessarily guarantee that total utility is maximized.

The duality gap between the dual and primal problems is $\overline{J}^* - U_{tot}^*$. For strictly concave utility functions, the duality gap is always 0, and the dual problem always gives the same solution as the primal problem; hence, development of resource allocation algorithms based on the dual problem is relatively straightforward. However, for semi-elastic flows utility is not concave. Our first lemma establishes when, under sigmoid utility, the dual problem gives the same solution as the primal problem:

Lemma 1 *If $\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} = P_T$ and $R_k = d_k\ \forall\ k$, then $\overline{J}^* = U_{tot}^*$.*

Proof *Under the hypotheses,* (9) *gives* $J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \mu) = \sum_{k=1}^{K} U_k(R_k)$, *and thus* (10) *gives* $\overline{J}^* = U_{tot}^*$.

The next theorem addresses when the Nash equilibrium of the user, network, and power allocation algorithms results in the optimal solution.

Denote the rates of the dual problem (10) at the Nash equilibrium as $\{\overline{R}_k^*\}$. Since from (11) user k will never select a rate $0 < d_k < R_k^{'}$, it follows from lemma 1 that the duality gap is 0 if $\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} = P_T$ and $\overline{R}_k^* = 0$ or $\overline{R}_k^* > R_k^{'}\,\forall k$. As a consequence, the Nash equilibrium maximizes total user utility:

Theorem 1 *If there exists a Nash equilibrium for the user, network and power allocation algorithms, then the duality gap is 0 and the Nash equilibrium is the optimal solution of the primary problem* (4).

Proof At the Nash equilibrium $\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} = P_T$ and $\overline{R}_k^* = 0$ or $\overline{R}_k^* > R_k^{'}\,\forall k$, and thus the duality gap is 0. $J(\mathbf{d}, \mathbf{p}, \boldsymbol{\lambda}, \mu) = \sum_{k=1}^{K} U_k(R_k)$ and the solution of dual problem is the optimal solution of the primary problem (4). □

We now turn to the issue of convergence. The challenge is that the user, network, and power allocation algorithms may not converge to a Nash equilibrium. This typically would occur when the solution to the primal problem

[2] Many optimization methods may be used; below we propose a bisection method.

[3] Many optimization methods may be used; below we propose a subgradient method.

includes at least one active user with a rate R_k below $R_k^{'}$. In this case, the duality gap is greater than 0, and the solution to the dual problem does not satisfy the power constraints. As a result, the algorithms may oscillate and not jointly converge to an equilibrium point. The algorithms must thus be modified to guarantee convergence.

One way is to force the algorithms to terminate by placing limits on the shadow costs. For the power allocation algorithm, we propose a subgradient method with bounds to update λ:

$$\lambda_k^{i+1} = \max[\min(\lambda_k^i + s_P^i(d_k^i - R_k^i), \overline{\lambda}_k), \underline{\lambda}] \tag{17}$$

with a suitable choice of step size. The lower bound $\underline{\lambda}$ can be set to the slope of the utility function at the rate that would be achieved if a single user were allocated all system resources. In practice, the lower bound may be set by the base station at a slightly higher level to decrease convergence time, but the influence of lower bound on convergence time is very slight. The duality gap will not be influenced by the lower bound. The upper bound can be set to $\overline{\lambda}_k = dU_k(R_k)/dR_k|(R_k = R_k^{'})$. If the utility function is not known by the base station, then these bounds may require that some minimal information if transmitted from the users to the base station. From (11), we know if $d_k > 0$ the user should be active. Thus, the active user set and resource allocation can be decided together.

If the problem were convex, then a diminishing step size can guarantee convergence, see e.g. [23]. However, for non-convex problems, convergence is not guaranteed; indeed, Hande et. al. [14] show that similar algorithms may fluctuate around the optimal point if a user's allocated rate is near the tangent point. Thus, here we force the iteration for $\boldsymbol{\lambda}$ to terminate when:

$$|\lambda_k^{i+1} - \lambda_k^i| < \delta \, \forall k \text{ or } R_k^{i+1} = R_k^i \, \forall i \tag{18}$$

where δ is a small constant.

For the update of μ in the network algorithm, we propose a bisection algorithm:

$$\begin{aligned} &\text{If } z_B^i > 0, \text{then } \mu^{i+1} = (\mu^i + \overline{\mu}^i)/2, \underline{\mu}^{i+1} = \mu^i, \overline{\mu}^{i+1} = \overline{\mu}^i \\ &\text{else } \mu^{i+1} = (\mu^i + \underline{\mu}^i)/2, \underline{\mu}^{i+1} = \underline{\mu}^i, \overline{\mu}^{i+1} = \mu^i \end{aligned} \tag{19}$$

where the initial lower bound $\underline{\mu}^0$ can be set to a small suitable constant and the initial upper bound $\overline{\mu}^0$ can be derived from (17) as $\overline{\mu}^0 = \max_{k,n} \, (B\overline{\lambda}_k H_{k,n}^2)/[(\delta^2 + I)\ln 2]$. The iteration for μ is terminated when:

$$|\mu^{i+1} - \mu^i| < \epsilon \tag{20}$$

where ϵ is a small constant.

We call the resulting algorithm, outlined in Table 2, Dual Iteration Search (DIS). Our algorithm is based on multi-layer decomposition. We first determine the optimal price per unit power μ, and then determine the optimal prices per unit rate $\boldsymbol{\lambda}$.

