Scheduling and Resource Allocation in OFDMA Wireless Systems

Jianwei Huang, Vijay Subramanian, Randall Berry, and Rajeev Agrawal

February 2009

Book chapter in "Orthogonal Frequency Division Multiple Access Fundamentals and Applications."

1 Introduction

Scheduling and resource allocation are essential components of wireless data systems. Here by scheduling we refer the problem of determining which users will be active in a given time-slot; resource allocation refers to the problem of allocating physical-layer resources such as bandwidth and power among these active users. In modern wireless data systems, frequent channel quality feedback is available enabling both the scheduled users and the allocation of physical layer resources to be dynamically adapted based on the users' channel conditions and quality of service (QoS) requirements. This has led to a great deal of interest both in practice and in the research community on various "channel aware" scheduling and resource allocation algorithms. Many of these algorithms 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 [1–4, 8, 9, 25, 28, 29]. One example is the "proportionally fair rule" [3,4] first proposed for CDMA 1xEVDO based on a logarithmic utility function of each user's throughput. A larger class of throughput-based utilities is considered in [2] where efficiency and fairness are allowed to be traded-off. The "Max Weight" policy (e.g. [6–8]) can also be viewed as a gradient-based policy, where the utility is now a function of a user's queue-size or delay.

Compared to TDMA and CDMA technologies, OFDMA divides the wireless resource into non-overlapping frequency-time chunks and offers more flexibility for resource allocation. It has many advantages such as robustness against intersymbol

interference and multipath fading as well as and lower complexity of receiver equalization. Owing to these OFDMA has been adopted the core technology for most recent broadband wireless data systems, such as IEEE 802.16 (WiMAX), IEEE 802.11a/g (Wireless LANs), and LTE for 3GPP.

This chapter discusses gradient-based scheduling and resource allocation in OFDMA systems. This builds on previous work specific to the single cell downlink [28] and uplink [25] setting (e.g., Fig. 1). The key contribution of the book chapter is providing a general framework that includes each of these as special cases and also applies to multiple cell/sector downlink transmissions (e.g., Fig. 2). In particular, several important practical constraints are included in this framework, namely, 1) integer constraints on the tone allocation, i.e., a tone can be allocated to at most one user; 2) constraints on the maximum SNR (i.e., rate) per tone, which models a limitation on the available modulation and coding schemes; 3) "self-noise" on tones due to channel estimation errors (e.g., [11]) or phase noise [24]; and 4) user-specific minimum and maximum rate constraints. We not only provide the optimal algorithm for solving the optimization problem corresponding to the generalized model, but also provide low complexity heuristic algorithms that achieve close to optimal performance.

Most previous work on OFDMA systems focused on solving the resource allocation problem without jointly considering the problem of user scheduling. We will briefly survey this work in the next section. Then we describe our general formulation together with the optimal and heuristic algorithms to solve the problem. Finally, we will summarize the chapter and outline some future research directions.

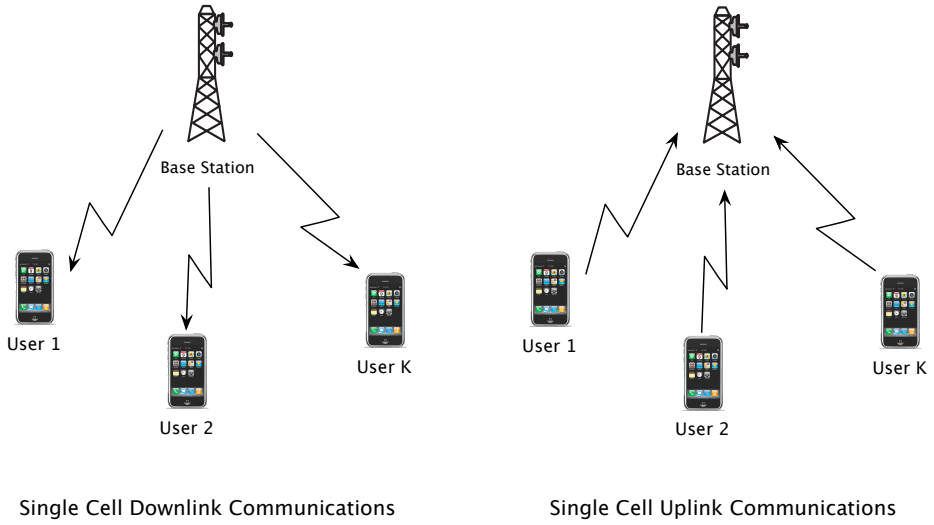

Figure 1: Example of a single cell downlink (left) and uplink scenerio (right).

