Downlink Scheduling and Resource Allocation for OFDM Systems

Jianwei Huang, *Member, IEEE*, Vijay G. Subramanian, *Member, IEEE*,
Rajeev Agrawal, and Randall A. Berry, *Member, IEEE*

***Abstract*—We consider scheduling and resource allocation for the downlink of a cellular OFDM system, with various practical considerations including integer tone allocations, different subchannelization schemes, maximum SNR constraint per tone, and "self-noise" due to channel estimation errors and phase noise. During each time-slot a subset of users must be scheduled, and the available tones and transmission power must be allocated among them. Employing a gradient-based scheduling scheme presented in earlier papers reduces this to an optimization problem to be solved in each time-slot. Using a dual formulation, we give an optimal algorithm for this problem when multiple users can time-share each tone. We then give several low complexity heuristics that enforce integer tone allocations. Simulations are used to compare the performance of different algorithms.**

***Index Terms*—Orthogonal frequency division multiplexing (OFDM), WiMax, cellular downlink, scheduling, resource allocation, nonlinear optimization, wireless communications.**

I. Introduction

MOST recent high-speed wireless data systems dynamically schedule users and allocate physical layer resources among them based on the users' channel conditions and quality of service (QoS) requirements. Many of the scheduling algorithms considered can be viewed as "gradient-based" algorithms, which select the transmission rate vector that maximizes the projection onto the (time-varying) gradient of the system's total utility [1]–[4]. Several such algorithms have been studied for time-division multiplexed (TDM) systems, including the "proportionally fair rule" [4], [6] which is based on a logarithmic utility function of each user's throughput. A larger class of throughput-based utilities is considered in [2], [5], where efficiency and fairness are allowed to be traded-off. The "Max Weight" policy (e.g. [7]–[9]) can also be viewed as a gradient-based policy, where the utility is also a function of the user's queue-size or delay.

In TDM systems, one only needs to schedule one user in a time-slot and choose the modulation and coding scheme for that user. However, in many current systems, multiple users may be multiplexed within a time-slot using Orthogonal Frequency Division Multiplexing (OFDM) (e.g. IEEE 802.16/WiMAX [11] and 3GPP LTE [12]). This paper addresses gradient-based scheduling and resource allocation for the downlink of such a system where in addition to determining which users are scheduled, the allocation of physical layer resources (e.g. transmission power and subcarriers) must also be specified.

Our approach is motivated by [10], where a gradient-based scheduling algorithm is used for a system which multiplexes users in a time-slot via code division multiple access (CDMA). Compared to CDMA, OFDM offers more degrees of freedom to allocate resources across (i.e., tone allocation in the frequency domain). This enables exploiting both multi-user diversity and frequency diversity at a finer granularity, but also significantly increases the complexity of the optimization.

At the beginning of each scheduling interval, the gradient-based scheduling algorithm maximizes the weighted throughput sum over the current set of feasible rates. In Section II, we give a model for this rate region, taking into account the following important practical considerations for OFDM systems: 1) different subchannelization techniques in which resource allocation is performed at a larger granularity (i.e, groups of tones or symbols) to reduce the channel measurement and feedback overhead; 2) constraints that each subchannel/tone can be allocated to at most one user; 3) constraints on the maximum rate per tone to model a limitation on the available modulation and coding schemes; and 4) "self-noise" due to channel estimation errors (e.g., [13]) or phase noise [23].

In Section III, we consider a dual formulation for the resulting optimization problem, which enables us to exploit the problem's structure and develop both optimal and simple suboptimal algorithms with low complexity. Simulation results are given in Section IV for these algorithms with dynamically varying weights under different choices of utility functions, subchannelization schemes, self-noise and per tone rate constraints. We conclude in Section V.

A number of related formulations without self-noise and per tone rate constraints for downlink OFDM resource allocation have been studied including [14]–[20]. In [15], the goal is

Manuscript received November 13, 2007; revised April 12, 2008 and July 31, 2008; accepted September 17, 2008. The associate editor coordinating the review of this paper and approving it for publication was D. I. Kim.

J. Huang is with the Dept. of Information Engineering, The Chinese University of Hong Kong (e-mail: jwhuang@ie.cuhk.edu.hk).

V. G Subramanian is with the Hamilton Institute, National University of Ireland (email: vijay.subramanian@nuim.ie).

R. Agrawal is with Motorola Inc. (e-mail: rajeev.agrawal@motorola.com).

R. A. Berry is with the Dept. of EECS, Northwestern University (e-mail: rberry@ece.northwestern.edu).

Part of this work was done while J. Huang and V. G. Subramanian were at Motorola. This work has been supported by the Competitive Earmarked Research Grants (Project Number 412308) established under the University Grant Committee of the Hong Kong Special Administrative Region, China, the Direct Grant (Project Number C001-2050398) of The Chinese University of Hong Kong, SFI grant IN3/03/I346, the Motorola-Northwestern Center for Seamless Communications, NSF CAREER award CCR-0238382, and the National Key Technology R&D Program (Project Number 2007BAH17B04) established by the Ministry of Science and Technology of the People's Republic of China.

Digital Object Identifier 10.1109/T-WC.2009.071266

to minimize the total transmit power given target bit-rates for each user. Sum-rate maximization is considered in [16], [18], [19], where [18], [19] also enforce a minimum bit-rate per user. Weighted sum-rate maximization (for a fixed set of weights) is studied in [14], [20]. In [14], a suboptimal algorithm with constant power per tone was shown in simulations to have little performance loss. Other heuristics that use a constant power per tone are given in [16]–[18]. We also consider such a heuristic in Section III-D. In [20], a similar dual-based algorithm to ours is considered and simulations are given which show that the duality gap of this problem quickly goes to zero as the number of tones increases; we will revisit this in Section III-B. Finally, in [21], the capacity region of a downlink broadcast channel with frequency-selective fading using a TDM scheme is given that covers our rate region without any maximum rate constraints or self-noise.

The previous papers optimize a static objective function while we are interested in the case where the objective changes according to a gradient-based algorithm. It is not *a priori* clear if a good heuristic for a static problem applied to each time-step, will be a good heuristic for the dynamic case, since the optimality result in [1]–[4], [7]–[9] is predicated on solving the optimization problem exactly in each time-slot. Our simulation results show that the heuristics continue to perform well, at least for the scenarios considered in this paper. In a companion paper [25], we use similar methods to solve the corresponding uplink problem. A more general solution framework that encompasses both the uplink and downlink cases is provided in [29].