Table 2 Dual iteration search

```
Initialize $\mu^0 = \underline{\mu}^0$, $\lambda_k^0 = \underline{\lambda} \, \forall k$
Repeat
    Repeat
      Allocate subcarrier by (16)
      Allocate power by (15)
      Update μ using (19)
    Until (20)
    If $\lambda_k^{i+1} = \overline{\lambda}_k$
      If $R_k^{i+1} = 0$ then $d_k^{i+1} = 0$ Else $d_k^{i+1} = R_k^{'}$
    Else calculate $d_k^{i+1}$ by (11)
    Update $\boldsymbol{\lambda}$ using (17)
Until (18)
```

If the Nash equilibrium of the user, network, and power allocation algorithms without the bounds and stopping rules introduced in (20), (17), and (18) exists, then the previous theorem states that the result is optimal. If such a Nash equilibrium does not exist, then the bounds and stopping rules guarantee termination in finite time, but result is likely not optimal. The suboptimality is caused by a duality gap greater than 0, which occurs when the optimal solution to the primal problem includes at least one active user with a rate R_k below $R_k^{'}$. In Sect. 4, this gap will be analyzed in detail. In this situation, the solution to the dual problem likely allocates such users a rate equal to $R_k^{'}$. However, this causes the power constraints of the primal problem to be violated. The DIS algorithm will respond to this constraint violation by raising the price per unit power μ until the constraint is satisfied. As a result, the DIS algorithm will allocate such users a rate below $R_k^{'}$, but will also allocate other users slightly lower rates than the solution to the dual problem. The result of the DIS algorithm thus approximately represents the dual solution projected into the primal constraint set.

The complexity of subgradient updates is polynomial in the dimension of the dual problem, and thus the complexity of the DIS algorithm is polynomial in the number of users K.

4 Properties of the solution to the dual problem

To understand the results more thoroughly, in this section we derive properties of the solution to the dual problem. These properties concern power and subcarrier allocation, the probability of outage, and the duality gap. The first two properties characterize the power and subcarrier allocation of the optimal solution to the dual problem.

Property 1 Power and subcarrier allocations are a function of the ratio of the optimal shadow costs, $\{\lambda_k/\ \mu\}$ and of user channels $\{H_{k,n}\}$.

Proof Power and subcarrier allocation are determined by maximization of $\Phi_{k,n}$, given in (18). Because the shadow cost for power, μ is the same for all users, dividing $\Phi_{k,n}$ by μ gives:

$$\frac{\Phi_{k,n}}{\mu} = \frac{\lambda_k}{\mu} B \log_2\left(\frac{B\lambda_k}{\mu \ln 2}\frac{H_{k,n}^2}{\delta^2 + I}\right)^+ - \left(\frac{B\lambda_k}{\mu \ln 2} - \frac{\delta^2 + I}{H_{k,n}^2}\right)^+ \tag{21}$$

As a result, subcarrier allocation is a function of $\{\lambda_k/\mu\}$ and of $\{H_{k,n}\}$. Power allocation, as given by (15), is also a function of $\{\lambda_k/\mu\}$ and of $\{H_{k,n}\}$. □

The ratio of the optimal shadow costs, $\{\lambda_k/\mu\}$, has units power per unit rate, and can be thought of as the efficiency of power use.

Property 2 (a) If all users have the same composite channel fading $H_{k,n}^2$ on subcarrier n, the subcarrier will be allocated to the user(s) with the highest shadow cost for rate, λ_k. (b) If all users have the same shadow cost for rate, λ_k, subcarrier n will be allocated to the user(s) with the highest composite channel fading $H_{k,n}^2$ on subcarrier n. (c) If all users have the same product of shadow cost for rate and composite channel fading on subcarrier n, $\lambda_k H_{k,n}^2$, the subcarrier will be allocated to the user(s) with the worst channel.

Proof

(a) The derivative of $\Phi_{k,n}$ to λ_k is

$$\frac{\partial \Phi_{k,n}}{\partial \lambda_k} = B \log_2\left(\frac{B\lambda_k}{\mu \ln 2}\frac{H_{k,n}^2}{\delta^2 + I}\right)^+ > 0$$

$\Phi_{k,n}$ is an increasing function of λ_k. Thus under the same channel fading condition, the user(s) with the highest λ_k will be assigned subcarrier n.

(b) The derivative of $\Phi_{k,n}$ to $H_{k,n}^2$ is

$$\frac{\partial \Phi_{k,n}}{\partial H_{k,n}^2} = \frac{2\mu}{H_{k,n}^2}\left(\frac{B\lambda_k}{\mu \ln 2} - \frac{\delta^2 + I}{H_{k,n}^2}\right) > 0$$

$\Phi_{k,n}$ is an increasing function of $H_{k,n}^2$. Hence if all the users have the same λ_k, the user(s) with the highest $H_{k,n}^2$ will be assigned subcarrier n.

(c) We further rearrange equation (24) as follows:

$$\frac{\Phi_{k,n}}{\mu} = \frac{1}{H_{k,n}^2}\left[\frac{B\lambda_k H_{k,n}^2}{\mu}\log_2\left(\frac{B\lambda_k}{\mu \ln 2}\frac{H_{k,n}^2}{\delta^2 + I}\right)^+ - \left(\frac{B\lambda_k H_{k,n}^2}{\mu \ln 2} - (\delta^2 + I)\right)^+\right]$$

The user(s) with the lowest $H_{k,n}^2$ will have the highest $\Phi_{k,n}$ and thus be assigned subcarrier n.

The next property concerns the probability of outage, $Pr(R_k = 0)$. Suppose the composite channel fading $H_{k,n}^2$ consists of a deterministic pathloss PL_k and random small scale fading $\alpha_{k,n}^2$, i.e. $H_{k,n}^2 = \alpha_{k,n}^2/\ PL_k$. The probability of outage can be related to the maximum average utility per unit rate, $\overline{\lambda}_k = dU_k(R_k)/dR_k|(R_k = R_k^{'})$:

Property 3 $Pr(R_k > 0)$ is an increasing function of $\overline{\lambda}_k$.