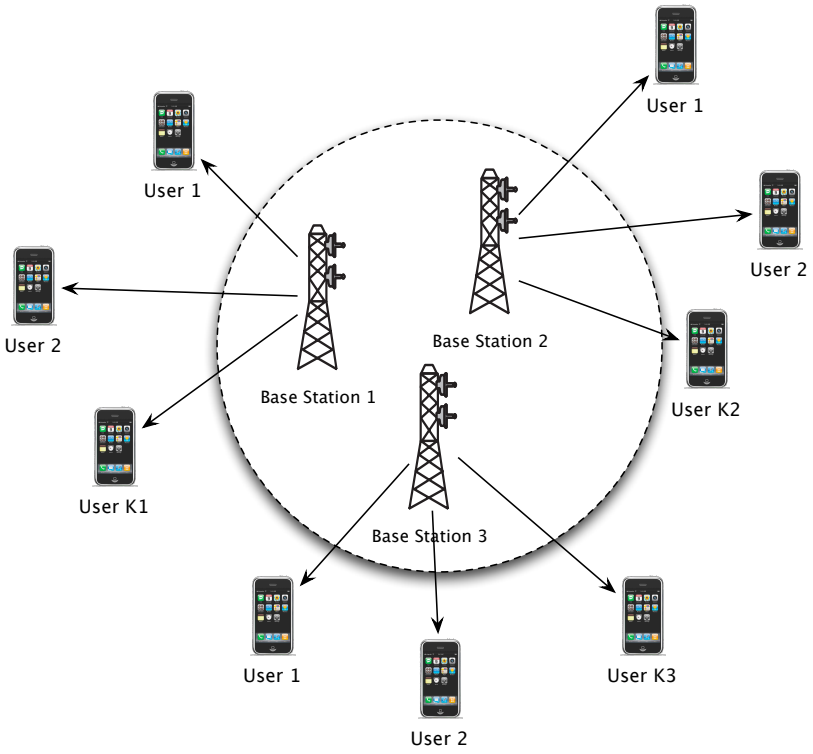

Figure 2: Example of a multiple cell/sector downlink scenerio (the different base stations could represent different sectors of the same base station shown by the circle).

2 Related Work on OFDMA resource allocation

A number of formulations for single cell downlink OFDMA resource allocation have been studied (e.g., [12–21]). In [13, 14], the goal is to minimize the total transmit power given target bit-rates for each user. In [14], the target bit-rates are determined by a fair queueing algorithm, which does not take into account the users' channel conditions. A number of papers including [15–18, 20, 21] have studied various sum-rate maximization problems, given a total power constraint. In [16–18] there is also a minimum bit-rate per user that must be met. [21] considers both minimum and maximum rate targets for each user and also takes into account several constraints that arise in Mobile WiMax. In [20], certain "delay sensitive" users are modeled as having fixed target bit-rate (i.e. their maximum and minimum rates are the same), while other "best effort" users have no bit-rate constraints. Thus the scheduler attempts to maximize the sum-rate of the best effort users while meeting the rate-targets of the delay sensitive ones. In [12, 19], weighted sum-rate maximization is considered. This is a special case of the resource allocation problem we study here for a given time-slot but does not account for constraints on the SNR per carrier, rate constraints, or self-noise. In [12], 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 [15–17]; we will briefly discuss a related approach in Section 4. In [19], a dual-based algorithm similar 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. In [22], the information theoretic capacity region of a single cell downlink broadcast channel with frequency-selective fading using a TDM scheme is given; the feasible rate region we consider, without any maximum SNR and rate constraints, can be viewed as a special case of this region. None of these papers consider self-noise, rate constraints or per user SNR constraints. Moreover, most of these papers optimize a static objective function, while we are interested in a dynamic setting where the objective changes over time 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 [1–3,6–8,29] is predicated on solving the weighted-rate optimization problem exactly in each time-slot. Simulation results in [28] show that this does hold for the heuristics presented in Section 4.