II. Problem Formulation

We consider downlink transmissions in an OFDM cell from a base station to a set $\mathcal{K} = \{1, \ldots, K\}$ of mobile users. In each time-slot, the scheduling and resource allocation decision can be viewed as selecting a rate vector $\boldsymbol{r}_t = (r_{1,t}, \ldots, r_{K,t})$ from the current feasible rate region $\mathcal{R}(\boldsymbol{e}_t) \subseteq \mathbb{R}_+^K$, where $\boldsymbol{e}_t$ indicates the time-varying channel state information available at the scheduler at time t. Following the gradient-based scheduling framework in [1]–[4], an $\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)$ is selected that has the maximum projection onto the gradient of a system utility function $U(\boldsymbol{W}_t) := \sum_{i=1}^K U_i(W_{i,t})$, where $U_i(W_{i,t})$ is an increasing concave utility function of user i's average throughput, $W_{i,t}$, up to time t. In other words, the scheduling and resource allocation decision is the solution to

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \nabla U(\boldsymbol{W}_t)^T \cdot \boldsymbol{r}_t = \max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_i U_i'(W_{i,t}) r_{i,t}, \quad (1)$$

where $U_i'(\cdot)$ is the derivative of $U_i(\cdot)$. For example, one class of utility functions given in [2], [5] is

$$U_i(W_{i,t}) = \begin{cases} \frac{c_i}{\alpha}(W_{i,t})^\alpha, & \alpha \leq 1,\ \alpha \neq 0, \\ c_i \log(W_{i,t}), & \alpha = 0, \end{cases} \quad (2)$$

where $\alpha \leq 1$ is a fairness parameter and c_i is a QoS weight. With equal class weights, $\alpha = 1$ results in the scheduling rule that maximizes the sum-rate during each slot; $\alpha = 0$ results in the proportionally fair rule.

In general, we consider the problem of

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_i w_{i,t} r_{i,t}, \quad (3)$$

where $w_{i,t} \geq 0$ is a time-varying weight assigned to the ith user at time t tied to the QoS requirements of the user [1]–[4], [7]–[9]. We note that (3) must be re-solved at each scheduling instance because of changes in both the channel state and the weights (e.g., the gradients of the utilities). While the former changes are due to the time-varying nature of wireless channels, the latter changes are due to new arrivals and past service decisions.

A. OFDM capacity regions

The solution to (3) depends on the channel state dependent rate region $\mathcal{R}(\boldsymbol{e})$, where for simplicity we suppress the dependence on time. We consider a model appropriate for downlink OFDM systems; related models have been considered in [14], [21]. In this model, $\mathcal{R}(\boldsymbol{e})$ is parameterized by the allocation of tones to users and the allocation of power across tones. In a traditional OFDM system, at most one user may be assigned to any tone. Initially, as in [15], we make the simplifying assumption that multiple users can share one tone using some orthogonalization technique (e.g. TDM).[1] In practice, if a scheduling interval contained multiple OFDM symbols, we can implement such sharing by giving a fraction of the symbols to each user. We discuss the case where only one user can use a tone in Section III-C.

Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of tones. For each $j \in \mathcal{N}$ and user $i \in \mathcal{K}$, let e_{ij} be the received signal-to-noise ratio (SNR) per unit power. We denote the power allocated to user i on tone j by p_{ij} and the fraction of that tone allocated to user i by x_{ij}. The total power allocation must satisfy $\sum_{i,j} p_{ij} \leq P$, and the total allocation for each tone j must satisfy $\sum_i x_{ij} \leq 1$. For a given allocation, with perfect channel estimation, user i's feasible rate on tone j is $r_{ij} = x_{ij} B \log(1 + \frac{p_{ij} e_{ij}}{x_{ij}})$, which corresponds to the Shannon capacity of a Gaussian noise channel with bandwidth $x_{ij}B$ and received SNR $p_{ij}e_{ij}/x_{ij}$. This SNR arises from viewing p_{ij} as the energy per time-slot user i uses on tone j; the corresponding transmission power becomes p_{ij}/x_{ij} when only a fraction x_{ij} of the tone is allocated. Without loss of generality we set $B = 1$ in the following.

In a realistic OFDM system, imperfect carrier synchronization and channel estimation may result in "self-noise" (e.g. [23], [13]). We model this in a similar way as [13]. Let the received signal on the jth tone of user i be given by $y_{ij} = h_{ij} s_{ij} + n_{ij}$, where h_{ij}, s_{ij} and n_{ij} are the (complex) channel gain, transmitted signal and additive noise, respectively, with $n_{ij} \sim \mathcal{CN}(0, \sigma^2)$. Assume that $h_{ij} = \tilde{h}_{ij} + h_{ij,\delta}$, where $\tilde{h}_{ij}$ is receiver i's estimate of h_{ij} and $h_{ij,\delta} \sim \mathcal{CN}(0, \delta_{ij}^2)$. After matched-filtering, the received signal will be $z_{ij} = \tilde{h}_{ij}^* y_{ij}$ resulting in an effective SNR of

$$\text{Eff-SNR} = \frac{\|\tilde{h}_{ij}\|^4 p_{ij}}{\sigma_{ij}^2 \|\tilde{h}_{ij}\|^2 + \delta_{ij}^2 p_{ij} \|\tilde{h}_{ij}\|^2} = \frac{p_{ij} \tilde{e}_{ij}}{1 + \beta_{ij} p_{ij} \tilde{e}_{ij}}, \quad (4)$$

[1]We focus on systems that do not use superposition coding and successive interference cancellation within a tone, as such techniques are generally considered too complex for practical systems.

where $p_{ij} = \mathrm{E}(\|s_{ij}\|^2)$, $\beta_{ij} = \frac{\delta_{ij}^2}{\|\tilde{h}_{ij}\|^2}$ and $\tilde{e}_{ij} = \frac{\|\tilde{h}_{ij}\|^2}{\sigma_{ij}^2}$.[2] Here, $\beta_{ij} p_{ij} \tilde{e}_{ij}$ is the self-noise term. As in the case without self-noise ($\beta_{ij} = 0$), the effective SNR is still increasing in p_{ij}. However, it now has a maximum of $1/\beta_{ij}$. For the sake of presentation, we assume that $\beta = \beta_{ij}$ for all i and j. The analysis is almost identical if users have different β_{ij}'s.

With self-noise, user i's feasible rate on tone j becomes $r_{ij} = x_{ij} \log(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij} \tilde{e}_{ij}})$, where again x_{ij} models time-sharing of a tone. Under these assumptions, we have

$$\mathcal{R}(\boldsymbol{e}) = \Bigg\{ \boldsymbol{r} : r_i = \sum_j x_{ij} \log\left(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right), \sum_{i,j} p_{ij} \le P, \sum_i x_{ij} \le 1 \ \forall j, (\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X} \Bigg\}, \quad (5)$$

where $\mathcal{X} := \prod_{j=1}^{N} \mathcal{X}_j$, and for all $j \in \mathcal{N}$,

$$\mathcal{X}_j := \left\{ (\boldsymbol{x}^j, \boldsymbol{p}^j) \ge \boldsymbol{0} : x_{ij} \le 1, p_{ij} \le \frac{x_{ij}\tilde{s}_{ij}}{\tilde{e}_{ij}} \ \forall i \right\}, \quad (6)$$

with $\boldsymbol{x}^j := (x_{ij}, \forall i \in \mathcal{K})$ and $\boldsymbol{p}^j := (p_{ij}, \forall i \in \mathcal{K})$. Here, $\tilde{s}_{ij} = \frac{\Gamma_{ij}}{1 - \Gamma_{ij}\beta}$, where $\Gamma_{ij} < 1/\beta$ is a maximum SNR constraint on tone j for user i, e.g., to model a constraint on the maximum rate per tone due to limited availability of modulation and coding schemes. At the cost of additional complexity, we could also include minimum rate constraints to model inelastic traffic, and maximum rate constraints to incorporate buffer sizes.

We assume that $\tilde{e}_{ij}$ is known by the scheduler for all i and j as is β (or δ_{ij}^2). In a frequency division duplex (FDD) system, this knowledge can be acquired by having the base station transmit pilot signals, from which the users can estimate their channel gains and feed them back to the base station. In a time division duplex (TDD) system, these gains can also be acquired by having the users transmit uplink pilots; the base station can then exploit reciprocity to measure the channel gains. In both cases, this feedback information would need to be provided within the channel's coherence time.