Proof Denote the relationship between $\Phi_{k,n}$ and $\alpha_{k,n}^2$, given in (21), as $\Phi_{k,n}(\alpha_{k,n}^2)$. We would like to define the inverse of this function, $\Phi_{\hat{k},n}^{-1}\left(\Phi_{k,n}(\alpha_{k,n}^2)\right)$. When $\Phi_{k,n} > 0$, $\Phi_{k,n}$ is a monotonically increasing function of $\alpha_{k,n}^2$, and thus $\Phi_{k,n}$ is a one-to-one function in this domain. However, there may be multiple values of $\alpha_{k,n}^2$ for which $\Phi_{k,n}(\alpha_{k,n}^2) = 0$. However, $\Phi_{k,n} = 0$ if and only if $p_{k,n} = 0$, and using (15) it follows that $\Phi_{k,n} = 0$ if and only if $\alpha_{k,n} \le PL_k \mu \ln 2(\delta^2 + I)/(B\lambda_k)$. Thus if we define $\Phi_{k,n}^{-1}(0) = PL_k \mu \ln 2(\delta^2 + I)/(B\lambda_k)$, then $\Phi_{k,n}(\alpha_{k,n}^2)$ is a one-to-one function and thus the inverse $\Phi_{\hat{k},n}^{-1}\left(\Phi_{k,n}(\alpha_{k,n}^2)\right)$ exists.

In the dual problem, the event $R_k = 0$ occurs if and only if $\lambda_k = \overline{\lambda}_k$. Thus:

$$Pr(R_k > 0) = N \int\limits_{\Phi_{k,n}^{-1}(0)}^{+\infty} \prod_{\hat{k}=1, \hat{k}\neq k}^{K} \int\limits_{\underline{\lambda}/\overline{\mu}}^{+\infty} \int\limits_{0}^{\Phi_{\hat{k},n}^{-1}\left(\Phi_{k,n}(\alpha_{k,n}^2)\right)} f(\alpha_{\hat{k},n}^2) d\alpha_{\hat{k},n}^2 \, f(\lambda_{\hat{k}}/\mu) d(\lambda_{\hat{k}}/\mu) f(\alpha_{k,n}^2) d\alpha_{k,n}^2 \tag{22}$$

Since $\Phi_{\hat{k},n}^{-1}\left(\Phi_{k,n}(\alpha_{k,n}^2)\right)$ is an increasing function of $\overline{\lambda}_k$ and $\Phi_{k,n}^{-1}(0)$ is a decreasing function of $\overline{\lambda}_k$, it follows that $Pr(R_k > 0)$ is an increasing function of $\overline{\lambda}_k$.

If the shadow cost per unit rate is low enough, a user will be active and will likely achieve a rate above $R_k^{'}$, the rate at the maximum average utility. The probability of outage thus depends critically on the maximum average utility per unit rate, $\overline{\lambda}_k$.

The final property provides a bound on the duality gap. Let R_k^l denote the rate in the convex portion of user k's utility curve at which the marginal utility equals $\overline{\lambda}_k$, i.e. $0 < R_k^l < R_k^f$ such that $dU_k(R_k)/dR_k|(R_k = R_k^l) = \overline{\lambda}_k$. The maximum duality gap for a particular user occurs when that user achieves rate R_k^l. It can be shown that the duality gap can be bounded by the sum of the differences between $\overline{\lambda}_k R_k^l$ and $U_k(R_k^l)$.

Property 4 The duality gap satisfies $\overline{J}^* - U^*_{tot} \leq \sum_{k=1}^{K} \left(\overline{\lambda}_k R_k^l - U_k(R_k^l)\right)$.

Proof Use $()^*$ to denote the optimal solution of the original packet scheduling problem (4) and $()^{DIS}$ to denote the result of the DIS algorithm. It is straightforward that $J^{DIS} \geq \overline{J}^* \geq U^*_{tot} \geq \sum_{k=1}^{K} U_k\left(R_k^{DIS}\right) = U^{DIS}$. It follows that the duality gap between the dual and primal problems is $\overline{J}^* - U^*_{tot} \leq J^{DIS} - U^{DIS}$. We can thus establish a bound on the duality gap if we can bound $J^{DIS} - U^{DIS}$.

The Lagrange function of the dual problem is given in (9). The difference between the allocated power and the available power approaches zero as the termination condition ϵ for the DIS algorithm approaches zero. It follows that:

$$J^{DIS} = \sum_{k=1}^{K} U_k(d_k) + \sum_{k=1}^{K} \lambda_k (R_k^{DIS} - d_k)$$

After the DIS algorithm terminates, denote by Ω_1 the set of users who achieved their desired rate, i.e. $\Omega_1 = \{k | R_k^{DIS} = d_k\}$, and denote by Ω_2 the set of users who did not achieve their desired rate, i.e. $\Omega_2 = \{k | R_k^{DIS} < d_k\}$. For users in Ω_2, $\lambda_k = \overline{\lambda}_k$, $d_k = R_k^{'}$ and $U_k(d_k) = \overline{\lambda}_k R_k^{'}$. It follows that

$$\begin{aligned} J^{DIS} &= \sum_{k \in \Omega_1} U_k(d_k) + \sum_{k \in \Omega_2} U_k(d_k) + \sum_{k \in \Omega_1} \lambda_k (R_k^{DIS} - d_k) \\ &\quad + \sum_{k \in \Omega_2} \overline{\lambda}_k (R_k^{DIS} - R_k^{'}) \\ &= \sum_{k \in \Omega_1} U_k(R_k^{DIS}) + \sum_{k \in \Omega_2} \overline{\lambda}_k R_k^{DIS} \end{aligned} \tag{23}$$

Similarly:

$$U^{DIS} = \sum_{k=1}^{K} U_k(R_k^{DIS}) = \sum_{k \in \Omega_1} U_k(R_k^{DIS}) + \sum_{k \in \Omega_2} U_k(R_k^{DIS}) \tag{24}$$

Subtracting (24) from (23) gives us a bound:

$$\overline{J}^* - U^*_{tot} \leq J^{DIS} - U^{DIS} = \sum_{k \in \Omega_2} \left(\overline{\lambda}_k R_k^{DIS} - U_k(R_k^{DIS})\right)$$

The maximum occurs when Ω_2 consists of all users and when $R_k^{DIS} = R_k^l \ \forall\ k$. The property directly follows.