Resource allocation for a single cell OFDMA uplink has been presented in [32–39]. In [32], a resource allocation problem was formulated in the framework of Nash Bargaining, and an iterative algorithm was proposed with relatively high complexity. The authors of [33] proposed a heuristic algorithm that tries to minimize each user's transmission power while satisfying the individual rate constraints. In [34], the author considered the sum-rate maximization problem, which is a special case of the problem

considered here with equal weights. The algorithm derived in [34] assumes Rayleigh fading on each subchannel; we do not make such an assumption here. In [35], an uplink problem with multiple antennas at the base station was considered; this enables spatial multiplexing of subchannels among multiple users. Here, we focus on single antenna systems where at most one user can be assigned per sub-channel. The work in [36–39] is closer to our model. The authors in [36] also considered a weighted rate maximization problem in the uplink case, but assumed static weights. They proposed two algorithms, which are similar to one of the algorithms described in this chapter. We propose several other algorithms that outperform those in [36] with similar or slightly higher complexity. Paper [37] generalized the results in [36] by considering utility maximization in one time-slot, where the utility is a function of the instantaneous rate in each time-slot. Another work that focused on per time-slot fairness is [39]. Finally, [38] proposed a heuristic algorithm based on Lagrangian relaxation, which has high complexity due to a subgradient search of the dual variables.

Resource allocation and interference management of multi-cell downlink OFDMA systems were presented in [42–49]. A key focus of these works is on interference management among multiple cells. Our general formulation includes the case where resource coordination leads to no interference among different cells/sectors/sites. In our model, this is achieved by dynamically partitioning the subchannels across the different cells/sectors/sites. In addition to being easier to implement, the interference free operation assumed in our model allows us to optimize over a large class of achievable rate regions for this problem. If the interference strength is of the order of the signal strength, as would be typical in the broadband wireless setting, then this partitioning approach could also be the better option in an information theoretic sense [31].[1]

3 OFDMA Scheduling and Resource Allocation

3.1 Gradient-based Wireless Scheduling and Resource Allocation Problem Formulation

Let us consider a network with a total of K users. In each time-slot t, 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 $\mathcal{R}(\boldsymbol{e}_t) \subseteq \mathbb{R}_+^K$. If a user is not scheduled his rate is simply zero. Here $\boldsymbol{e}_t$ indicates the time-varying channel state

[1] We note that our discussions do not directly apply to the case of frequency reuse, where different non-adjacent cells may use the same frequency bands. In practice, frequency reuse is typically considered together with fixed frequency allocations, while here we consider dynamic frequency allocations across different cells.

information of all users available at the scheduler at time t. The decision on the rate vector is made according to the gradient-based scheduling framework in [1–3, 29] that is basically a stochastic version of the conditional gradient/Frank-Wolfe algorithm [26]. Namely, an $\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)$ is selected that has the maximum projection onto the gradient of the system's total utility function

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

where $U_i(\cdot)$ is an increasing concave utility function that measures user i's satisfaction for different values of throughput, and $W_{i,t}$ is user i's average throughput up to time t. In other words, the scheduling and resource allocation decision is the solution to

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \nabla U(\boldsymbol{W}_t)^T \cdot \boldsymbol{r}_t = \max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_{i=1}^{K} U_i'(W_{i,t}) r_{i,t}, \tag{2}$$

where $U_i'(\cdot)$ is the derivative of $U_i(\cdot)$. As a concrete example, it is useful to consider the class of commonly used iso-elastic utility functions given in [2, 5],

$$U_i(W_{i,t}) = \begin{cases} \frac{c_i}{\alpha}(W_{i,t})^{\alpha}, & \alpha \leq 1,\ \alpha \neq 0, \\ c_i \log(W_{i,t}), & \alpha = 0, \end{cases} \tag{3}$$

where $\alpha \leq 1$ is a fairness parameter and c_i is a QoS weight. In this case, after taking derivatives, (2) becomes

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_{i} c_i (W_{i,t})^{\alpha-1} r_{i,t}. \tag{4}$$

With equal class weights ($c_i = c$ for all i), setting $\alpha = 1$ results in a scheduling rule that maximizes the total throughput during each slot. For $\alpha = 0$, this results in the proportionally fair rule, and as α increases without bound, we get closer to a max-min fair solution. Thus, this family of utility functions yields a flexible class of policies: the α parameter allows for the choice of an appropriate fairness objective while the c_i parameter allows one to distinguish relative priorities within each fairness class.

However, more generally, we consider the problem of

$$\max_{\boldsymbol{r}_t \in \mathcal{R}(\boldsymbol{e}_t)} \sum_{i} w_{i,t} r_{i,t}, \tag{5}$$

where $w_{i,t} \geq 0$ is a time-varying weight assigned to the ith user at time t. In the case of (4), we let $w_{i,t} = c_i (W_{i,t})^{\alpha-1}$. In (4) these weights are given by the gradients of throughput-based utilities; however, other methods for generating the weights (possibly depending upon queue-lengths and/or delays [6–8]) are also possible. We note

that (5) 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.

