Downlink Scheduling and Resource Allocation for OFDM Systems

Jianwei Huang
Dept. of EE
Princeton University
Princeton, NJ, 08544

Vijay Subramanian and Rajeev Agrawal
Mathematics of Communication Networks
Motorola Inc.
Arlington Heights, IL 60004

Randall Berry
Dept. of EECS
Northwestern University
Evanston IL 60208

***Abstract*— We consider scheduling and resource allocation for the downlink of a OFDM-based wireless network. During each time-slot the scheduling and resource allocation problem involves selecting a subset of users for transmission, determining the assignment of available subcarriers to selected users, and for each subcarrier determining the transmission power and the coding and modulation scheme used. We address this in the context of a utility-based scheduling and resource allocation scheme presented in earlier papers. Scheduling and resource allocation is determined by solving an optimization problem, which is convex for a reasonable model of the feasible rates. By exploiting the structure of this problem, we give optimal and sub-optimal algorithms for its solution. We provide simulation results comparing different algorithms and parameter settings.**

I. Introduction

Channel-aware scheduling and resource allocation has become an essential component for high-speed wireless data systems. In these systems, the active users and the allocation of physical layer resources among them are dynamically adapted based on the users' current 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 gradient of the system's total utility [2]. The utility is a function of each user's throughput and is used to quantify fairness and other QoS considerations. Several such gradient-based policies have been studied for time-division multiplexed (TDM) systems, such as the the "proportional fair rule" [3]–[5], first proposed for CDMA 1xEVDO, which is based on a logarithmic utility function. In [6], a larger class of utility functions is considered that allows efficiency and fairness to be traded-off. Generalized $c\mu$-policies [7]–[9], such as a Max Weight policy [10]–[12], [14], can also be viewed as a type of gradient-based policy, where the utility is a function of a user's queue-size or delay.

In TDM systems, the scheduling and resource allocation decision is simply which user to schedule in a time-slot and what modulation and coding scheme to use for the scheduled user. In many recent systems, users may be further multiplexed within a time-slot using another technique. In particular, Orthogonal Frequency Division Multiplexing (OFDM) is seen as a promising option for broadband wireless networks due in part to its ability for combating intersymbol interference in frequency selective channels and for avoiding interference among scheduled users by assigning different subcarriers to different users.[1] OFDM is utilized in a number of emerging wireless data systems such as IEEE 802.16 (WiMAX). This paper addresses gradient-based scheduling and resource allocation for the downlink in a single OFDM cell. In this setting, in addition to determining which users are scheduled, the allocation of physical layer resources including the transmission power and the assignment of tones to users must be specified.

In prior work [13], we considered gradient-based scheduling and resource allocation when code division multiple access (CDMA) was used to multiplex users within a time-slot, as in CDMA 1xEVDV. In the CDMA case, the physical layer resources are the number of spreading codes assigned to each user and the transmission power; allocating these according to a gradient-based policy requires maximizing the weighted sum rate across all users in each time-slot, where the weights dynamically vary based on the gradient of the system utility. When the users' SINR and rate per code are related via the Shannon capacity formula, the resulting problem is a tractable convex optimization problem, enabling the development of low complexity near-optimal algorithms and the characterization of key properties of the solution. Here, we follow a similar approach for an OFDM-based system; the main difference being that we have more degrees of freedom (i.e., the subcarriers) to allocate resources across, which increases the complexity of the optimization problem. We also allow for different subchannelization techniques, in which the resource allocation must be specified in terms of groups of subcarriers or symbols. Such techniques allow for trading off the granularity of resource allocation with the overhead requirements for channel measurement and feedback.

As in [13], within each time-slot the gradient-based scheduling policy requires maximizing the weighted sum throughput over the set of feasible rates. Here, the set of feasible rates depends on what subchannelization is used, the current channel state information, and the resource allocation decisions. We also allow constraints on the maximum SINR or rate per subcarrier, which can model a limitation on the available

Part of this work was done while J. Huang was a summer intern at Motorola. R. Berry was supported in part by the Motorola-Northwestern Center for Telecommunications and NSF CAREER award CCR-0238382.

[1] In the following we use the terms tone and subcarrier synonymously.

modulation order. When the rate per sub-carrier is given via the Shannon capacity formula and users are allowed to time-share each sub-carrier, again this becomes a tractable convex optimization problem.[2] A special case of this problem for a fixed set of weights and no constraints on the SINR per carrier was considered in [15]; there a suboptimal algorithm with constant power per sub-carrier was given and shown in simulations to have little performance loss. Here, we first consider a dual formulation for this problem, which enables us characterize some structural properties and leads to both optimal and reduced complexity sub-optimal algorithms. We also present simulation results in a system where the scheduling weights are dynamically adjusted according to a gradient-based scheduling rule.

In related work, in addition to [15], a number of other formulations for downlink OFDM resource allocation have been studied including [16]–[21]. In [16], [19] the goal is to minimize the total transmit power given target bit rates for each user. In [19], the target bit-rates are determined by a fair queueing algorithm, which does not take into account the users' channel conditions. In [18], [20], [21], the focus is on maximizing the sum-rate given a required minimum bit rate per user; [17] also considers maximizing the sum-rate, without any minimum bit-rate target. [17], [18], [20] also consider suboptimal heuristics that use a constant power per sub-carrier. Finally, in [22], the capacity region of a downlink broadcast channel with frequency-selective fading using a TDM scheme is given; the feasible rate region we consider, without any maximum SINR constraints, can be viewed as a special case of this region.