5 Heuristic search algorithm

The resource allocation algorithm proposed in the previous two sections generates an allocation this is sub-optimal with a bounded duality gap with a complexity that is polynomial in the number of users K. In this section, we seek an algorithm with reduced complexity at the cost of some additional decrease in total user utility. The idea is to separate the decisions about which users should be active and what downlink power and subcarriers to allocate to each such active user. The first stage will focus on active user selection, and use a greedy algorithm that attempts to ensure that every active user will obtain a rate at or above the tangent point rate $R_k^{'}$. The second stage will take the active user set as given and use standard convex optimization techniques to allocate power and subcarriers.

This decomposition will come at the cost of some performance, but we have reason to believe that this degradation should be relatively small. The reasoning is as follows. If utility functions are concave, then total user utility is maximized when all users are active. The optimal subcarrier and power allocation can be accomplished using marginal cost pricing. If utility functions are convex, then total user utility is maximized when only a subset of users are active and almost all active users are assigned their maximum rate.

A sigmoid utility function is convex for $R_k < R_k^f$ and concave for $R_k > R_k^f$. Total user utility is maximized when only a subset of users are active and almost all active users are assigned a rate above their rate at maximum average utility $R_k^{'}$. We can create a decomposition based on this observation. If user selection and subcarrier allocation are done first, and rate scheduling is then done based on the results of the subcarrier allocation, then this reduces complexity at the cost of performance. The complexity reduction comes from not having to search simultaneously for both types of shadow costs. The performance reduction comes from no longer being able to consider subcarrier allocation on the basis of the final assigned rate. However, this performance reduction should be minor since it should primarily affect the small number of users who in the optimal allocation would be assigned rates below $R_k^{'}$.

The first portion of the algorithm must identify users who should be active. Recall from (11) that active users will maximize user utility minus user cost. It follows that except for a small number of users, if user k is active, it should be assigned a rate not only above the inflection point R_k^f but also above $R_k^{'}$, the rate of user k at maximum average utility. The algorithm should thus allocate sets of subcarriers and power to attempt to ensure that active users have rates $R_k \geq R_k^{'}$. A simpler approach, for the purposes of subcarrier assignment, is to assume that the total power P_T is divided equally among all subcarriers and to assign sets of subcarriers to attempt to ensure that active users have rates $R_k \geq R_k^{'}$. However, even this approach is too complex since it requires assigning multiple subcarriers at a time.

For guidance, we look back to the DIS algorithm. If a user currently has been allocated a rate $R_k^i < R_k^{'}$, then $\lambda_k^i =$

$\overline{\lambda}_k$. The power allocation problem (12) for such users thus becomes

$$\max_{\mathbf{p}\in\mathbf{A}} \sum_{\{k|R_k<R_k'\}} \overline{\lambda}_k R_k$$
$$s.t. \quad P_T - \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} = 0$$

We thus adopt a greedy approach, while users have $R_k < R_k'$, by maximizing the weighted rate $\overline{\lambda}_k r_{k,n}$ that can be achieved on each subcarrier one at a time. This greedy approach avoids the difficulties of bin-packing algorithms. If there are sufficient subcarriers to move each user above R_k', then residual subcarriers can be assigned based on the maximum incremental utility, as shadow cost pricing would do. Such a subcarrier assignment algorithm is shown in Table 3.

A further simplification of this algorithm can be achieved, at some performance cost, by proceeding sequentially through the subcarriers rather than searching for the unassigned subcarrier with the highest marginal rate. This algorithm is outlined in Table 4.

At the end of either of these algorithms only the subcarrier assignment is saved; the rates are discarded since they were calculated on the basis of an equal power assignment to each subcarrier. The rate scheduling portion of the algorithm then follows. Recall from (13) that power should be allocated so as to maximize the revenue from selling rate minus the cost of power. This will be iteratively done using steps of power of ΔP. Given the subcarrier assignment, the rate scheduling should attempt to ensure that each active user is ultimately assigned a rate of at least R_k'. A greedy approach is to assign each increment of power to the user with a rate less than R_k' that can gain the greatest $\overline{\lambda}_k \Delta r_{k,n}$ where $\Delta r_{k,n} = r_{k,n}(p_{k,n} + \Delta P) - r_{k,n}(p_{k,n})$ and where $r_{k,n}(\cdot)$ is determined by (1). If all users can be assigned rates of at least R_k', assign incremental power to the user that can gain the greatest utility $\Delta U_{k,n} = U_k(R_k + \Delta r_{k,n}) - U_k(R_k)$. Such a rate scheduling algorithm is outlined in Table 5.