3.2 General OFDMA rate regions

The solution to (5) depends on the channel state dependent rate region $\mathcal{R}(\boldsymbol{e})$, where we suppress the dependence on time for simplicity. We consider a model appropriate for general OFDMA systems including single cell downlink and uplink as well as multiple cell/sector/site downlink with frequency sharing; related single cell downlink and uplink models have been considered in [12, 22, 25, 28]. 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 OFDMA system at most one user may be assigned to any tone. Initially, as in [13, 14], we make the simplifying assumption that multiple users can share one tone using some orthogonalization technique (e.g. TDM).[2] In practice, if a scheduling interval contains multiple OFDMA symbols, we can implement such sharing by giving a fraction of the symbols to each user; of course, each user will be constrained to use an integer number of symbols. Also, with a large number of tones, adjacent tones will have nearly identical gains, in which case this time-sharing can also be approximated by frequency sharing. The two approximations becomes tight as the number of symbols or tones increases, respectively. We discuss the case where only one user can use a tone in Section 4.

Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of tones[3] and $\mathcal{K} = \{1, 2, \ldots, K\}$ the set of users. 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 transmit power. We denote the transmit power allocated to user i on tone j by p_{ij}, and the fraction of that tone allocated to user i by x_{ij}. As tones are shared resources, 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

$$r_{ij} = x_{ij} B \log\left(1 + \frac{p_{ij} e_{ij}}{x_{ij}}\right),$$

which corresponds to the Shannon capacity of a Gaussian noise channel with bandwidth $x_{ij}B$ and received SNR $p_{ij}e_{ij}/x_{ij}$.[4] This SNR arises from viewing p_{ij} as the

[2] 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.

[3] In practice, tones may be grouped into subchannels and allocated at the granularity of subchannels. As discussed in [28], our model can be applied to such settings as well by appropriately redefining the sub-channel gains $\{e_{ij}\}$ and interpreting $\mathcal{N}$ as the set of sub-channels.

[4] To better model the achievable rates in a practical system we can re-normalize e_{ij} by γe_{ij}, where

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 bandwidth is allocated. Similarly this can also be explained by time-sharing as follows: a channel of bandwidth B is used only a fraction x_{ij} of the time with average power p_{ij} which leads to the power during channel usage to be p_{ij}/x_{ij}. Without loss of generality we set $B = 1$ in the following.

3.2.1 Self-noise

In a realistic OFDMA system, imperfect carrier synchronization and channel estimation may result in "self-noise" (e.g. [11, 24]). We follow a similar approach as in [11] to model self-noise. 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 $n_{ij} \sim \mathcal{CN}(0, \sigma^2)$.[5] 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

$$\text{Eff-SNR} = \frac{\|\tilde{h}_{ij}\|^4 p_{ij}}{\sigma_{ij}^2\|\tilde{h}_{ij}\|^2 + \delta_{ij}^2 p_{ij}\|\tilde{h}_{ij}\|^2} = \frac{p_{ij}e_{ij}}{1 + \beta_{ij}p_{ij}e_{ij}}, \tag{6}$$

where $p_{ij} = \mathrm{E}(\|s_{ij}\|^2)$, $\beta_{ij} = \frac{\delta_{ij}^2}{\|\tilde{h}_{ij}\|^2}$ and $e_{ij} = \frac{\|\tilde{h}_{ij}\|^2}{\sigma_{ij}^2}$.[6] Here, $\beta_{ij}p_{ij}e_{ij}$ 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 $1/\beta_{ij}$.

In general, β_{ij} may depend on the channel quality e_{ij}. For example, this happens when self-noise arises primarily from estimation errors. The exact dependence will depend on the details of channel estimation. As an example, using the model in [23, Section IV] it can be shown that when the pilot power is either constant or inversely proportional to channel quality subject to maximum and minimum power constraints (modeling power control), β is inversely proportional to the channel condition for large e. On the other hand $\beta_{ij} = \beta$ is a constant when self-noise is due to phase noise as in [24]. For simplicity of presentation, we assume constant $\beta_{ij} = \beta$ in the remainder of the paper (except in Fig. 4 where we we allow $\beta(e) \propto 1/e$ to illustrate the impact

$\gamma \in [0, 1]$ represents the system's "gap" from capacity.

[5] We use the notation $x \sim \mathcal{CN}(0, b)$ to denote that x is a 0 mean, complex, circularly-symmetric Gaussian random variable with variance $b := \mathrm{E}(\|x\|^2)$.

[6] This is slightly different from the Eff-SNR in [11] 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, (6) seems more reasonable in that the resource allocation will depend only on $\tilde{h}_{ij}$ and not on h_{ij}. We also note that (6) is shown in [23] to give an achievable lower bound on the capacity of this channel.

of self-noise on the optimal power allocation). The analysis is almost identical if users have different β_{ij}'s.