B. Subchannelization

With many tones and users, providing pilots and/or feedback per tone can require excessive overhead; e.g., in IEEE 802.16e [11], a channel with bandwidth 1.25Mhz to 20Mhz is divided from 128 to 2048 tones. One way to reduce this overhead is for feedback and resource allocation to be done at the granularity of *subchannels* of disjoint sets of tones, i.e., constant power is used and coding is done across the tones in the same subchannel. Our model can be adapted to this setting by viewing $\mathcal{N}$ as the set of subchannels and $\tilde{e}_{ij}$ as the effective SNR per unit power for user i on the jth subchannel. Specifically, assuming that k tones are bundled into subchannel j, $\tilde{e}_{ij}$ is chosen so that the total rate (given by $x_{ij} \sum_{j_l \in \mathcal{N}_j} \log(1 + \frac{p_{ij}\tilde{e}_{ij_l}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij_l}})$ where $\mathcal{N}_j$ is the set of tones in the jth subchannel and $\tilde{e}_{ij_l}$ is the SNR per unit power for tone j_l) for user i in this subchannel is approximately $k x_{ij} \log(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}})$. Since $\log(1 + \frac{pe}{x + \beta pe})$ is a concave function of e, using Jensen's inequality the rate achieved over a subchannel is upper bounded by taking $\tilde{e}_{ij}$ to be the *arithmetic average* of the channel gains of tones in subchannel j. The rate can be lower bounded using the strict convexity of $\log(l + \exp(y))$ for $y \in \Re$ (with $l > 0$) and Jensen's inequality. If $\beta = 0$, taking $y = \log\left(\frac{pe}{x}\right)$ and $l = 1$ we lower bound the rate by setting $\tilde{e}_{ij}$ equal to the *geometric average* of the subchannel gains. When $\beta > 0$ we take $y = -\log\left(1 + \frac{x}{\beta pe}\right)$ and $l = \beta$, apply Jensen's inequality followed by the arithmetic-mean geometric-mean inequality to lower bound the rate by setting $\tilde{e}_{ij}$ equal to the *harmonic average* of the subchannel gains. The gap between the upper and lower bounds is quite small for reasonable values of pe; for the SNRs achieved by scheduled users in our simulations, we do not see much difference.[3] From here onwards we will use the terms tone/carrier/subchannel to mean the basic allocation unit; the specific distinctions will be clear from the context.

We consider the following subchannelizations: (i) adjacent channelization, where adjacent tones are grouped together as in the optional "band AMC mode" in IEEE 802.16d/e [11]; (ii) interleaved channelization, where tones are (perfectly) interleaved as in the interleaved channelization in IEEE 802.16d/e [11]; and (iii) random channelization, where tones are randomly assigned as in systems that employ frequency hopping as in the Flash OFDM system [24]. Adjacent channelization enables the resource allocation to better exploit frequency diversity. Interleaved or random channelization reduces the variance of the effective SNR across subchannels for each user; when the variance is small, user i can simply feed back a single e_i value. Random channelizations also aid in managing inter-cell interference.

III. Optimal and Suboptimal Algorithms

From (3) and (5), the scheduling and resource allocation problem can be stated as:

$$\max_{(\boldsymbol{x},\boldsymbol{p}) \in \mathcal{X}} V(\boldsymbol{x}, \boldsymbol{p}) := \sum_{i,j} w_i x_{ij} \log\left(1 + \frac{p_{ij}\tilde{e}_{ij}}{x_{ij} + \beta p_{ij}\tilde{e}_{ij}}\right) \quad \text{subject to:} \quad \sum_{i,j} p_{ij} \le P, \text{ and } \sum_i x_{ij} \le 1, \forall j \in \mathcal{N}, \quad (7)$$

where we still assume that users can time-share subchannels. Next we show how to solve (7) via a dual formulation.

[2] This is slightly different from the Eff-SNR in [13] in which the signal power is instead given by $\|h_{ij}\|^4 p_{ij}$; the following analysis works for such a model as well by a simple change of variables. For the problem at hand, (4) seems more reasonable in that the resource allocation will depend only on $\tilde{h}_{ij}$ and not on h_{ij}. We also note that (4) is shown in [22] to give an achievable lower bound on the capacity of this channel.

[3] For example, in our simulations of the optimal algorithm with $\beta = 0.01$, the differences between achieved utilities under arithmetic average and harmonic average approximations are 0.005%, 0.1%, and 0.4% under adjacent, interleaved and random subchannelizations, respectively.

A. Optimal Dual Solution

Consider the Lagrangian, $L(\boldsymbol{x},\boldsymbol{p},\lambda,\boldsymbol{\mu}) := \lambda P + \sum_{j=1}^{N} L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j)$, where

$$L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j) := \sum_{i=1}^{K} w_i x_{ij} \log\left(1+\frac{p_{ij}\tilde{e}_{ij}}{x_{ij}+\beta p_{ij}\tilde{e}_{ij}}\right) + \mu_j\left(1-\sum_{i=1}^{K} x_{ij}\right) - \lambda\sum_{i=1}^{K} p_{ij}, \quad (8)$$

and $\boldsymbol{\mu} = (\mu_j)_{j=1}^{N}$. The corresponding dual function $L(\lambda,\boldsymbol{\mu}) := \max_{(\boldsymbol{p},\boldsymbol{x})\in\mathcal{X}} L(\boldsymbol{x},\boldsymbol{p},\lambda,\boldsymbol{\mu})$ can then be written as

$$L(\lambda,\boldsymbol{\mu}) = \lambda P + \sum_{j=1}^{N} \max_{(\boldsymbol{p}^j,\boldsymbol{x}^j)\in\mathcal{X}_j} L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j). \quad (9)$$

By directly evaluating the Hessian of $x\log(1+\frac{p}{x+\beta p})$ it can be seen that this is jointly concave in (x,p). It follows that Problem (7) is convex and satisfies Slater's condition. Hence, there is no duality gap and so $V^* := \min_{\lambda\geq 0,\boldsymbol{\mu}\geq\mathbf{0}} L(\lambda,\boldsymbol{\mu})$ is the optimal objective value [26].

Next we give a closed-form representation of $L(\lambda,\boldsymbol{\mu})$ in (9). We then show that minimizing $L(\lambda,\boldsymbol{\mu})$ over $\boldsymbol{\mu}$ only requires searching for the maximum of user dependent metrics for each tone j. The only numerical search needed is for the minimization over λ, which is a one-dimensional search.