II. Problem Formulation

We consider the downlink of a single cell in an OFDM system with K users. Time is divided into TDM time-slots that contain an integer number of OFDM symbols. 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 $\mathcal{R}(\boldsymbol{e}_t) \subset \mathbb{R}_+^K$, where $\boldsymbol{e}_t$ indicates the time-varying channel state information available at the scheduler. This decision is made according to the gradient-based scheduling framework in [2], [6]. In this framework, $\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)$ is selected that has the maximum projection onto the gradient of a system utility function $\nabla U(\boldsymbol{W}_t)$, where

$$U(\boldsymbol{W}_t) = \sum_{i=1}^{K} U_i(W_{i,t}),$$

and $U_i(W_{i,t})$ 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

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \nabla U(\boldsymbol{W}_t)^T \cdot \boldsymbol{r}_t = \max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_i \dot{U}_i(W_{i,t}) r_{i,t}. \tag{1}$$

For example, one class of utility functions given in [6] is

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

where $\alpha \leq 1$ is a fairness parameter and c_i is a QoS weight. In this case, (1) becomes

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_i c_i (W_{i,t})^{\alpha-1} r_{i,t}. \tag{3}$$

With equal class weights, setting $\alpha = 1$ results in a scheduling rule that maximizes the total throughput during each slot. For $\alpha = 0$, this results in the proportional fair rule.

More generally, the utility can depend on other parameters for each user such as the queue size or the delay of the head-of-line packet, as in the "Max Weight" policies mentioned in the introduction. In general, we consider the problem of

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_i w_{i,t} r_{i,t}, \tag{4}$$

where $w_{i,t} \geq 0$ is a time-varying weight assigned to the ith user at time t. In the above examples these weights are given by the gradient of the utility; however, other methods for generating these weights are also possible. We note that (4) must be re-solved at each scheduling instant because of changes in both the channel state and the weights (e.g., the gradient of the utility).

A. OFDM capacity regions

Solving (4) depends on the state dependent capacity region $\mathcal{R}(\boldsymbol{e})$.[3] We focus on a model appropriate for downlink OFDM systems; similar models have been considered in [15], [22]. In this model, $\mathcal{R}(\boldsymbol{e})$ is parameterized by the allocation of subcarriers to users and the allocation of power across sub-carriers. In a traditional OFDM system, at most one user may be assigned to any subcarrier. Here, as in [16], [19], we make the simplifying assumption that multiple users can share one tone using some orthogonalization technique (e.g. via TDM). In practice, if a scheduling interval contained multiple OFDM symbols, we could implement such time-sharing by giving a fraction of the symbols to each user; of course, each user would be constrained to use an integer number of symbols and the required signaling overhead would increase. Given a solution to this problem, we can obtain a feasible solution allowing only one user per tone by applying an appropriate projection. For the simulations in Section IV, we choose only one user per tone.

Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of subcarriers. For each subcarrier j and user i, let e_{ij} denote the received signal to interference plus noise ratio (SINR) per unit power. We denote the power allocated to user i on subchannel j by p_{ij} and the fraction of that subchannel allocated to user i by x_{ij}. These must satisfy a total power constraint, $\sum_{i,j} p_{ij} \leq P$, and for all subcarriers j, $\sum_i x_{ij} \leq 1$, i.e., the total fraction of each sub-carrier allocated must be no greater than one. For a given allocation, user i's feasible rate on subcarrier j is given by

[2] We focus on systems that do not use superposition coding and successive interference cancellation within a sub-carrier. While such techniques are necessary for achieving the multiuser capacity of a broadcast channel [22], they are generally considered too complex for practical systems.

[3] To simplify notation we have suppressed the time-dependence.

$r_{ij} = x_{ij} B \log(1 + \frac{p_{ij}e_{ij}}{x_{ij}})$. This corresponds to the Shannon capacity of a Gaussian noise channel with bandwidth $x_{ij}B$ and received SINR $\frac{p_{ij}e_{ij}}{x_{ij}}$.[4] Without loss of generality we set $B = 1$. The achievable rate region is then

$$\begin{aligned}\mathcal{R}(\boldsymbol{e}) = \Big\{ \boldsymbol{r} :\ r_i = \sum_j x_{ij} \log\left(1 + \frac{p_{ij}e_{ij}}{x_{ij}}\right), \\ \sum_{ij} p_{ij} \le P,\ \sum_i x_{ij} \le 1\ \forall\, j,\ (\boldsymbol{x},\boldsymbol{p}) \in \mathcal{X} \Big\},\end{aligned} \tag{5}$$

where[5]

$$\mathcal{X} := \left\{ (\boldsymbol{x},\boldsymbol{p}) \ge \boldsymbol{0} : 0 \le x_{ij} \le 1, 0 \le p_{ij} \le \frac{x_{ij}s_{ij}}{e_{ij}}\ \forall i,j \right\}. \tag{6}$$

Here we have also included a maximum SINR constraint of s_{ij} on tone j for user i. For example, this can model a constraint on the maximum rate per tone due to a limitation on the available modulation order.

In the above model, the channel state $\boldsymbol{e} = \{e_{ij}\}$, which we assume is known by the scheduler for all users and tones. In a frequency division duplex (FDD) system, this knowledge can be acquired by having the base station transmit pilot signals, from which the users estimate their channel gains and feed this 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 would need to be done within the channel's coherence time.

In a system with many tones and users, providing pilots and/or feedback per tone can require excessive overhead. One approach for reducing this overhead is by forming *subchannels* that are disjoint sets of tones. Feedback and resource allocation is then done at the granularity of these subchannels. The above model can be adapted to this setting, by viewing $\mathcal{N}$ as the set of subchannels and e_{ij} as the effective SINR per unit power for user i within the jth subchannel. In other words, assuming that k tones are bundled into subchannel j, e_{ij} is chosen so that the rate for user i in this subchannel is approximately $kx_{ij}\log(1 + \frac{p_{ij}e_{ij}}{x_{ij}})$. For our simulations, we set e_{ij} to be the average of the SINR per unit power of all the tones in a subchannel.[6]

Subchannels can be formed in various ways; in our simulations, we consider the following three approaches: (1) adjacent channelization, where adjacent tones are grouped into a sub-channel; (2) interleaved channelization, where tones are interleaved to form subchannels; and (3) random channelization, where tones are randomly assigned to subchannels. In IEEE 802.16d/e, interleaved channelization is primarily used; the optional "band AMC mode" allows for adjacent channelization. Randomized channelization can model systems that employ frequency hopping as in Flarion's Flash OFDM system. Under adjacent channelization, if the adjacent channels lie approximately within a coherence band, then this enables the resource allocation to better exploit frequency diversity. Using interleaved or random channelization reduces the variance in the channel gains across subchannels for each user. One advantage to this is that when the variance is small, the user can simply feedback a single e_{ij} value that will be representative of each subchannel, further reducing the required overhead.[7] Another advantage of random channelization is in managing other cell interference.