Table 3 Subcarrier assignment

```
Initialize $R_k^0 = 0\ \forall\ k$ and allocate power $P_T/N$ to each subcarrier.
$\mathbf{C} = \{n|$ subcarrier $n$ has not been assigned$\}$
Repeat
    $\mathbf{B} = \{k|R_k^i < R_k'\}$
    If $\mathbf{B}$ is not empty
        Allocate subcarrier $n$ to user
        arg $\max_{k\in\mathbf{B}} \max_{n\in\mathbf{C}} \overline{\lambda}_k r_{k,n}$.
    Else
        Allocate subcarrier $n$ to user
        arg $\max_k \max_{n\in\mathbf{C}} [U_k(R_k^i + r_{k,n}) - U_k(R_k^i)]$.
    Update $R_k^{i+1} = R_k^i + r_{k,n}$.
    $\mathbf{C} = \{n|$ subcarrier $n$has not been assigned$\}$
Until $\mathbf{C}$ is empty
```

Table 4 Sequential subcarrier assignment

```
Initialize $R_k^0 = 0\ \forall\ k$ and allocate power $P_T/N$ to each subcarrier.
For n=1:N
    $\mathbf{B} = \{k|R_k^i < R_k'\}$
    If $\mathbf{B}$ is not empty
        Allocate subcarrier $n$ to user arg $\max_{k\in\mathbf{B}} \overline{\lambda}_k r_{k,n}$.
    Else
        Allocate subcarrier $n$ to user
        arg $\max_k [U_k(R_k^i + r_{k,n}) - U_k(R_k^i)]$.
    Update $R_k^{i+1} = R_k^i + r_{k,n}$.
End For
```

Table 5 Rate Scheduling

```
Initialize $R_k^0 = 0\ \forall\ k,\ p_{k,n} = 0\ \forall\ k, n$.
Repeat
    $\mathbf{B} = \{k|R_k^i < R_k'\}$
    If $\mathbf{B}$ is not empty
        Allocate power $\Delta P$ to user and subcarrier
        arg $\max_{\{k\in\mathbf{B},n\}} \overline{\lambda}_k \Delta r_{k,n}$.
    Else
        Allocate power $\Delta P$ to user and subcarrier
        arg $\max_{\{k,n\}} \Delta U_{k,n}$.
  Update $R_k^{i+1} = R_k^i + \Delta r_{k,n}$.
    Update $P_T = P_T - \Delta P$.
Until $P_T = 0$.
```

We call the combination of the subcarrier assignment algorithm in Table 3 with the rate scheduling algorithm in Table 5 the heuristic search (HS) algorithm, and call the combination of the sequential subcarrier assignment algorithm in Table 4 with the rate scheduling algorithm in Table 5 the heuristic sequential search (HSS) algorithm. The complexity of the HS algorithm is $O(KN(N+1)/2 + NP/\Delta P)$. The complexity of the HSS algorithm is $O(KN + NP/\Delta P)$. The complexity of both algorithms is linear in the number of users; however the HS algorithm is more sensitive to the number of subcarriers.

6 Simulation results

In this section, we examine via simulation the performance of the dual iteration search (DIS) algorithm, the heuristic search (HS) algorithm, and the simplified heuristic search with sequential subcarrier allocation (HSS). For purposes of the simulation, we set $B = 20$KHz. The channel fading

$H_{k,n}$ is determined by an urban propagation model [24][4]. As with previous research on resource allocation, see e.g. [20], the sum of the interference and noise power can be set to an arbitrary level; we use $I + \delta^2 = 1$.[5]

Users have sigmoid utility functions of the form:

$$U_k(R_k) = \begin{cases} aR_k^2, \text{ if } R_k < R_k^f \\ c(R_k + b)^d, \text{ else} \end{cases} \qquad (25)$$

Two utility functions are used

Type 1: $a = (5/6)^{1/3}/25$, $b = -25/6$, $c = 1$, $d = 1/3$, $R_k^f = 5$kbps, $R_k^{'} = 6.25$kbps, $\bar{\lambda}_k = 0.2043$.

Type 2: $a = 1/4 * (2/5)^{1/3}/(12/5)^2$, $b = -2$, $c = 1/4$, $d = 1/3$, $R_k^f = 12/5$kbps, $R_k^{'} = 3$ kbps, $\bar{\lambda}_k = 0.0833$.

The parameters have been chosen so that these two utility functions have different slope at the maximum average utility point.

6.1 Users with identical pathlosses

In this section, we investigate the case in which all users have identical pathlosses $PL_k = 1 \quad (0 \text{ dB}) \quad \forall\ k$, but independent small scale fading. We simulate $K = 10$ users using $N = 500$ subcarriers. The base station downlink power budget P_T is varied from 0.2 to 15, and for the heuristic algorithms we use a power increment $\Delta P = P_T/\ 4{,}000$.

First consider the scenario that all the users have the type 1 utility function. The total utility of all users under each algorithm is shown in Fig. 3. We see that the total utility is an increasing concave function of the power P_T for all three algorithms. This form is to be expected under any reasonable algorithm. Greedy algorithms attempt to allocate capacity in decreasing order of returns, with small errors given by imperfect bin-packing.

It would be beneficial to have a comparison between the DIS algorithm and the optimal solution to (4). Determination of the optimal solution is computationally prohibitive, due to the requirement of solving a large set of fixed point equations. However, we can construct an upper bound using the dual problem. The DIS algorithm uses the solution of the dual problem as an approximation to the solution of the primal problem. The two solutions are identical if the duality gap is zero. The duality gap exists when the primal and dual problems disagree on users with rates below $R_k^{'}$. However, as discussed above, this disagreement should be relatively small. Because $\bar{J}(\lambda, \mu) \geq \bar{J}^* \geq U_{tot}^*$, substituting the solution of DIS algorithm into (9) provides an upper bound on the optimal total user utility of the original problem. In Fig. 3, the total user utility achieved by DIS overlaps with this upper bound. We thus conclude that the duality gap is near zero and that the total user utility achieved by the DIS algorithm is equal to that of the optimal solution.