We assume that e_{ij} is known by the scheduler for all i and j as is β. For example, in a frequency division duplex (FDD) downlink system, this knowledge can be acquired by having the base station transmit pilot signals, from which the users can estimate their channel gains and feedback to the base station. In a time division duplex (TDD) system, these gains can also be acquired by having the users transmit uplink pilots; for the downlink case, 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.

With self-noise, user i's feasible rate on tone j becomes

$$r_{ij} = x_{ij} \log\left(1 + \frac{p_{ij}e_{ij}}{x_{ij} + \beta p_{ij}e_{ij}}\right) =: x_{ij} f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right), \tag{7}$$

where again x_{ij} models time-sharing of a tone and the function $f(\cdot)$ is given by

$$f(s) = \log\left(1 + \frac{1}{\beta + 1/s}\right), \quad \beta \geq 0. \tag{8}$$

More generally, we assume that a user i's rate on channel j is given by

$$r_{ij} = x_{ij} f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right), \tag{9}$$

for some function $f : \mathbb{R}_+ \to \mathbb{R}_+$ that is non-decreasing, twice continuously differentiable and concave with $f(0) = 0$, (without loss of generality)[7] $f'(0) := \frac{df}{ds}(0) = \lim_{s\downarrow 0} \frac{f(s)}{s} = \sup_{s>0} \frac{f(s)}{s} = 1$, and $\lim_{t\to+\infty} \frac{df}{ds}(t) = 0$. We also assume by continuity[8] that $xf(p/x)$ is 0 at $x = 0$ for every $p \geq 0$. From the assumptions on the function $f(\cdot)$ it follows that $xf(p/x)$ is jointly concave in x, p; this can be easily proved by showing that the Hessian is negative semidefinite [26, 27]. It is easy to verify that f given by (8) satisfies the above properties. We should, however, point out that using the theory of subgradients [26,27], our mathematical results easily extend to a general

[7]Using the idea that Shannon capacity $\log(1 + s)$ is a natural upper bound for $f(s)$, it follows that $0 < \frac{df}{ds}(0) \leq 1$. Therefore, if $f'(0) \neq 1$, then we can solve the problem using a scaled version of function, i.e., $\tilde{f}(s) = f(s)/\frac{df}{ds}(0)$, after scaling the rate constraints by the same amount; the power and subchannel allocations will be the same in the two cases. The Shannon capacity upper bound also yields that $0 \leq \lim_{t\to+\infty} \frac{df}{ds}(t) \leq \lim_{s\to+\infty} \frac{f(s)}{s} \leq \lim_{s\to+\infty} \frac{\log(1+s)}{s} = 0$, as concavity of $f(\cdot)$ and $f(0) = 0$ imply that $\frac{df}{ds}(t) \leq \frac{f(t)}{t}$ for all $t > 0$.

[8]Using the Shannon capacity function, $\log(1 + s)$, upper bound, we have for $p > 0$, that $\lim_{x\downarrow 0} xf(p/x) = p \lim_{t\uparrow+\infty} \frac{f(t)}{t} \leq p \lim_{t\uparrow+\infty} \frac{\log(1+t)}{t} = 0$. For $p = 0$, we directly get the property from $f(0) = 0$.

$f(\cdot)$ that is only non-decreasing and concave. For instance, it can be easily proved from first principles that $xf(p/x)$ is jointly concave in (x, p) if $f(\cdot)$ is merely concave. We consciously choose the simpler setting of twice continuously differentiable functions to keep the level of discussion simple, but to aid a more interested reader, we will strive to point out the loosest conditions needed for each of our results. Before proceeding we should point out that, operationally, $f(\cdot)$ is a function of the received signal-to-noise ratio, and thus, abstracts the usage of all possible single-user decoders, including the optimal decoder that yields Shannon capacity.