III. Optimal and Suboptimal Algorithms

From (4) and (5), the scheduling and resource allocation problem can be stated as:

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

where $\mathcal{X}$ is given in (6).

A. Optimal Dual Solution

We first solve this problem using duality methods. It can be shown that (7) is convex and Slater's condition holds, so there is no duality gap and the optimal solution is characterized by the Karush-Khun-Tucker conditions [1].

Consider the Lagrangian given by

$$\begin{aligned}L(\boldsymbol{x},\boldsymbol{p},\lambda,\boldsymbol{\mu}) := \sum_i w_i \sum_j x_{ij} \log\left(1 + \frac{p_{ij}e_{ij}}{x_{ij}}\right) \\ + \lambda\Big(P - \sum_{i,j} p_{ij}\Big) + \sum_j \mu_j \Big(1 - \sum_i x_{ij}\Big).\end{aligned}$$

Optimizing over $\boldsymbol{p}$ given $\boldsymbol{x}$, $\boldsymbol{\mu}$ and λ yields[8]

$$p^*_{ij} = \frac{x_{ij}}{e_{ij}} \left[\left(\frac{w_i e_{ij}}{\lambda} - 1 \right)^+ \wedge s_{ij} \right]. \tag{8}$$

Substituting this into $L(\boldsymbol{x},\boldsymbol{p},\lambda,\boldsymbol{\mu})$, we have

$$\begin{aligned}&L(\boldsymbol{x},\boldsymbol{p}^*,\lambda,\boldsymbol{\mu}) \\ &= \sum_{ij} x_{ij}\left(w_i h\left(\lambda, w_i e_{ij}, s_{ij}\right) - \mu_j\right) + \sum_j \mu_j + \lambda P,\end{aligned}$$

Here, as in [13],

$$h(x,y,z) = \begin{cases} 0, & x \ge y, \\ \frac{x}{y} - 1 - \log\frac{x}{y}, & \frac{y}{1+z} \le x < y, \\ \log(1+z) - \frac{x}{y}z, & x < \frac{y}{1+z}, \end{cases} \tag{9}$$

[4] As in [13], to better model the achievable rates in a practical system we can re-normalize e_{ij} by γe_{ij}, where $\gamma \in [0, 1]$ represents the system's "gap" from capacity.

[5] Here and in the following we use boldfaced symbols to denote the vector of all the corresponding values across users/tones, e.g. $\boldsymbol{x}$ is the vector of all x_{ij}'s.

[6] Using the concavity of the logarithm, it can be seen that this upper-bounds the average rate that can be achieved in a subchannel. The average rate can be lower bounded by using the geometric mean of the SINRs of the tones in a subchannel.

[7] When every channel is identical, the resource allocation problem becomes equivalent to the CDMA problem in [13].

[8] The notation $(x)^+ = \max(x, 0)$ and $x \wedge y = \min(x, y)$.

where $x \geq 0$, $y > 0$ and $z \geq 0$.

Optimizing $L(\boldsymbol{x}, \boldsymbol{p}^*, \lambda, \boldsymbol{\mu})$ over $\boldsymbol{x}$ we get the corresponding dual function

$$\begin{aligned} L(\lambda, \boldsymbol{\mu}) &:= L(\boldsymbol{x}^*, \boldsymbol{p}^*, \lambda, \boldsymbol{\mu}) \\ &= \sum_{ij} \left(w_i h\left(\lambda, w_i e_{ij}, s_{ij}\right) - \mu_j\right)^+ + \sum_j \mu_j + \lambda P. \end{aligned}$$

Since there is no duality gap, it follows that minimizing this over λ and $\boldsymbol{\mu}$ yields an optimal solution to (7). We follow a similar procedure as in [13] to accomplish this. First considering the optimal $\boldsymbol{\mu}$, we have:

Lemma 3.1: For all $\lambda \geq 0$,

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

where for all j, the minimizing value of μ_j is

$$\mu_j^*(\lambda) = \max_i w_i h\left(\lambda, w_i e_{ij}, s_{ij}\right). \tag{10}$$

The proof of this follows from a similar argument as in [13], and by noting that $L(\lambda, \boldsymbol{\mu})$ is a separable function of $\boldsymbol{\mu}$. Note that (10) requires a sort of all the users according to the metrics $\mu_{ij} := w_i h(\lambda, w_i e_{ij}, s_{ij})$ for each sub-channel j.

As in [13], $L(\lambda)$ can be shown to be a convex function of λ; hence it can be minimized using an iterated one dimensional search, like the Golden Section method. At the minimizing value λ^*, $L(\lambda^*)$ gives the optimal solution to (7).

B. Optimal primal variables with time-sharing

Next we turn to finding optimal values of the primal variables $(\boldsymbol{x}, \boldsymbol{p})$. For a given $\lambda \geq 0$, let

$$(\boldsymbol{x}^*, \boldsymbol{p}^*) := \arg \max_{(\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X}} L\left(\boldsymbol{x}, \boldsymbol{p}, \lambda, \boldsymbol{\mu}^*(\lambda)\right), \tag{11}$$

which can be solved using the same procedure as in deriving the dual function. Given that $\lambda = \lambda^*$, it follows from duality theory, that if the corresponding $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ are primal feasible and satisfy complimentary slackness, then they are optimal primal values. However, in (10) there can be multiple users in a given subchannel whose metrics μ_{ij} are tied at the maximum value. In this case, it can be shown that there will be multiple primal values that satisfy (11), not all of which may be feasible. Thus, breaking these ties to settle on a specific primal solution is necessary to find the optimal solution. A key point to note is that when ties occur at a given λ, $L(\lambda)$ is not differentiable at that λ. However, since $L(\lambda)$ is a convex function, subgradients exists.