1) Computing the Dual Function: For a given $\boldsymbol{x}^j$, μ_j and λ, the $\boldsymbol{p}^j$ which obtains the maximum in (9) is given by

$$p_{ij}^*(\boldsymbol{x},\lambda,\boldsymbol{\mu}) = x_{ij}\tilde{p}_{ij}(\lambda) \quad \text{with}$$
$$\tilde{p}_{ij}(\lambda) := \frac{1}{\tilde{e}_{ij}}\left[q\left(\beta,\left(\frac{w_i\tilde{e}_{ij}}{\lambda}-1\right)^+\right)\wedge\tilde{s}_{ij}\right], \quad (10)$$

where $(x)^+ = \max(x,0)$, $a\wedge b = \min(a,b)$, and

$$q(\beta,z) := \begin{cases} z, & \text{if } \beta = 0; \\ \left(\frac{2\beta+1}{2\beta(\beta+1)}\right)\left(\sqrt{1+\frac{4\beta(\beta+1)}{(2\beta+1)^2}z}-1\right), & \text{if } \beta > 0. \end{cases}$$

Figure 1 shows p_{ij}^* in (10) as a function of $\tilde{e}_{ij}$ for $\beta = 0, 0.01$, and 0.1. When $\beta = 0$, (10) becomes a "water-filling" solution in which $p_{ij}^*(\boldsymbol{x},\lambda,\boldsymbol{\mu})$ is non-decreasing in $\tilde{e}_{ij}$. For a fixed $\beta > 0$, due to self-noise, less power may be allocated to "better" subchannels. The constant β case is applicable when the self-noise is due to phase noise as in [23]. On the other hand, when self-noise arises primarily from estimation errors, β may not be constant but could depend on the channel quality. The exact dependence will depend on the details of channel estimation. As an example, we also show a curve for when $\beta(e) = 10/e$, which is motivated by the analysis in [22, Section IV] for the estimation error of a Gauss-Markov channel from a pilot with known power. For that model, when the pilot power is either constant or inversely proportional to channel quality subject to maximum and minimum power constraints (modeling power control), β will be inversely proportional to e. It can be seen that the curve has a different shape and amplitude compared to the $\beta = 0$ case. For simplicity of presentation, we assume constant β's in the remainder of the paper.

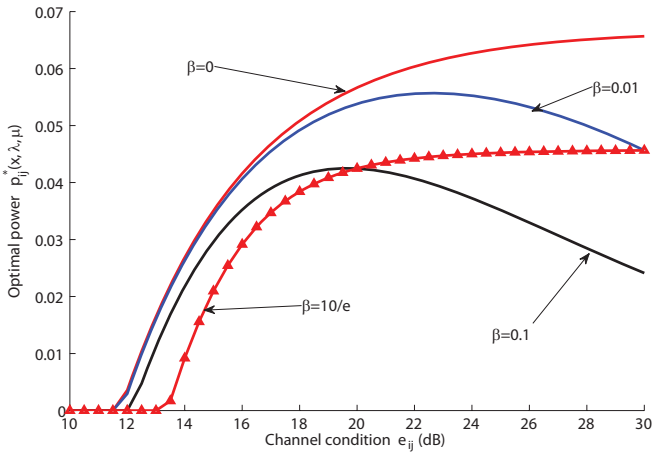

Fig. 1. Optimal power $\tilde{p}_{ij}(15)$ with $w_i = 1$ versus channel condition e_{ij}.

Notice from (10) that the optimal value of p_{ij}^* is always a linear function of x_{ij}. Substituting (10) into $L_j(\boldsymbol{x}^j,\boldsymbol{p}^j,\lambda,\mu_j)$ also results in a linear function of x_{ij}, namely,

$$L_j(\boldsymbol{x}^j,\boldsymbol{p}^{j,*},\lambda,\mu_j) = \sum_i x_{ij}\left(\mu_{ij}(\lambda)-\mu_j\right)+\mu_j,$$

where $\mu_{ij}(\lambda) := w_i h\left(\beta,\frac{w_i\tilde{e}_{ij}}{\lambda},\tilde{s}_{ij}\right)$, and

$$h(\beta,\omega,\tilde{s}_{ij}) := \log\left(1+\frac{q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}}{1+\beta\left(q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}\right)}\right) - \frac{1}{\omega}\left(q\left(\beta,(\omega-1)^+\right)\wedge\tilde{s}_{ij}\right).$$

From this it follows that any choice

$$x_{ij}^*(\lambda,\boldsymbol{\mu}) \in \begin{cases} \{1\}, & \text{if } \mu_{ij}(\lambda) > \mu_j; \\ [0,1], & \text{if } \mu_{ij}(\lambda) = \mu_j; \\ \{0\}, & \text{if } \mu_{ij}(\lambda) < \mu_j, \end{cases} \quad (11)$$

will maximize $L_j(\boldsymbol{x}^j,\boldsymbol{p}^{j,*},\lambda,\mu_j)$. Hence, $L(\lambda,\boldsymbol{\mu}) := \lambda P + \sum_{j=1}^{N} L_j(\lambda,\mu_j)$, where

$$L_j(\lambda,\mu_j) = \sum_i \left(\mu_{ij}(\lambda)-\mu_j\right)^+ + \mu_j. \quad (12)$$

2) Optimizing the Dual Function over λ and $\boldsymbol{\mu}$: Lemma 1 characterizes the optimization of $L(\lambda,\boldsymbol{\mu})$ over $\boldsymbol{\mu}$.

Lemma 1: For all $\lambda \geq 0$,

$$L(\lambda) := \min_{\boldsymbol{\mu}\geq\mathbf{0}} L(\lambda,\boldsymbol{\mu}) = \lambda P + \sum_j \mu_j^*(\lambda), \quad (13)$$

where for every tone j, the minimizing value of μ_j^* is

$$\mu_j^*(\lambda) = \max_i \mu_{ij}(\lambda). \quad (14)$$

Proof of Lemma 1 is similar to the proof in [10]. For each tone j, (14) computes the maximum of user metric μ_{ij}.

Since $L(\lambda)$ is the minimum of a convex function over a convex set, it is a convex function of λ; hence, it can be minimized using an iterated one dimensional search (e.g., the Golden Section method [26] for which the computation complexity is $O(\log(1/\epsilon))$, where ϵ is the target relative error bound). Since there is no duality gap, this minimization gives the optimal objective value in (7).

B. Optimal primal variables with time-sharing

Now we find optimal values of the primal variables $(\boldsymbol{x}, \boldsymbol{p})$. For every $\lambda \geq 0$, with $\boldsymbol{\mu}^*(\lambda)$ as in (14), let

$$(\boldsymbol{x}^*, \boldsymbol{p}^*) := \arg \max_{(\boldsymbol{x},\boldsymbol{p})\in\mathcal{X}} L\left(\boldsymbol{x}, \boldsymbol{p}, \lambda, \boldsymbol{\mu}^*(\lambda)\right); \qquad (15)$$

note that these satisfy (10) and (11).

Given that $\lambda = \lambda^*$, it follows from duality theory, that if the $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ satisfying (15) are primal feasible and satisfy complimentary slackness, then they are primal optimal primal. In particular, if for each tone j there exists a unique user i that achieves the maximum in (14), then since there is no duality gap, allocating tone j only to that user must be primal optimal. In general, given $\lambda \geq 0$, let $\mathcal{A}_j := \{i | \mu_{ij}^*(\lambda) = \max_i \mu_{ij}^*(\lambda)\}$ be the set of users who achieve the maximum on tone j, and $|\mathcal{A}_j|$ be the size of $\mathcal{A}_j$. From (11) it follows that all $\boldsymbol{x}^*$ that solve (15) are those that satisfy the following properties: (i) for $i \notin \mathcal{A}_j$, $x_{ij}^* = 0$; (ii) if $|\mathcal{A}_j| = 1$, then $x_{ij}^* = 1$ for $i \in \mathcal{A}_j$; and (iii) if $|\mathcal{A}_j| > 1$, then for all $i \in \mathcal{A}_j$, $x_{ij}^* \in [0, 1]$ and $\sum_{i\in\mathcal{A}_j} x_{ij}^* = 1$. In case (*iii*), not all tone allocations satisfying $\sum_{i\in\mathcal{A}_j} x_{ij}^* = 1$ may be primal feasible (e.g., $\sum_{ij} p_{ij}^*$ maybe larger than P). Breaking these ties is necessary to find a primal optimal solution. A key point is that when ties occur at a given λ, $L(\lambda)$ may not be not differentiable at that λ. However, since $L(\lambda)$ is a convex function, subgradients exist [27].