3.2.2 General power constraint - single cell downlink, uplink and multi-cell downlink with frequency sharing

Let $\{\mathcal{K}_m\}_{m=1}^M$ be non-empty subsets of the set of users $\mathcal{K}$ that form a covering, i.e., $\cup_{m=1}^M \mathcal{K}_m = \mathcal{K}$. We assume that there is a vector of non-negative power budgets $\{P_m\}_{m=1}^M$ associated with these subsets, so that $\sum_{i\in\mathcal{K}_m}\sum_j p_{ij} \le P_m$ for each m. This condition ensures that there is no user who is unconstrained in its power usage. This provides a common formulation of the single cell downlink and uplink scheduling problems as described in [28] and [25], respectively. For the single cell downlink problem $M = 1$ and $\mathcal{K}_1 = \mathcal{K}$, and for the single cell uplink problem $M = K$ and $\mathcal{K}_i = \{i\}$ for $i \in \mathcal{K}$. More generally, if $\{\mathcal{K}_m\}_{m=1}^M$ is a partition, i.e., mutually disjoint, then we can view the "transmitters" for users $i \in \mathcal{K}_m$ as co-located with a single power amplifier. For example, such a model may arise in the downlink case where $\mathcal{M} := \{1, 2, \ldots, M\}$ represents sectors or sites across which we need to allocate common frequency/channel resources, but which have independent power budgets. A key assumption, however, is that we can make the transmissions from the different sectors/sites non-interfering by time-sharing or by some other suitable orthogonalization technique.

3.2.3 Capacity Region - max SNR and min/max rate constraints

Under these assumptions, the rate region can be written as

$$\begin{aligned}\mathcal{R}(\boldsymbol{e}) = \Bigg\{\boldsymbol{r} : r_i = \sum_j x_{ij} f\left(\tfrac{p_{ij}e_{ij}}{x_{ij}}\right) \text{ and } R_i^{\min} \le r_i \le R_i^{\max},\ \forall i,\\ \sum_{i\in\mathcal{K}_m}\sum_j p_{ij} \le P_m,\ \forall m, \sum_i x_{ij} \le 1,\ \forall j, (\boldsymbol{x},\boldsymbol{p}) \in \mathcal{X}\Bigg\},\end{aligned} \tag{10}$$

where

$$\mathcal{X} := \left\{(\boldsymbol{x},\boldsymbol{p}) \ge \boldsymbol{0} : x_{ij} \le 1, p_{ij} \le \tfrac{x_{ij}s_{ij}}{e_{ij}}\ \forall i, j\right\}. \tag{11}$$

Here and in the following, a boldfaced symbol will indicate the vector of the corresponding scalar quantities, e.g. $\boldsymbol{x} := (x_{ij})$ and $\boldsymbol{p} := (p_{ij})$. Also, any inequality such

as $\boldsymbol{x} \geq \mathbf{0}$ should be interpreted componentwise. The linear constraint on (x_{ij}, p_{ij}) in (11) using s_{ij} models a constraint on the maximum rate per subchannel due to a limitation on the available modulation and coding schemes; if user i can send at a maximum rate of $\tilde{r}_{ij}$ on tone j, then $s_{ij} = f^{-1}(\tilde{r}_{ij})$. We have also assumed that each user $i \in \mathcal{K}$ has maximum and minimum rate constraints $R_i^{\max}$ and $R_i^{\min}$, respectively. In order to have a solution we assume that the vector of minimum rates $\{R_i^{\min}\}_{i\in\mathcal{K}}$ is feasible. For the vector of maximum rates, it is more convenient to assume that $\{R_i^{\max}\}_{i\in\mathcal{K}}$ is infeasible. Otherwise the optimization problem associated with feasibility (see Section 3.5) will yield an optimal solution. Typically we will set $R_i^{\min} = 0$ and $R_i^{\max}$ to be the (time-varying) buffer occupancy. However, with tight minimum throughput demands one can imagine using a non-zero $R_i^{\min}$ to guarantee this.

3.3 Optimal Algorithms

From (5) and (10), the optimal scheduling and resource allocation problem can be stated as:

$$\max_{(\boldsymbol{x},\boldsymbol{p})\in\mathcal{X}} V(\boldsymbol{x},\boldsymbol{p}) := \sum_i w_i \sum_j x_{ij} f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right) \tag{P2}$$

$$\text{subject to: } \sum_j x_{ij} f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right) \geq R_i^{\min} \qquad \forall i \in \mathcal{K} \qquad (\eta_i)$$

$$\sum_j x_{ij} f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right) \leq R_i^{\max} \qquad \forall i \in \mathcal{K} \qquad (\gamma_i)$$

$$\sum_i x_{ij} \leq 1 \quad \forall j \in \mathcal{N} \qquad (\mu_j)$$

$$\sum_{i\in\mathcal{K}_m} \sum_j p_{ij} \leq P_m \qquad \forall m = 1, 2, \ldots, M \qquad (\lambda_m)$$

where set $\mathcal{X}$ is given in (11). As a rule, variables at the right of constraints will indicate the dual variables that we will use to relax those constraints while constructing the dual problem later.