Definition 3.1: A scalar $d \in \Re$ is a *subgradient* of $L(\lambda)$ at λ if

$$L(\tilde{\lambda}) \geq L(\lambda) + \left(\tilde{\lambda} - \lambda\right) d, \quad \forall \tilde{\lambda} \geq 0.$$

For an arbitrary λ, a solution to (11) that also satisfies $\sum_j \mu_j^*(\lambda) \left(1 - \sum_i x_{ij}\right) = 0$ and $\sum_i x_{ij} \leq 1$ for all j, can be used to find a sub-gradient of $L(\lambda)$.

Proposition 3.1: Let $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{p}})$ be a solution to (11) for a given λ that satisfies $\sum_i \hat{x}_{ij} \leq 1$ for all j and $\sum_j \mu_j^*(\lambda) \left(1 - \sum_i \hat{x}_{ij}\right) = 0$. Then $P - \sum_{ij} \hat{p}_{ij}$ is a subgradient of $L(\lambda)$ at λ.

For a given λ, let $\mathcal{A}_j := \{i : i = \arg\max_i \mu_{ij}(\lambda)\}$. From the previous analysis it follows that the set of all $\boldsymbol{x}$ that solve (11) 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 $i \in \mathcal{A}_j$.
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*), we must break ties to determine the allocation for that sub-carrier. We refer to an allocation satisfying (*i*)-(*iii*) as an *extreme point* if additionally $x_{ij} \in \{0, 1\}$ for all i and j; such an allocation can be represented by a function $f : \mathcal{N} \rightarrow \mathcal{M}$, where $\mathcal{M}$ is the set of users, so that $f(j)$ indicates the user allocated channel j, i.e. $x_{f(j)j} = 1$. To satisfy (*i*)-(*iii*), it must be that $f(j) \in \mathcal{A}_j$ for all j. 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 $f(j)$ can take on only one value. For each subchannel $j \in \mathcal{B}^c$, there are $|\mathcal{A}_j|$ users we can allocate the subchannel to, and so $f(j)$ can take $|\mathcal{A}_j|$ values. It follows that the total number of extreme points is $\prod_{j \in \mathcal{B}^c} |\mathcal{A}_j|$.

Each extreme point satisfies Proposition 3.1 and so provides a subgradient for $L(\lambda)$. Let

$$\tilde{p}_{ij} := \left[\left(\frac{w_i e_{ij}}{\lambda} - 1\right)^+ \wedge s_{ij}\right] \frac{1}{e_{ij}}.$$

Given an extreme point f, from (8), the resulting power allocation to subchannel j is given by $p_{ij} = \tilde{p}_{ij}$ for $i = f(j)$ and $p_{ij} = 0$ for $i \neq f(j)$. Hence the corresponding subgradient $d(f)$ is given by

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

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

$$f(j) = \arg \min_{i \in \mathcal{A}(j)} \tilde{p}_{ij}, \ \forall j. \tag{12}$$

Likewise, the minimum subgradient is given by

$$f(j) = \arg \max_{i \in \mathcal{A}(j)} \tilde{p}_{ij}, \ \forall j. \tag{13}$$

Lemma 3.2: There exists a primal optimal solution where $\boldsymbol{x}$ is given by time-sharing between the extreme points in (12) and (13) with $\lambda = \lambda^*$, and $\boldsymbol{p}$ is given by (8).

Note that this lemma implies that there is always an optimal primal solution for which at most two users time-share any tone. Moreover, each tone that is time-shared is shared in the same proportion. The optimal time-sharing factor can be found by simply finding a convex combination of the subgradients corresponding to (12) and (13) that is equal to zero.

The above steps give us an algorithm for finding the optimal solution to (7). Namely, first use a one-dimension search to find the optimal λ^* that minimizes $L(\lambda)$. Next find the corresponding optimal primal solution as in Lemma 3.2.

C. Single user per tone

Next we consider the case where we restrict our final allocation to allow only one user to transmit on each subchannel. In this case, we can still find the optimal λ^*. If there are no ties as discussed above, then the optimal solution will only allow one user/tone. If there are ties, then a reasonable heuristic is to simply choose one extreme point allocation. In our simulations we choose the extreme point that corresponds to the subgradient with the smallest non-negative value; i.e. the extreme point f, for which $\sum_{jin\mathcal{N}} \tilde{p}_{f(j)j}$ is closest to P, without exceeding it. Other mechanisms for choosing a extreme point could also be used.

For a given extreme point, the total power constraint using the powers $\tilde{p}_{f(j)j}$ will be over-shot or under-shot (unless this point is optimal). In this case we consider re-optimizing the power allocation for a given fixed tone allocation $\boldsymbol{x}$, i.e. we want to solve

$$\max_{\boldsymbol{p}:(\boldsymbol{p},\boldsymbol{x})\in\mathcal{X}} V(\boldsymbol{n},\boldsymbol{p}) \quad \text{s.t.} \sum_{ij} p_{ij} \le P \tag{14}$$

Let $L_{\boldsymbol{x}}(\lambda)$ be the dual function for this problem. Given that $\tilde{\lambda} = \arg\min_{\lambda\ge 0} L_{\boldsymbol{x}}(\lambda)$, the optimal power allocation to (14) is given by (8) with $\lambda = \tilde{\lambda}$ and the given tone allocation. The following lemma gives an alternative characterization for the correct λ.