Proposition 1: For any $\lambda \geq 0$, d is a subgradient of $L(\lambda)$ if and only if there exists $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ satisfying (15), $\sum_i x_{ij}^* \leq 1$ for all j, $\mu_j^*(\lambda)\left(1 - \sum_i x_{ij}^*\right) = 0$ for each j, and $P - \sum_{i,j} p_{ij}^* = d$.

The proof of Proposition 1 can be found in [28] and follows by observing that that dual function is the maximum of a set of Lagrangian functions which are linear in λ and that the gradient of each of the Lagrangian functions (with respect to λ) is given by $P - \sum_{i,j} p_{ij}$. At any given λ, we need to restrict attention to the maximizing $\boldsymbol{x}^*, \boldsymbol{p}^*$ to obtain the set of subgradients of $L(\lambda)$. The rest follows by observing that the resulting subgradient $P - \sum_{i,j} \tilde{p}_{ij}(\lambda) x_{ij}^*$ is linear in x_{ij}^*, which takes values in a convex set (product of simplexes).

Thus, in order to find the dual optimal, we need to search for λ^* which has a zero subgradient (if $\lambda^* > 0$; and non-negative if $\lambda^* = 0$). From Proposition 1, this will also be the check for primal feasibility and complimentary slackness for the power constraint. Next we provide a solution for this check. We refer to an allocation as an *extreme point* if it satisfies (*i*)-(*iii*) and $x_{ij}^* \in \{0, 1\}$ for all i and j; such an allocation can be represented by a function $f : \mathcal{N} \to \mathcal{K}$, so that $f(j) \in \mathcal{A}_j$ indicates the user who is allocated channel j, i.e., $x_{f(j)j}^* = 1$. Let $\mathcal{B} = \{j : |\mathcal{A}_j| = 1\}$ and $\mathcal{B}^c = \mathcal{N} \setminus \mathcal{B}$. For each $j \in \mathcal{B}$, there are no ties, and so $f(j)$ is unique. For each tone $j \in \mathcal{B}^c$, there are $|\mathcal{A}_j|$ users in the tie, and so the total number of extreme points is $\prod_{j\in\mathcal{B}^c} |\mathcal{A}_j|$. Each extreme point satisfies Proposition 1 and so provides a subgradient for $L(\lambda)$. From Proposition 1 it follows that all the subgradients of $L(\lambda)$ can be obtained as a convex combination of the values at the extreme points. Given an extreme point f, from (10), it follows that the corresponding subgradient $d(f)$ is given by

$$d(f) = P - \sum_{j\in\mathcal{B}} \tilde{p}_{f(j)j} - \sum_{j\in\mathcal{B}^c} \tilde{p}_{f(j)j}. \qquad (16)$$

Choosing different extreme points only effects the last term on the right of (16). It follows that the maximum subgradient of $L(\lambda)$ corresponds to the extreme points given by

$$\hat{f}(j) := \arg \min_{i\in\mathcal{A}(j)} \tilde{p}_{ij}, \ \forall j. \qquad (17)$$

The minimum subgradient corresponds to the extreme points

$$\bar{f}(j) := \arg \max_{i\in\mathcal{A}(j)} \tilde{p}_{ij}, \ \forall j. \qquad (18)$$

At λ^*, the maximum subgradient (using (17)) is always nonnegative, and the minimum subgradient (using (18)) is always non-positive. If either is zero, an integer primal optimal solution is found. In general, we have the following:

Proposition 2: There exists an optimal primal solution $(\boldsymbol{x}^*(\lambda^*), \boldsymbol{p}^*(\lambda^*))$, where $\boldsymbol{x}^*(\lambda^*)$ is given by time-sharing between the two extreme points in (17) and (18) so that the convex combination of the corresponding subgradients is equal to zero, and $\boldsymbol{p}^*(\lambda^*)$ is given by (10).

Proposition 2 implies that each time-shared tone is shared in the same proportion.

The above steps give an algorithm for finding the optimal solution to (7) in two stages. First, find λ^* that minimizes $L(\lambda)$ as in Section III-A. This involves evaluating $L(\lambda)$ for a fixed value of λ as an inner loop, and a one-dimensional search over λ as an outer loop. The outer loop has a complexity that is independent of N and K. The inner loop has a complexity of $O(NK)$ due to searching for the maximum of K metrics (14) on each of the N tones. Thus the total complexity of this stage is $O(NK)$. Second, given λ^*, we compute the maximum and minimum extreme points and find the optimal primal variables as in Proposition 2 which also has a complexity of $O(NK)$. Hence, the overall complexity of the optimal algorithm is $O(NK)$.

In our simulations, the actual complexity of the second stage is typically much smaller than $O(NK)$ because "typically" only a few ties occur. [4] However, the number of extreme points can be very large under interleaved channelization. This is because if two users are tied on one subchannel, it is very likely that they will also be tied on other subchannels since all subchannels have roughly the same channel gain for the same user. However, if all the ties are due to the same two users, we can just allocate all subchannels with a tie to the same user and this will lead to either the largest or smallest subgradient. These observations are consistent with [20], which argues that an OFDM system with $\beta = 0$ in which no time-sharing is allowed will have a certain "duality gap" that is small for a reasonable number of sub-channels. Problem (7) can be viewed as the dual of the dual problem in [20, eqn. (9)] and the duality gap in [20] can be viewed as a measure of the accuracy of approximating the OFDMA scheduling problem by the time-sharing version of it from (7). When there is exactly one extreme point, the duality gap is clearly zero (since we have an integer solution). The arguments in [20] for a vanishing duality gap roughly correspond to showing

[4] For example, extensive simulation results show that for a system of 64 subchannels (grouped from 512 tones) and 40 users in a high mobility environment, there are on average only two extreme points typically on one subchannel involving two users, at each scheduling interval (averaged over 3000 scheduling intervals) under either adjacent or random channelizations.

that the spread in the power consumption of different extreme points (i.e., the maximum difference in subgradient values) is typically small for a reasonable number of carriers. When this spread is small, one expects that fewer ties occur which is consistent with the above discussion. Discussions above also argue that the conclusions in [20] extend to the $\beta > 0$ case.

C. Single user per tone

We now consider the case where no time-sharing is allowed, i.e., $x_{ij} \in \{0, 1\}$ for all i and j. Suppose we still find the optimal λ^* as in Section III-A. If there are no ties on any of the tones or if there is an extreme point with $\sum_{j\in\mathcal{N}} \tilde{p}_{f(j)j} = P$, the optimal primal solution given in Section III-B only has one user per tone, and we are done. If not, Proposition 2 will no longer give a solution that satisfies the integer constraints. In this case, a reasonable heuristic is to simply choose one extreme point allocation. In our simulations, we choose the extreme point corresponding to the subgradient with the smallest non-negative value; i.e., the extreme point f, for which $\sum_{j\in\mathcal{N}} \tilde{p}_{f(j)j}$ is closest to P, without exceeding it. Other rules for choosing an extreme point can also be used. Note that this requires searching over all extreme points, which has a worst-case complexity of $O(K^N)$ (if all users were tied on every tone). However, as discussed above, typically there are only two users tied on one tone and so this has almost constant complexity. If instead the largest or smallest subgradient was used, the worst-case complexity would again be $O(NK)$.