One important point to note is that as described above, the optimization problem (P2) is not convex and so we can not appeal to standard results such as Slater's conditions to guarantee that is has zero duality gap [26, 27]. In particular, note that the maximum rate constraints have a concave function on the left side. To show that we still have no duality gap, we will consider a related convex problem in higher dimensions that has the same primal solution and the same dual. The new

optimization problem (P1) is given by

$$\max \sum_i w_i r_i \tag{P1}$$

$$\begin{aligned}
\text{subject to: } & r_i \le \sum_j x_{ij} f\left(\frac{p_{ij} e_{ij}}{x_{ij}}\right), \qquad \forall i \in \mathcal{K} && (\alpha_i)\\
& \sum_i x_{ij} \le 1, \qquad \forall j \in \mathcal{N} && (\mu_j)\\
& \sum_{i \in \mathcal{K}_m} \sum_j p_{ij} \le P_m, \qquad \forall m = 1, 2, \ldots, M && (\lambda_m)\\
& R_i^{\min} \le r_i \le R_i^{\max}, \qquad \forall i \in \mathcal{K}\\
& (\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X}.
\end{aligned}$$

This problem is easily seen to be convex due to the joint concavity of $xf(p/x)$ as a function of (x, p) and also will satisfy Slater's condition.[9] Hence, it will have zero duality gap [26, 27]. The problem (P1) can be practically motivated as follows: the physical (PHY) layer gives the scheduler (at the MAC layer) a maximum rate that it can serve per user based upon power and subchannel allocations, and the scheduler then drains from the queue an amount that obeys the minimum and maximum rate constraints (imposed by the network layer) and the maximum rate constraint from the PHY layer output. If the scheduler chooses not to use the complete allocation given by the PHY layer, then the final packet sent by the MAC layer is assumed to be constructed using an appropriate number of padded bits. However, we will now show that at the optimal, there is no of loss optimality in assuming that the scheduler never sends less than what the PHY layer allocates, i.e., the first constraint in Problem (P1) is always made tight at an optimal solution. This point of view is exemplified in schematic shown in Figure 3.

Assume that there is an optimizer of (P1) at which for some user i, $r_i < \sum_j x_{ij} f(\frac{p_{ij} e_{ij}}{x_{ij}})$. We will now construct another feasible solution that will satisfy the above relationship with equality. Let $\gamma \in [0, 1]$ and set $\tilde{p}_{ij} := \gamma p_{ij}$. Note that by convexity, both the power and subchannel constraints are satisfied for every value of γ. Now $\sum_j x_{ij} f(\gamma \frac{p_{ij} e_{ij}}{x_{ij}})$ is a non-decreasing and continuous function of γ taking values 0 at $\gamma = 0$ and $\sum_j x_{ij} f(\frac{p_{ij} e_{ij}}{x_{ij}})$ at $\gamma = 1$. Therefore, there exists a $\gamma^* \in (0, 1)$ such that $r_i = \sum_j x_{ij} f(\gamma^* \frac{p_{ij} e_{ij}}{x_{ij}})$ as desired. This procedure can be followed for every user i for whom $r_i < \sum_j x_{ij} f(\frac{p_{ij} e_{ij}}{x_{ij}})$, so that at the end we satisfy $r_i = \sum_j x_{ij} f(\frac{\tilde{p}_{ij} e_{ij}}{x_{ij}})$ for a feasible $(\boldsymbol{x}, \tilde{\boldsymbol{p}})$. Therefore both the optimal value and an optimizer of problem (P1)

[9] More precisely, Slater's condition will be satisfied provided that the minimum rate ($R_i^{\min}$) are strictly in the interior of rate-region $\mathcal{R}(\mathbf{e})$. If $R_i^{\min} = 0$ for all i this will trivially be true.

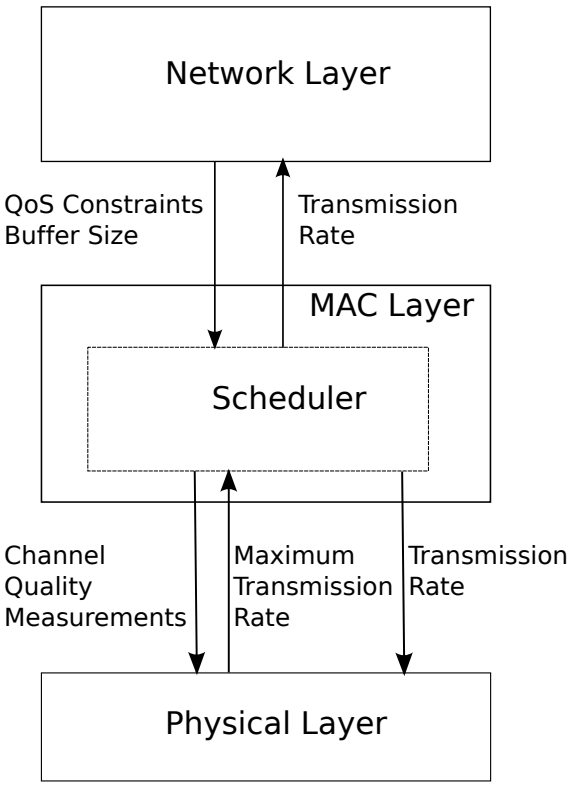

Figure 3: Schematic of a scheduler that has cross-layer visibility.