Lemma 3.3: A given λ is the solution to the dual problem $\min_{\lambda\ge 0} L_{\boldsymbol{x}}(\lambda)$ if and only if

$$\lambda = \frac{\sum_{i,j} x_{ij} w_i 1_{\mathcal{W}_{ij}}(\lambda)}{P - \sum_{i,j} \frac{s_{ij}}{e_{ij}} 1_{\mathcal{Y}_{ij}}(\lambda) + \sum_{i,j} \frac{1}{e_{ij}} 1_{\mathcal{W}_{ij}}(\lambda)}, \tag{15}$$

where $\mathcal{W}_{ij} = \left[\frac{x_{ij}w_i e_{ij}}{1+s_{ij}}, x_{ij}w_i e_{ij}\right)$, $\mathcal{Y}_{ij} = \left[0, \frac{x_{ij}w_i e_{ij}}{1+s_{ij}}\right)$, and $1_{\{\cdot\}}$ is the indicator function.

This lemma provides an algorithm for finding $\tilde{\lambda}$, which we describe next. First check if the power constraint is violated when all users use maximum power on the allocated tones, i.e. if $\sum_{(i,j)} p^*_{ij} = \frac{x_{ij}}{e_{ij}} s_{ij} > P$. If this is not true, the problem is solved. If this is true, we need to search for $\tilde{\lambda}$.

Let $\boldsymbol{a}$ be a vector of length $2N$ containing the values of $x_{ij}w_i e_{ij}$ and $\frac{x_{ij}w_i e_{ij}}{1+s_{ij}}$ for all (i,j) such that $x_{ij} = 1$, sorted in descending order. Define two additional vectors $\boldsymbol{z}$ and $\boldsymbol{y}$ such that for any $k = 1, ..., 2N$,

$$z(k) = 1, \text{ if } a(k) = \frac{x_{ij}w_i e_{ij}}{1+s_{ij}} \text{ for some } (i,j),$$
$$y(k) = \{i,j\}, \text{ if } a(k) = x_{ij}w_i e_{ij} \text{ or } \frac{x_{ij}w_i e_{ij}}{1+s_{ij}}.$$

The complete λ search algorithm is given in Algorithm 1. The basic idea is to start from the largest λ, and calculate the right-hand side of (15). If it is less than the current value of λ, decrease λ and recalculate, until a fixed-point is found. It can be shown that the algorithm will stop in at most $2N$ steps with $\lambda(k) = \tilde{\lambda}$.

Algorithm 1 Search Algorithm for Optimal λ

```
1) Initialization: $k = 0$, $G_{xw} = 0$, $G_{s/e} = 0$ and $G_{1/e} = 0$.
2) $k = k + 1$.
3) Let $\{i(k), j(k)\} = y(k)$.
4) if $z(k) = 0$, then
       $G_{xw} = G_{xw} + x_{i(k)j(k)} w_{i(k)}$,
       $G_{1/e} = G_{1/e} + \frac{1}{e_{i(k)j(k)}}$,
   otherwise,
       $G_{s/e} = G_{s/e} + \frac{s_{i(k)j(k)}}{e_{i(k)j(k)}}$,
       $G_{xw} = G_{xw} - x_{i(k)j(k)} w_{i(k)}$,
       $G_{1/e} = G_{1/e} - \frac{1}{e_{i(k)j(k)}}$.
5) Let $\lambda(k) = G_{xw} / \left(P - G_{s/e} + G_{1/e}\right)$. If $k = 2N$,
   stop. Otherwise, go to step (6).
6) If $\lambda(k) \le a(k)$ and $\lambda(k) \ge a(k+1)$, stop. Otherwise,
   go to step (2).
```

D. Single sort suboptimal algorithm

The optimal sub-carrier allocation is determined by assigning each tone j to the user with the largest metric $\mu_j^*(\lambda^*)$ on that tone (breaking any ties as discussed above). This requires iterating to find the optimal Lagrange multiplier λ^*. We give a sub-optimal algorithm that is based instead doing a single sort of the users on each tone according to a different metric. Here, we consider using the metric $w_i \bar{R}_{ij}$, where $\bar{R}_{ij}$ is the rate that user i could achieve on this channel under a constant power allocation, i.e.,

$$\bar{R}_{ij} = \log[1 + (s_{ij} \wedge (e_{ij}P/N))].$$

The tone is allocated to the user with the largest metric, with ties broken arbitrarily. After the tone allocation is made, the optimal power allocation is done as in the optimal algorithm. This metric was motivated in part the prior work in [15], [17], [18], [20] where a uniform power allocation was shown to be nearly optimal. Some other suboptimal algorithms are discussed in the Appendix.

IV. Simulation Study

In this section we report some simulation results for the algorithm which finds the optimal λ^* and then chooses a tone-allocation with one user per tone as described in Section III-C. We also consider the sub-optimal algorithm described in Section III-D. We simulate a single cell with $M = 40$ users. The channel gains e_{ij} are the product of a fixed location-based term for each user i and a frequency-selective fast fading term. The location-based components were picked using an empirically obtained distribution for many users in a large system. The fast-fading term was generated using a block-fading model based upon the Doppler frequency (for the block-length in time) and a standard reference mobile delay-spread

TABLE I

SIMULATION RESULTS FOR DIFFERENT CHOICES OF α (64 SUBCHANNELS, ADJACENT SUBCHANNELIZATION, NO PER USER SINR CONSTRAINTS).

α	Algorithm	Utility	Log U	Rate(kbps)	Num.
0.5	FULL	1236	12.58	497.8	5.40
0.5	MO-wR	1234	12.56	498.3	5.17
0	FULL	12.69	12.69	396.8	5.75
0	MO-wR	12.68	12.68	393.0	5.47
1	FULL	716955	8.04	719.3	3.04
1	MO-wR	716955	8.04	719.3	3.04

model (for variation in frequency). For a user's fast-fading term, each multi-path component was held fixed for $2msec$ and an independent value was generated for the next block, corresponding to a 250MHz Doppler frequency. The delay-spread was 1μsec. The user's channel conditions averaged over the applicable channelization scheme are fed back to the scheduler for all the channels.