We now move to consideration of the heuristic algorithms. The performance reduction from the DIS algorithm to the HS algorithm comes from no longer being able to consider subcarrier allocation on the basis of the final assigned rate. The magnitude of this reduction also depends on disagreements between the DIS and HS algorithms on which users are assigned rates below $R_k^{'}$. We expect that the HS algorithm will have only limited such disagreements with the DIS algorithm, and thus the performance degradation should be small. In the simulation, the HS algorithm performs almost identically to the DIS algorithm, with notable degradation only at small values of P_T. In contrast, the HSS algorithm, which reduces complexity further by assigning subcarriers in sequence, does result in a performance degradation of 10–20 %.

To further investigate these differences, in Table 6 we show which users are assigned at least one subcarrier and some power for the same range of P_T. Users who achieve a rate lower than $R_k^{'}$ are marked with a "$-$" superscript, and users activated under the HS or HSS algorithms but not under the DIS algorithm are written in parentheses. Under all three algorithms, decreasing the power P_T generally reduces the set of active users, as expected. When $P_T \geq 5$, all three algorithms activate all 10 users. When $P_T = 1$, both heuristic algorithms make the mistake of activating user 1 (who has the worst average channel) rather than allocating additional subcarriers and power to the other users. When $P_T = 0.5$, all three algorithms agree on not activating user 1 or user 8 (who has the second worst average channel) and on assigning user 3 (who has the worst max channel gain among

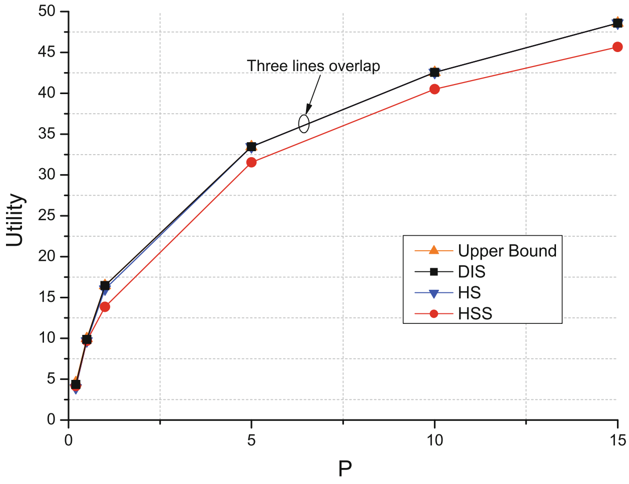


Fig. 3 Simulation result of users with one kind of utility

[4] Rayleigh channel with 6 paths with delays = [0, 0.2, 0.5, 1.6,2.3,5.0] $*10^{-6}$ sec and fading = [1, 0.3, 0.6, $-$ 0.6, $-$ 0.8, $-$ 1]dB, generated using the Matlab routine *rayleighchan*.

[5] Power will scale linearly with $I + \delta^2$.

Table 6 Active users under various algorithms

	DIS	HS	HSS
$P_T = 15$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 10$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 5$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 1$	2, 3, 4, 5, 6, 7, 8, 9, 10	(1), 2, 3, 4, 5, 6, 7, 8, 9, 10	(1), 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 0.5$	2, 3^{-}, 4, 5, 6, 7, 9, 10	2, 3^{-}, 4, 5, 6, 7, 9, 10	2, 3^{-}, 4, 5, 6, 7, 9, 10
$P_T = 0.2$	2, 4, 6^{-}, 9	2, (5^{-}), 6, (10)	2, 4, 6^{-}, 9

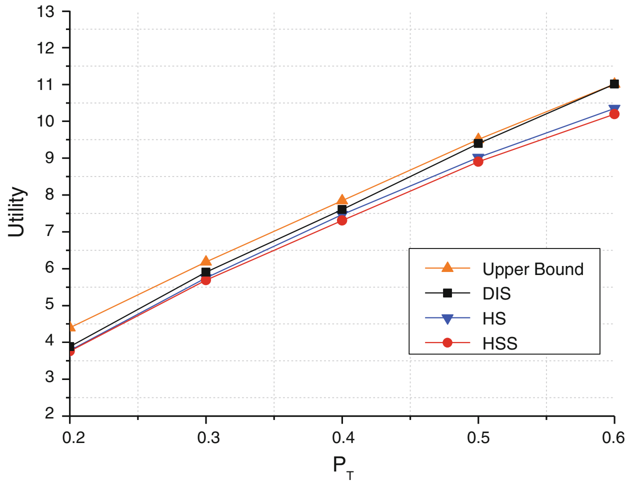


Fig. 4 Simulation result of users with low power

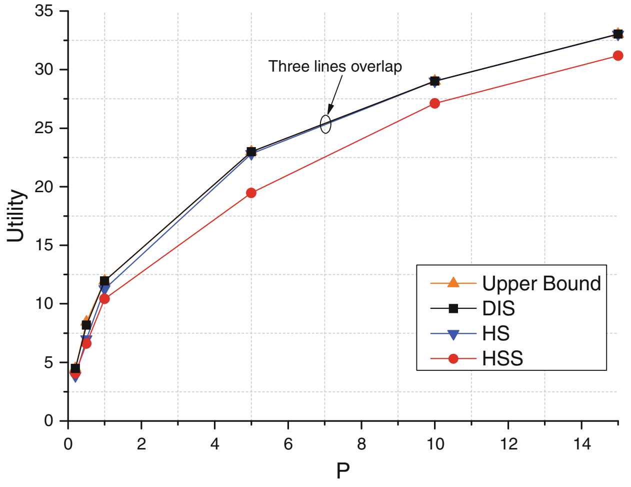


Fig. 5 Simulation result of users with two kinds of utilities

active users) a rate lower than $R_k^{'}$. However, when $P_T = 0.2$, the HS algorithm differs substantially from the DIS algorithm on which users to activate as well as which single user to assign a rate below $R_k^{'}$; it allocates subcarriers to user 5, 10 and try to ensure that $R_5 \geq R_5^{'}$ and $R_{10} \geq R_{10}^{'}$. In contrast the DIS algorithm allocates these subcarriers to users 4 and 9 to maximize profit.