For a given extreme point f, the total transmit power $\sum_{j\in\mathcal{N}} \tilde{p}_{f(j)j}$ will be either greater or less than the constraint P (unless this point is optimal). We then need to re-optimize the power allocation for the given fixed feasible tone allocation $\boldsymbol{x}$ (i.e., $x_{ij} = 1$ if $i = f(j)$, otherwise x_{ij}=0), i.e., solve

$$\max_{\boldsymbol{p}:(\boldsymbol{p},\boldsymbol{x})\in\mathcal{X}} V(\boldsymbol{x}, \boldsymbol{p}) \quad \text{s.t.} \sum_{i,j} p_{ij} \le P. \tag{19}$$

Let $L_{\boldsymbol{x}}(\lambda)$ be the dual function for this problem. Given $\tilde{\lambda} = \arg\min_{\lambda\ge 0} L_{\boldsymbol{x}}(\lambda)$, the optimal power allocation to (19) is given by (10) with $\lambda = \tilde{\lambda}$ and the given tone allocation $\boldsymbol{x}$. A simple one-dimensional search once again yields the optimal λ. This will have a complexity of $O(N)$ (to get within ϵ of the optimal) since each tone has at most one user.

When the self-noise term $\beta = 0$, we can actually find the optimal $\tilde{\lambda}$ in finite steps based on the following alternative characterization of $\tilde{\lambda}$, the proof of which is based on a similar argument as in [10].

Proposition 3: For $\beta = 0$ a given $\hat{\lambda}$ is the unique optimal solution to the dual problem $\min_{\lambda\ge 0} L_{\boldsymbol{x}}(\lambda)$ if and only if

$$\hat{\lambda} = \frac{\sum_{i,j} x_{ij} w_i 1_{\{\hat{\lambda}\in\mathcal{W}_{ij}\}}}{P - \sum_{i,j} \frac{\Gamma_{ij}}{e_{ij}} 1_{\{\hat{\lambda}\in\mathcal{Y}_{ij}\}} + \sum_{i,j} \frac{1}{e_{ij}} 1_{\{\hat{\lambda}\in\mathcal{W}_{ij}\}}}, \tag{20}$$

where $\mathcal{W}_{ij} = \left[\frac{x_{ij} w_i e_{ij}}{1+\Gamma_{ij}}, x_{ij} w_i e_{ij}\right)$, and $\mathcal{Y}_{ij} = \left[0, \frac{x_{ij} w_i e_{ij}}{1+\Gamma_{ij}}\right)$.

Proposition 3 suggests the following algorithm [28] for finding $\tilde{\lambda}$. First check if the power constraint is violated when all users use maximum power on the allocated tones, i.e., if $\sum_{(i,j)} \frac{x_{ij}}{e_{ij}} \Gamma_{ij} > P$. If this is false, the problem is solved. Otherwise, we need to search for $\tilde{\lambda}$ by starting from the largest λ, and calculating the right side of (20). If the result is less than the chosen value of λ, then we decrease λ and recalculate, until a fixed-point is found. It can be shown that the algorithm will stop [28] in at most $2N$ steps at the correct $\tilde{\lambda}$. This algorithm sorts $2N$ values and thus, has a complexity of $O(N \log N)$ which is larger than the $O(N)$ complexity of the one-dimensional search, but yields the exact optimal solution in finite time as opposed to an ϵ-optimal solution. However, regardless of how the power is allocated, we first need to find the optimal λ^*. It follows that if the largest or smallest subgradients are used to break ties, the overall algorithm will have a complexity of $O(NK)$ or $O(NK + N \log N)$ depending on how the power is re-optimized.

D. Single sort suboptimal algorithm

Now we introduce two sub-optimal algorithms that do not require finding the optimal λ^* iteratively. Instead, a carrier allocation is determined by a single sort on each tone based on some easily calculated metric. These heuristic algorithms are much faster than the previous algorithms, although their complexity is again $O(NK)$.

1) HEURISTIC 1: Each subchannel j is allocated to the user with the largest value of $w_i \bar{R}_{ij}$, where

$$\bar{R}_{ij} = \log\left[1 + \left(\tilde{s}_{ij} \wedge \left(\frac{\tilde{e}_{ij} P/N}{1 + \beta \tilde{e}_{ij} P/N}\right)\right)\right]$$

is the rate user i could achieve on subchannel j under power allocation P/N. Any ties are broken arbitrarily, and power allocation P/N is used. This metric was motivated in part by work in [14], [16] where a uniform power allocation (not necessarily over all tones) was shown to be nearly optimal.

2) HEURISTIC 2: Here subchannels are allocated as in HEURISTIC 1. However, after this procedure, an optimal power allocation is performed as in Section III-C (instead of power allocation P/N). It may turn out that no power is assigned to some subchannels.

IV. Simulation Study

We report simulation results based on a realistic OFDM simulator with assumptions and parameters commonly used in IEEE 802.16 standards [11]. We focus on the following algorithms: the OPTIMAL algorithm which finds the optimal λ^* and then chooses a tone-allocation with one user per tone as described in Section III-C[5], and HEURISTIC 1 and HEURISTIC 2 from Section III-D.

We simulate a single OFDM cell with $M = 40$ users and a total transmission power of $P = 6$W at the base station. The channel gains e_{ij}'s are the product of a fixed location-based term for each user i and a frequency-selective fast-fading term. The location-based components are picked using an empirically obtained distribution for many users in a large system. The fast-fading term is generated using a block-fading model based upon the Doppler frequency (for

[5]We simulated both the algorithms in Section III-B and III-C, and found that they have identical performance under all parameter choices. This could be due to the fact that the gap in making the time-sharing assumption is small owing to there being very few significantly different extreme points at each scheduling interval as discussed at the end of Section III-B. We thus refer to the algorithm in Section III-C simply as the OPTIMAL algorithm.

the block-length in time) and a standard reference mobile delay-spread model (for variation in frequency). For the fast-fading terms, each multi-path component is held fixed for $2m$sec (i.e., a fading block length), which corresponds to a 250Hz Doppler frequency. The delay-spread is set to 1μsec. The users' channel conditions are averaged over the applicable subchannelization scheme and fed back to the scheduler.

We consider a system bandwidth of 5MHz consisting of 512 OFDM tones, grouped into 64 subchannels (8 tones per subchannel). The symbol duration is 100μsec with a cyclic prefix of 10μsec, which roughly corresponds to 20 OFDM symbols per fading block (i.e., $2m$sec). This is one of the allowed configurations in the IEEE 802.16 standards [11]. Resource allocation (i.e. solving (1)) is done once per fading block. All the results are averaged over the last 2000 OFDM symbols out of 60000 OFDM symbols (i.e., 3000 fading blocks) by which time we can be reasonably confident that the system has reached stationarity. All users are infinitely back-logged and assigned a throughput-based utility as in (2) with parameter $c_i = 1$ and the same fairness parameters (α) across users.