We considered a system bandwidth of 5MHz corresponding to 512 OFDM carriers/tones. The symbol duration was 100μsec with a cyclic prefix of 10μsec. This roughly corresponds to 20 OFDM symbols per fading block. The resource allocation 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), at this time the system has reached a stationary operation point. All users were infinitely back-logged. We assigned each user a throughput-based utility with the form given in (2); for a given simulation all the users have identical QoS weights $(c_i = 1)$ and fairness parameters (α).

The first set of simulations we consider are for a system where the tones are grouped into 64 subchannels with adjacent subchannelization, i.e. adjacent sets of 8 tones are grouped into subchannels. We initially assume that there are no per user SINR constraints (i.e., $s_{ij} = \infty$). Table I shows results for both the algorithm with the optimal λ^* (FULL) and the suboptimal algorithm (MO-$w\bar{R}$) for different choices of the utility parameter α. The column "utility" gives the average utility per user for each algorithm. The column labeled "log U" shows the log utility per user; this gives some indication of the "fairness" of the resulting allocation (for $\alpha = 0$ this is the same as the utility.). The column labeled "Rate" is the average throughput per user, and the final column is the average number of users scheduled. We note that for each choice of α, the two algorithms perform nearly the same for each of these metrics; when $\alpha = 1$ (maximum throughput) they have identical performance. In Figure 1, we plot the throughput CDF for both algorithms, as well as two other sub-optimal algorithms for $\alpha = 0.5$. Both FULL and MO-$w\bar{R}$ have nearly identical CDF's; under the other algorithms the CDF is more spread out, indicating a higher variance in the users' rates.

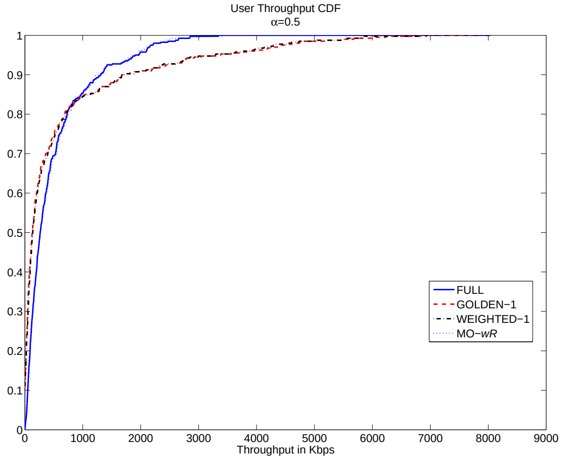

Fig. 1. Empirical CDF of users' throughputs for $\alpha = 0.5$.

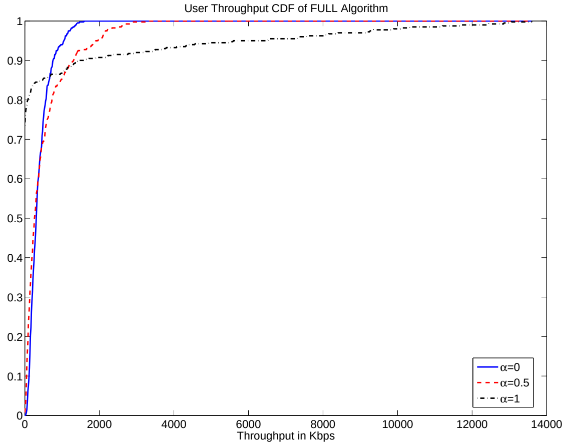

Fig. 2. Empirical CDF of users' throughputs for different values of α.

In Figure 2, concentrating on the FULL algorithm, we compare the effects of different α's on the throughput CDF. Since an α close to 1 emphasizes total throughput more than fairness, the distributions get more spread out as α increases.

Next we consider the effect of different channelization schemes. Table II shows the performance of the two algorithms for the adjacent (Adj.), randomized (Ran.), and Interleaved (Int.) channelization schemes described in Section II-A. The parameters here are the same as in Table I, with $\alpha = 0.5$. Again, MO-$w\bar{R}$ performs nearly the same as FULL and in the interleaved case even achieves a slightly higher utility. For both algorithms, the random channelization results in lower utility than the adjacent, and the interleaved results in yet lower utility. This is likely due to the decreased frequency diversity with each scheme. Indeed, for the channel model used here, in the interleaved case all subchannels can be shown to be identical, which explains why both schemes only schedule one user.

Finally, we consider the effect of varying the number of tones per subchannel and the effect of a per user SINR

TABLE II

SIMULATION RESULTS OF DIFFERENT CHANNELIZATION SCHEMES (64 SUBCHANNELS, NO PER USER SINR CONSTRAINTS, $\alpha = 0.5$).

Chan.	Algorithm	Utility	Log U	Rate (kbps)	Num.
Adj.	FULL	1236	12.58	497.8	5.40
Adj.	MO-$w\bar{R}$	1234	12.56	498.3	5.17
Ran.	FULL	1171	12.42	465.2	4.08
Ran.	MO-$w\bar{R}$	1167	12.40	465.5	3.64
Int.	FULL	1136	12.32	447.1	1
Int.	MO-$w\bar{R}$	1142	12.33	455.2	1

TABLE III

SIMULATION RESULTS ($\alpha = 0.5$, 32 SUBCHANNELS, ADJACENT SUBCHANNELIZATION, NO PER USER SINR CONSTRAINTS).