To further investigate the duality gap, we increase the number of users to 20 and focus on the scenario when the total power P_T is low. In Fig. 4, we show the total utility performance. When total power is very low, the rates of most active users are lower than the tangent rate and the duality gap is significant. At higher (but still low) total power, under DIS all users obtain a rate higher than the tangent point; the duality gap decreases to zero when $P_T = 0.6$. In contrast, at similar power levels, the HS and HSS algorithms continue to allocate rates lower than the tangent point to some users. In low power regions, the performance difference between the HS and HSS algorithms is small. At higher power levels than shown, both DIS and HS performance approach the upper bound, but HSS performance continues to be lower than the upper bound.

We now consider a scenario in which the first five users have type 1 utility functions and the next five users have type 2 utility functions. Correspondingly, the maximum average utility per unit rate $\overline{\lambda}_k$ of type 1 users is higher than that of type 2 users. The total utility of all users under each algorithm is shown in Fig. 5. The active users are shown in Table 7. The upper bound is still obtained by substituting the solution of DIS algorithm into (9), which provides an upper bound on the optimal total user utility of the original problem. When $P_T = 0.2$, resources are severely constrained and the only active users are type 1. When $P_T = 0.5$, DIS activates all type 1 users, except user 1 who has a bad channel; it also activates user 9, but does not have enough resources to get this user above $R_9^{'}$. In contrast, when all users have identical utility functions, a greater number of type 2 users are active.

This scenario can also be used to illustrate property 4. We set total power $P_T = 1$ and run the simulation 10,000 time slots. We observe that the proportion of time slots in which type 1 users are assigned zero rate is approximately one half of that of type 2 users.

Table 7 Active users under various algorithms

	DIS	HS	HSS
$P_T = 15$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 10$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 5$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 1$	1, 2, 3, 4, 5, 6, 7, 9, 10	1, 2, 3, 4, 5, 6, 7, (8), 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$P_T = 0.5$	2, 3, 4, 5, 9^-	(1), 2, 3, 4, 5, (6), (10^-)	2, 3, 4, 5, (6), (7^-), 9, (10)
$P_T = 0.2$	2, 4, 5	2, (3), 4^-, 5	2, (3^-), 4, 5

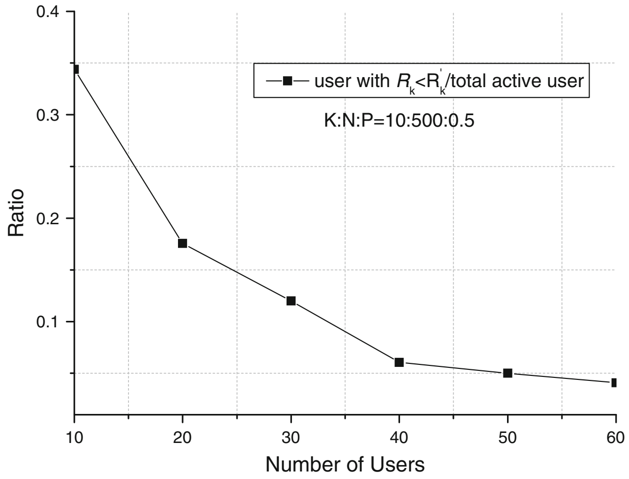


Fig. 6 Large scale behavior

6.2 Users with independent pathlosses

In this last subsection, we investigate the case in which users have independent pathlosses and independent small scale fading. Users are uniformly distributed in a circular cell with radius 1km. The pathloss $PL(x) = 10^{-2}/x^2$ where x is the distance from the user to the base station measured in km.

We are interested in the relationship between the duality gap and the scale of the network. As discussed above, the duality gap exists when the primal and dual problems disagree on users with rates below $R_k^{'}$. the size of the duality gap is related to the number of such users and on the difference between $\overline{\lambda}_k R_k$ and $U_k(R_k)$ for each such user. In property 4, we gave an upper bound on the duality gap for the DIS algorithm. (The duality gap also affects the HS and HSS algorithms, but can be most clearly illustrated with the DIS algorithm.) We conjecture that the percentage of users with rates below $R_k^{'}$ to be decreasing with the size of the network.

To investigate this conjecture, we vary the number of users, K, from 10 to 60. All users have type 1 utility functions. We scale the number of subcarriers and the base station downlink power linearly with the number of users: $N = 50K$ and $P_T = 0.05K$. The ratio of users with rates below $R_k^{'}$ under DIS over total active users as a function of N is illustrated in Fig. 6. The conjecture appears to be correct.

7 Conclusion and discussions

In this paper, we addressed user selection and resource allocation in wireless networks for semi-elastic applications such as video conferencing. We posed a utility maximization problem, and showed that the dual formulation can be used to reduce complexity by exchanging price and demand for power between a power allocation module and the network, and by exchanging price and demand for rate between users and the power allocation module. By solving the dual problem, we can jointly decide the user selection and resource allocation result. We also proposed a heuristic algorithm that further reduces complexity by decomposing the user selection, subcarrier allocation and rate scheduling problems. The heuristic has a computational complexity linear in the number of users, and is shown by simulation to produce near-optimal results.