The rate of user i on subchannel j is calculated as

$$r_{ij} = 0.28Bx_{ij}\log\left(1+\frac{0.56p_{ij}\tilde{e}_{ij}}{x_{ij}+\beta p_{ij}\tilde{e}_{ij}}\right),$$

where B is the subchannel bandwidth. Here 0.56 accounts for the "SNR gap" due to limited modulation and coding choices and 0.28 accounts for various factors such as hybrid ARQ transmission scheme and the overhead due to guard tones and control symbols, etc. While the scheduling is based on the geometric average for $\beta = 0$ and harmonic average for $\beta > 0$, the decoded rate is based on per tone channel conditions.

The first set of simulation results are for a system with adjacent channelization, no self-noise ($\beta = 0$), and no per-user SNR constraints (i.e., $\Gamma_{ij} = \infty$ for all i and j). Table I shows the results for all three algorithms under different choices of the utility parameter α. The column "Utility" gives the average utility per user for each algorithm. The column "log U" shows the log utility per user; this gives an alternate indication of the "fairness" of the resulting allocation (same as utility for $\alpha = 0$). The column "Rate" is the average throughput per user in Kbps, and the final column is the average number of users scheduled per scheduling interval. For each choice of α, the three algorithms perform close to each other for each of these metrics. HEURISTIC 2 performs better than HEURISTIC 1, since the former re-optimizes the power allocation after tone allocation, and the latter just uses constant power allocation. When $\alpha = 1$ (maximum throughput), all three algorithms have almost identical performance.

Next we consider the effect of different subchannelization schemes. Table II shows the performance of the three algorithms for the adjacent, random, and interleaved channelization schemes from Section II-A. We set $\alpha = 0.5$, $\beta = 0$, and $\Gamma_{ij} = \infty$ for all i and j. Again, both HEURISTIC algorithms perform close to the OPTIMAL algorithm. In all cases, interleaved and random channelizations result in lower utility than the adjacent channelization. This is likely due to higher frequency diversity with adjacent channelization.

TABLE I
PERFORMANCE FOR DIFFERENT CHOICES OF α (ADJACENT CHANNELIZATION, NO-SELF-NOISE, NO SNR CONSTRAINTS).

α	Algorithm	Utility	Log U	Rate	Num.
0	OPTIMAL	10.74	10.74	60.8	7.73
0	HEURISTIC 1	10.66	10.66	54.6	7.29
0	HEURISTIC 2	10.72	10.72	57.3	7.35
0.5	OPTIMAL	545.2	10.83	105.9	7.32
0.5	HEURISTIC 1	528.8	10.73	99.3	7.20
0.5	HEURISTIC 2	542.8	10.81	103.2	7.01
1	OPTIMAL	261677	6.79	261.7	2.58
1	HEURISTIC 1	261676	6.79	261.7	2.58
1	HEURISTIC 2	261676	6.77	261.7	2.58

TABLE II
PERFORMANCE OF DIFFERENT SUBCHANNELIZATION SCHEMES ($\alpha = 0.5$, NO SELF-NOISE, NO SNR CONSTRAINTS).

Channelization	Algorithm	Utility	Log U	Rate	Num.
Adjacent	OPTIMAL	545.15	10.83	105.9	7.32
Adjacent	HEURISTIC 1	528.83	10.73	99.3	7.20
Adjacent	HEURISTIC 2	542.84	10.81	103.2	7.01
Interleaved	OPTIMAL	494.61	10.53	92.4	1.79
Interleaved	HEURISTIC 1	486.40	10.47	88.4	1.14
Interleaved	HEURISTIC 2	487.02	10.48	87.8	1.15
Random	OPTIMAL	487.53	10.53	89.2	4.89
Random	HEURISTIC 1	479.07	10.46	84.2	4.39
Random	HEURISTIC 2	485.63	10.51	86.5	4.34

Indeed, for the channel model used here, in the interleaved case all subchannels can be shown to be almost identical, explaining why it typically schedules only one or two users.

Next we consider the case when the self-noise coefficient $\beta = 0.0056$ in Table III. Here we assume $\alpha = 0.5$, and no per-user SNR constraint. The performance gap between the three algorithms is slightly larger compared to the case without self-noise in Table II.

Figure 2 shows the throughput CDFs for all three algorithms, with $\beta = 0.0056$ and $\beta = 0$. Here adjacent channelization is used, $\alpha = 0.5$, and $\tilde{s}_{ij} = \infty$ for all i and j. It is clear that users achieve better throughput when there is no self-noise ($\beta = 0$). For each β the OPTIMAL algorithm always achieves better rates compared to the HEURISTIC ones.

Table IV illustrates the effect of SNR constraints. In particular, we choose the SNR constraint to be ∞, 32.5dB, and 22.5dB, respectively, and the same across all users and all tones. We choose adjacent channelization with utility parameter $\alpha = 0.5$ and no self-noise. Compared to the no SNR constraints case, a constraint of 32.5dB does not change the results significantly, while a constraint of 22.5dB substantially decreases the achievable rates (13% for the OPTIMAL algorithm and 27% for HEURISTIC 1 algorithm).

TABLE III
PERFORMANCE OF DIFFERENT SUBCHANNELIZATION SCHEMES ($\alpha = 0.5$, $\beta = 0.0056$, NO SNR CONSTRAINTS).

Channelization	Algorithm	Utility	Log U	Rate	Num.
Adjacent	OPTIMAL	512.20	10.82	82.5	7.52
Adjacent	HEURISTIC 1	489.32	10.70	73.7	7.40
Adjacent	HEURISTIC 2	504.00	10.78	77.2	7.22
Interleaved	OPTIMAL	467.00	10.51	73.5	1.98
Interleaved	HEURISTIC 1	453.16	10.43	66.8	1.26
Interleaved	HEURISTIC 2	454.59	10.44	66.9	1.27
Random	OPTIMAL	460.53	10.51	71.6	5.60
Random	HEURISTIC 1	446.58	10.42	64.7	4.89
Random	HEURISTIC 2	453.51	10.48	66.1	4.85

TABLE IV
PERFORMANCE OF DIFFERENT SNR CONSTRAINTS (ADJACENT CHANNELIZATION, $\alpha = 0.5$, NO SELF-NOISE).

SNR Max	Algorithm	Utility	Log U	Rate	Num.
∞	OPTIMAL	545.15	10.83	105.9	7.32
∞	HEURISTIC 1	528.83	10.73	99.3	7.20
∞	HEURISTIC 2	542.84	10.81	103.2	7.01
32.5dB	OPTIMAL	542.78	10.83	102.97	7.33
32.5dB	HEURISTIC 1	519.81	10.72	91.87	7.25
32.5dB	HEURISTIC 2	535.89	10.81	96.35	7.10
22.5dB	OPTIMAL	522.48	10.82	88.11	7.40
22.5dB	HEURISTIC 1	483.50	10.66	72.60	7.09
22.5dB	HEURISTIC 2	505.81	10.77	78.61	6.92

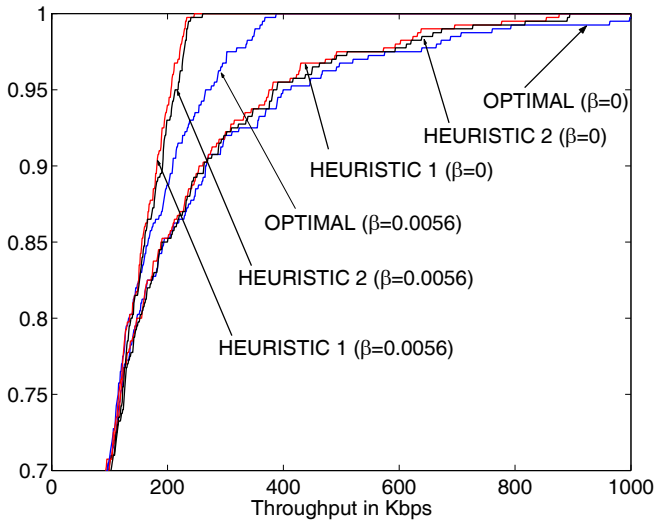

Fig. 2. Empirical CDF of users' throughputs (adjacent channelization, $\alpha = 0.5$, no per-user SNR constraints).