Algorithm	Utility	Log U	Rate (kbps)	Num
FULL	1234	12.57	496.6	5.22
MO-$w\bar{R}$	1232	12.56	497.2	5.02

constraint. Table III shows the case where the OFDM tones are grouped into 32 subchannels instead of 64 as in Table I (i.e. 16 tones/subchannel). Comparing to the 64 subchannel case, both algorithms achieve slightly less utility; again this can be explained by a slight decrease in the frequency diversity. As long as there is enough diversity, our simulations suggest that the overall performance is not very sensitive to the number of subchannels. Table IV shows the performance of the algorithms when each user can only transmit at a maximum SINR of 6.5dB on each subchannel.[9] Here, the performance gaps between the FULL algorithm and MO-$w\bar{R}$ slightly increases compared with Table I (MO-$w\bar{R}$ achieves 96.6% of the maximum utility, as opposed to 99.8% as in Table I).

V. CONCLUSIONS

We considered scheduling and resource allocation for the downlink of OFDM systems. Using a gradient-based scheduling framework, we formulated an optimal scheduling and resource allocation problem, which was shown to be a convex problem. Using a dual formulation, we characterized the optimal solution, and used this to develop optimal and sub-optimal algorithms. The algorithms can be applied across different channelization schemes and accommodate per user SINR constraints. We presented simulation results showing that a sub-optimal algorithm, in which users are sorted once per tone based on a uniform power allocation, performs nearly the same as under the optimal algorithm.

[9] This choice is motivated by the IEEE 802.16e standard, in which the maximum rate/tone is achieved with 64-QAM modulation and 5/6 code rate.

TABLE IV

SIMULATION RESULTS ($\alpha = 0.5$, 64 SUBCHANNELS, ADJACENT SUBCHANNELIZATION, MAXIMUM SINR EQUAL TO 6.5DB).

Algorithm	Utility	Log Utility	Rate (kbps)	User Scheduled
FULL	1137	12.60	349.8	5.94
MO-$w\bar{R}$	1098	12.50	333.3	5.21

REFERENCES

[1] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, Massachusetts: Athena Scientific, 1999.

[2] R. Agrawal and V. Subramanian, “Optimality of Certain Channel Aware Scheduling Policies,” *Proc. of 2002 Allerton Conference on Communication, Control and Computing*, Oct. 2002.

[3] A. Jalali, R. Padovani, and R. Pankaj, “Data throughput of CDMA-HDR a high efficiency - high data rate personal communication wireless system.,” in *Proc. VTC ’2000*, Spring, 2000.

[4] P. Viswanath, D. Tse, and R. Laroia, “Opportunistic Beam-forming using Dumb Antennas,” *IEEE Trans. on Information Theory*, vol. 48, June 2002.

[5] H. Kushner and P. Whiting, “Asymptotic Properties of Proportional-Fair Sharing Algorithms,” *Proc. of 2002 Allerton Conference on Communication, Control and Computing*, Oct. 2002.

[6] R. Agrawal, A. Bedekar, R. La, and V. Subramanian, “A Class and Channel-Condition based Weighted Proportionally Fair Scheduler,” *Proc. of ITC 2001*, Salvador, Brazil, Sept. 2001.

[7] J. A. Van Mieghem, “Dynamic Scheduling with Convex Delay Costs: the Generalized $c\mu$ Rule,” *Annals of Applied Probability*, 5(3), 1995.

[8] A. Mandelbaum and A. L. Stoylar, “$Gc\mu$ Scheduling of Flexible Servers: Asymptotic Optimality in Heavy Traffic,” *Proc. of 2002 Allerton Conference on Communication, Control and Computing*, Oct. 2002.

[9] P. Liu, R. Berry, and M. Honig, “A Fluid Analysis of a Utility-Based Wireless Scheduling Policy,” to appear, IEEE Trans. on Information Theory.

[10] A. L. Stolyar, “MaxWeight scheduling in a generalized switch: state space collapse and equivalent workload minimization in Heavy Traffic,” submitted, 2001.

[11] L. Tassiulas and A. Ephremides, “Dynamic server allocation to parallel queue with randomly varying connectivity”, in *IEEE Transactions on Information Theory*, Vol. 39, pp. 466-478, March 1993.

[12] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar and P. Whiting, “Providing Quality of Service over a Shared Wireless Link”, *IEEE Communications Magazine*, Vol. 39, No. 2, pp. 150-154,2001.

[13] R. Agrawal, V. Subramanian and R. Berry, ”Joint Scheduling and Resource Allocation in CDMA Systems,” *Proc. of 2nd Workshop on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt ’04)*, Cambridge, UK, March 24-26, 2004.

[14] R. Leelahakriengkrai and R. Agrawal, “Scheduling in Multimedia Wireless Networks,” *17th International Teletraffic Congress*, Salvador da Bahia, Brazil, Dec. 2-7, 2001.

[15] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, “Multiuser Transmit Optimization for Multicarrier Broadcast Channels: Asymptotic FDMA Capacity Region and Algorithms,” *IEEE Trans. on Communications*, vol. 52, no. 6, pp. 922-930, June 2004.

[16] C. Y. Wong, R. S. Cheng, K. B. Letaief and R. D. Murch, “Multiuser OFDM with Adaptive Subcarrier, Bit and Power Allocation,” *IEEE Journal on Selected Areas in Communications* vol. 17, no. 10, Oct. 1999.

[17] J. Jang and K. B. Lee, “Transmit Power Adaptation for Multiuser OFDM System,” *IEEE Journal on Selected Areas in Communications* vol. 21, no. 2, pp. 171-178, Feb. 2003.

[18] Y. J. Zhang and K. B. Letaief, “Multiuser Adaptive Subcarrier-and-Bit Allocation With Adaptive Cell Selection for OFDM Systems,” *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, Sept. 2004.

[19] Y. J. Zhang and K. B. Letaief, “Adaptive Resource Allocation and Scheduling for Multiuser Packet-based OFDM Networks,” in *Proc. of IEEE ICC*, pp. 2949-2953, June 2004.