In future research, we will address call access control for semi-elastic applications. We envision a system in which CAC admits users if capacity is available and perhaps on the basis of the resulting utility over time, and in which packet scheduling maximizes instantaneous total utility of admitted users, as accomplished here. We envision that the packet scheduling algorithm will report back to CAC the resulting outage rate of sigmoid users, and that CAC will use this information when deciding whether to admit future users.

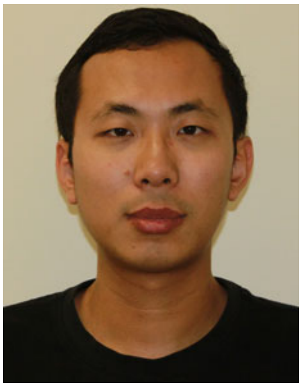
Acknowledgments This work has been supported by the National Science Foundation.

References

1. Sandvine. (2011). Global Internet phenomena report. http://www.sandvine.com
2. Cioffi J. M. (2003). Digital communication. *EE379 course reader*. Stanford: Stanford University.
3. Liu, P., Zhang, P., Jordan, S., & Honig, M. (2004). Single-cell forward link power allocation using pricing in wireless networks. *IEEE Transactions on Wireless Communications. 3*(2), 533–543.

4. Jang, J., & Lee, K. B. (2003). Transmit power adaptation for multiuser OFDM systems. *IEEE Journal on Selected Areas in Communications. 21*(2), 171–178.
5. Shen, Z., Andrews, J., & Evans, B. (2005). Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. *IEEE Transactions on Wireless Communications. 4*(6), 2726–2737.
6. Wong, C. Y., Cheng, R., Lataief, K., & Murch, R. (1999). Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE Journal on Selected Areas in Communications. 17*(10), 1747–1758.
7. Tsang, Y. M., & Cheng, R. S. (2004). Optimal resouce allocation in SDMA/MIMO/OFDM systems under QoS and power constraints. In *Proc. WCNC*, pp. 1595–1600.
8. Gao, X., Nandagopal, T., & Bharghavan, V. (2001). Achieving application level fairness through utility-based wireless fair scheduling. In Proceedings of the GlobeCom, pp. 3257–3261.
9. Feng, N., Mau, S. C., & Mandayam, N. (2004). Pricing and power control for joint network-centric and user-centric radio resource management. *IEEE Transactions on Wireless Communications 52,* 1547–1557.
10. Song, G., & Li, Y. (2005). Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks. *IEEE Communications Magazine 43*(12), 127–134.
11. Zhou, C., Honig, M., & Jordan, S. (2005). Utility-based power control for a two-cell CDMA data network. *IEEE Transactions on Wireless Communications 4*(6), 2764–2776.
12. Yang, C., Wang, W., & Zhang, X. (2009). Multi-service transmission in multiuser cooperative networks. In Proceedings of the WCNC, pp. 1–5.
13. Lee, J., Mazumdar, R., & Shroff, N. (2005). Downlink power allocation for multi-class wireless systems. *IEEE/ACM Transactions on Networking 13*(4), 854–867.
14. Hande, P., Shengyu, Z., & Mung, C. (2007). Distributed rate allocation for inelastic flows. *IIEEE/ACM Transactions on Networking 15*(6), 1240–1253.
15. Cheung, M. H., Mohsenian-Rad, A.-H., Wong, V., & Schober, R. (2010). Random access for elastic and inelastic traffic in WLANs. *IEEE Transactions on Wireless Communications 9*(6), 1861–1866.
16. Abbas, G., Nagar, A. K., & Tawfik, H. (2011). On unified quality of service resource allocation scheme with fair and scalable traffic management for multiclass Internet services. *IET Transaction on Communication 5*(16), 2371–2385.
17. Jin, J., Sridharan, A., Krishnamachari, B., & Palaniswami, M. (2010). Handling inelastic traffic in wireless sensor networks. *IEEE Journal on Selected Areas in Communications 28*(7), 1105–1115.
18. Song, G. C., & Li, Y. (2005). Cross-layer optimization for OFDM wireless networks-part I: Theoretical framework. *IEEE Transactions on Wireless Communications 4*(2), 614–624.
19. Zhou, C., Zhang, P., Honig, M., & Jordan, S. (2004). Two-cell power allocation for downlink CDMA. *IEEE Transactions on Wireless Communications 3*(6), 2256–2266.
20. Ng, T. C.-Y., & Yu, W. (2007). Joint optimization of relay strategies and resource allocations in cooperative cellular networks. *IEEE Journal on Selected Areas in Communications 25*(2), 328–339.
21. Boyd S., Vandenberghe L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
22. Chu, T. G., & Wang, L. (2009). Self-learning PD game with imperfect information on networks. In Proceedings of the CDC, pp. 6864–6869.
23. Boyd, S., Xiao, Li., & Mutapcic, A. (2003). Subgradient methods. *Lecture notes of EE392o*. Stanford: Stanford University.
24. ETSI. (1993). GSM specification 05.05 annex c.

Author Biographies



Chao Yang received the B.S degree in Electronic and Information Engineering and the M.S. degree in Signal and Information Processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2006 and 2009 respectively. He is currently working toward the Ph.D. degree at the University of California, Irvine. His research interests include resource allocation and admission control for both computer networks and cellular networks.

Scott Jordan received the B.S./A.B., the M.S., and Ph.D. degrees from the University of California, Berkeley, in 1985, 1987, and 1990, respectively. From 1990 until 1999, he served as a faculty member at Northwestern University. Since 1999, he has served as a faculty member at the University of California, Irvine. During 2006, he served as an IEEE Congressional Fellow, working in the United States Senate on Internet and telecommunications policy issues. His research interests currently include net neutrality, pricing and differentiated services in the Internet, and resource allocation in wireless multimedia networks.