V. CONCLUSIONS

We have considered the problem of gradient-based scheduling and resource allocation for a downlink OFDM system, which essentially reduces to solving a convex optimization problem in each time-slot. We studied this problem for a model that accommodates various choices for user utility functions, different subchannelization techniques, and self-noise due to imperfect channel estimates or phase noise. Using duality theory we first gave an optimal algorithm for solving a relaxed version of this problem in which users can time-share each subchannel. This involves finding a maximum of a per user (closed-form) metric for each subchannel and a one-dimensional search of an optimal dual variable. More interestingly, this algorithm typically automatically yields an integer carrier allocation (except on one or two tones). To enforce such a constraint on all tones, we further proposed an algorithm that picks an integer carrier allocation and re-optimizes the power allocation accordingly. The numerical performance of this algorithm is almost identical to the optimal solution of the relaxed problem. Finally, we proposed two even simpler suboptimal algorithms that only perform a single sort on each of the tones and avoid any iterative calculations. Simulations show that the suboptimal algorithms achieve close to optimal performance under a wide range of scenarios, and the performance gap widens when per user SNR constraints or channel estimation errors are considered.

REFERENCES

[1] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. 2002 Allerton Conference*, 2002.
[2] R. Agrawal, A. Bedekar, R. La, and V. Subramanian, "A class and channel-condition based weighted proportionally fair scheduler," in *Proc. ITC 2001*, Salvador, Brazil, Sept. 2001.
[3] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, 2005.
[4] H. Kushner and P. Whiting, "Asymptotic properties of proportional-fair sharing algorithms," in *Proc. 40th Annual Allerton Conference*, 2002.
[5] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.
[6] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE VTC*, Spring, 2000.
[7] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queue with randomly varying connectivity," *IEEE Trans. Inform. Theory*, vol. 39, pp. 466–478, 1993.
[8] M. Andrews, *et al.*, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.
[9] A. L. Stolyar, "Maximizing queueing network utility subject to stability: greedy primal-dual algorithm," *Queueing Syst.*, vol. 50, no. 4, pp. 401–457, 2005.
[10] R. Agrawal, V. Subramanian, and R. Berry, "Joint scheduling and resource allocation in CDMA systems," in *Proc. WiOpt*, Cambridge, UK, Mar 2004 (journal version under submission).
[11] "IEEE 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005" (http://www.ieee802.org/16/).
[12] "Long Term Evolution of the 3GPP radio technology," `http://www.3gpp.org/Highlights/LTE/LTE.htm`.
[13] H. Jin, R. Laroia, and T. Richardson, "Superposition by position," in *Proc. IEEE ITW 2006*, Mar. 2006.
[14] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms," *IEEE Trans. Commun.*, vol. 52, no. 6, pp. 922–930, 2004.
[15] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, 1999.
[16] J. Jang and K. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Select. Areas Commun.*, vol. 21, no. 2, pp. 171–178, 2003.
[17] Y. Zhang and K. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1566–1575, 2004.
[18] T. Chee, C. Lim, and J. Choi, "Adaptive power allocation with user prioritization for downlink orthogonal frequency division multiple access systems," in *Proc. ICCS 2004*, pp. 210–214, 2004.
[19] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *Proc. IEEE Globecom*, 2000.
[20] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. IEEE ISIT*, pp. 1394–1398, 2006.

[21] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels–part I: ergodic capacity," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.
[22] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 935-946, May 2000.
[23] J. Lee, H. Lou and D. Toumpakaris, "Analysis of phase noise effects on time-direction differential OFDM receivers," in *Proc. IEEE GLOBECOM*, 2005.
[24] QUALCOMM Flarion Technologies, http://www.qualcomm.com/qft/.
[25] J. Huang, V. Subramanian, R. Berry, and R. Agrawal, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Select. Areas Commun.*, to appear, 2009.
[26] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.
[27] R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.
[28] J. Huang, V. Subramanian, R. Berry, and R. Agrawal, "Downlink scheduling and resource allocation for OFDM systems," tech. report. Available: http://personal.ie.cuhk.edu.hk/~jwhuang/publication/OFDM_DL_Technical_Report.pdf.
[29] J. Huang, V. G. Subramanian, R. Berry, and R. Agrawal, "Scheduling and Resource Allocation in OFDMA Wireless Systems," book chapter, submitted.

Jianwei Huang (S'01-M'06) is an Assistant Professor in Information Engineering Department at the Chinese University of Hong Kong. He received the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Northwestern University in 2003 and 2005, respectively. From 2005 to 2007, He worked as a Postdoctoral Research Associate at Princeton University. His main research interests lie in the area of modeling and performance analysis of communication networks, including cognitive radio networks, OFDM and CDMA systems, wireless medium access control, multimedia communications, network economics, and applications of optimization theory and game theory. Dr. Huang is an Associate Editor of JOURNAL OF COMPUTER & ELECTRICAL ENGINEERING, a Guest Editor for IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS and JOURNAL OF ADVANCES IN MULTIMEDIA, and a TPC Co-Chair of the International Conference on Game Theory for Networks (GameNets'09).

Vijay G. Subramanian (M'01) received his Ph.D. degree in Electrical Engg. from the University of Illinois at Urbana-Champaign, in 1999. From 1999 to 2006, he was with the Networks Business, Motorola, Arlington Heights, IL, USA. Since May 2006 he is a Research Fellow at the Hamilton Institute, NUIM, Ireland. His research interests include information theory, communication networks, queueing theory, and applied probability and stochastic processes.

Rajeev Agrawal is a Fellow of the Technical Staff at Motorola where his responsibilities include the architecture, design and optimization of Motorola's next generation wireless systems. Prior to joining Motorola in 1999, Rajeev was Professor of Electrical and Computer Engg. and Computer Science departments at the University of Wisconsin - Madison. He also spent a sabbatical year at IBM TJ Watson Research, British Telecom Labs, and INRIA-Sophia Antipolis. Rajeev received his M.S. (1987) and Ph.D. (1988) degrees in Electrical Engg.-systems from the University of Michigan, Ann Arbor and his B.Tech. (1985) degree in Electrical Engg. from the Indian Institute of Technology, Kanpur.

Randall A. Berry (S'93-M'00) received the M.S. and PhD degrees in Electrical Engg. and Computer Science from the Massachusetts Institute of Technology in 1996 and 2000, respectively. In September 2000, he joined the faculty of Northwestern University, where he is currently an Associate Professor of Electrical Engg. and Computer Science. Dr. Berry is the recipient of a 2003 NSF CAREER award. He is currently serving on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.