[20] T. Chee, C. C. Lim, and J. Choi, “Adaptive Power Allocation with User Prioritization for Downlink Orthogonal Frequency Division Multiple Access Systems,” in *Proc. of 9th IEEE International Conf. on Communication Systems*, pp. 210-214, Sept. 2004.

[21] H. Yin and H. Liu, “An Efficient Multiuser Loading Algorithm for OFDM-based Broadband Wireless Systems,” in *Proc. of IEEE Globecom*, pp. 103-107, Nov. 2000.

[22] L. Li and A. Goldsmith, "Optimal Resource Allocation for Fading Broadcast Channels- Part I: Ergodic Capacity," *IEEE Trans. on Information Theory*, vol. 47, no. 3, pp. 1083-1102, March 2001.

APPENDIX I
OTHER SUB-OPTIMAL ALGORITHMS

In this appendix we provide several other sub-optimal algorithms. As in the algorithm in Section III-D, each of these algorithms can be viewed as determining some metric of each user/tone and then assigning the tone to the user with the largest metric. After the tone assignment, the optimal power allocation is performed across tones (power could also be uniformly allocated).

1) Suboptimal without Full Iteration: The first class of algorithms we consider is based on attempting to estimate the optimal Lagrange mulitplier λ^* as in the optimal algorithm. Recall this requires an iterative search to minimize $L(\lambda)$; for these sub-optimal algorithms, we estimate this by only apply a limited number of iterations for a given search technique. Let $\hat{\lambda}$ be the value of λ obtained after these iterations. We then use the metric $\mu_{ij} = w_i h(\hat{\lambda}, w_i e_{ij}, s_{ij})$ to determine the tone assignment. We have considered the following two search techniques for generating new values of λ:

a) Golden section search method: [10] In this case, we use the golden search method for minimizing the convex function $L(\lambda)$ (e.g. [1]). This method keeps track of an interval containing the optimal λ starting from the initial interval $[\lambda_{\min}, \lambda_{\max}] = \left[0, \max_{(i,j)} e_i w_{ij}\right]$. At each iteration, one of the boundary points on the interval is updated by comparing the value of $L(\lambda)$ at two points in the interior of the interval and the value at the boundary points. This can be done in such as way that only one new value of $L(\lambda)$ needs to be constructed in each iteration. No subgradient information is needed during the iterations.

b) Subgradient-weighted search method: This search method uses the subgradients of $L(\lambda)$ to guide the search. Starting from interval $[\lambda_{\min}, \lambda_{\max}] = \left[0, \max_{(i,j)} e_i w_{ij}\right]$, each iteration consists of two steps updates:

- First calculate the subgradients of the two boundary points, say $sb_{\min}$ and $sb_{\max}$.[11] Then calculate

$$\tilde{\lambda}^1 = \frac{\lambda_{\min} |sb_{\max}| + \lambda_{\max} |sb_{\min}|}{|sb_{\min}| + |sb_{\max}|}. \qquad (16)$$

  The rational behind (16) is the following: if the subgradient $|sb_{\min}|$ is much smaller than $|sb_{\max}|$, then it is reasonable to believe that the optimal value of λ is much closer to $\lambda_{\min}$ than to $\lambda_{\max}$. Next, find [12]

$$\left(\boldsymbol{x}(\tilde{\lambda}^1), \boldsymbol{p}(\tilde{\lambda}^1)\right) = \arg \max_{(\boldsymbol{x},\boldsymbol{p}) \in \mathcal{X}} L\left(\boldsymbol{x}, \boldsymbol{p}, \tilde{\lambda}^1, \boldsymbol{\mu}^*(\tilde{\lambda}^1)\right).$$

  Update one the boundary points of $[\lambda_{\min}, \lambda_{\max}]$ based on the subgradient value, $P - \sum_{(i,j)} p_{ij}(\tilde{\lambda}^1)$.

TABLE V
SIMULATION RESULTS ($\alpha = 0.5$, 64 SUBCHANNELS, ADJACENT SUBCHANNELIZATION, NO PER USER SINR CONSTRAINTS)

Algorithm	Utility	Log Utility	Rate (kbps)	User Scheduled
FULL	1236	12.58	497.8	5.40
GOLDEN-1	1146	12.03	596.5	4.26
GOLDEN-4	1218	12.45	537.8	4.83
WEIGHTED-1	1161	12.09	596.7	4.35
WEIGHTED-4	1236	12.58	497.6	5.37
MO-we	1108	11.80	629.5	4.29
MO-wR	1234	12.56	498.3	5.17

- Based on $\boldsymbol{x}(\tilde{\lambda}^1)$, we then perform a second update of λ, $\tilde{\lambda}^2$, using Algorithm 1. Again we update one the boundary points of $[\lambda_{\min}, \lambda_{\max}]$ based on the subgradient of $\tilde{\lambda}^2$.

2) Single sort with other metrics: A second class of sub-optimal algorithms involves sorting the users as in Section III-D, but using a different metric. One other example we have considered is using $w_i e_{ij}$.

A. Performance Comparison

Table V shows simulation results for 5 different suboptimal algorithms along with the FULL and MO-$w\bar{R}$ algorithms. The settings are the same as those in Table I. The algorithm GOLDEN-x is based on using the golden section search with at most x iterations. Likewise, WEIGHTED-x is based on using the subgradient-weighted search method with at most x iterations.[13] The algorithm MO-we uses a single sort with the $w_i e_{ij}$ metric. It can be seen that the WEIGHTED-4 algorithm performs closest to the FULL algorithm, while all other algorithms perform slightly worse than the MO-$w\bar{R}$ algorithm. Similar trends were observed for other system parameters.

[10] In the simulation results for the FULL algorithm, this search technique was also used.

[11] In case of ties, pick any subgradient in the tie. If any boundary point has both negative and positive subgradients, then it is the optimal value of λ.

[12] Again, if ties exist, just pick any of them.

[13] Note that a single iteration of the WEIGHTED algorithm requires two updates of λ, compared with one update of λ with the GOLDEN algorithm.