coincides with those for problem (P2). The loosest condition needed for the above to hold is $f(\cdot)$ being non-decreasing and concave with $f(0) = 0$. Henceforth, we will only work with Problem (P1).

Before proceeding to solve the problem by dual methods, we first define some key notation. For two numbers, $x, y \in \mathbb{R}$ we set $x \wedge y := \min(x, y)$, $x \vee y := \max(x, y)$ and $(x)_+ = [x]_+ := x \vee 0$.

3.3.1 Dual of Problem

We now proceed to derive a closed-form expression for the dual function for problem (P1). The Lagrangian obtained by relaxing the marked constraints of (P1) using the corresponding dual variables is given by

$$\begin{aligned} L(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \sum_i (w_i - \alpha_i) r_i + \sum_j \mu_j + \sum_{m=1}^{M} \lambda_m P_m + \sum_{i,j} \alpha_i x_{ij} f\left(\frac{p_{ij} e_{ij}}{x_{ij}}\right) \\ &\quad - \sum_j \mu_j \sum_i x_{ij} - \sum_m \lambda_m \sum_{i \in \mathcal{K}_m} \sum_j p_{ij}. \end{aligned} \tag{12}$$

The corresponding dual function is then given by maximizing this Lagrangian over $\boldsymbol{r}, \boldsymbol{x}$ and $\boldsymbol{p}$. First optimizing over rate $r_i \in [R_i^{\min}, R_i^{\max}]$ and noting that the Lagrangian is linear in r_i we get

$$\begin{aligned} L(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = &\sum_i (w_i - \alpha_i)_+ R_i^{\max} - \sum_i (\alpha_i - w_i)_+ R_i^{\min} + \sum_j \mu_j + \sum_{m=1}^{M} \lambda_m P_m \\ &+ \sum_{i,j} \alpha_i x_{ij} f(\frac{p_{ij} e_{ij}}{x_{ij}}) - \sum_j \mu_j \sum_i x_{ij} - \sum_m \lambda_m \sum_{i \in \mathcal{K}_m} \sum_j p_{ij}. \end{aligned}$$

The optimizing $\boldsymbol{r}^*$ is given by the following

$$\forall i \in \mathcal{K},\ r_i^* \in \begin{cases} \{R_i^{\max}\} & \text{if } \alpha_i < w_i; \\ \{R_i^{\min}\} & \text{if } \alpha_i > w_i; \text{ and} \\ [R_i^{\min}, R_i^{\max}] & \text{if } \alpha_i = w_i \end{cases} \tag{13}$$

Note that the last term of equation (12) can be rewritten as

$$\sum_m \lambda_m \sum_{i \in \mathcal{K}_m} \sum_j p_{ij} = \sum_{i,j} p_{ij} \sum_{m: i \in \mathcal{K}_m} \lambda_m = \sum_{i,j} p_{ij} \hat{\lambda}_i \tag{14}$$

where $\hat{\lambda}_i := \sum_{m: i \in \mathcal{K}_m} \lambda_m$.

Now maximizing the Lagrangian over power $\boldsymbol{p}$ requires us to maximize

$$\alpha_i x_{ij}\left[f\left(\frac{p_{ij}e_{ij}}{x_{ij}}\right)-\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\frac{p_{ij}e_{ij}}{x_{ij}}\right] \tag{15}$$

over p_{ij} for each i,j. From the assumptions on the function f, it is easy to check that the maximizing p_{ij}^* will be of the form

$$\frac{p_{ij}^* e_{ij}}{x_{ij}} = g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right)\wedge s_{ij}, \tag{16}$$

for some function $g:\mathbb{R}_+\to[0,\infty]$ with $g(x)=0$ for $x\geq f'(0)$. Specifically if df/ds is monotonically decreasing, we may show that $g(\cdot)=\left(\frac{df}{ds}\right)^{-1}(\cdot)$, i.e., the inverse of the derivative of $f(\cdot)$. Otherwise, since df/ds is still a non-increasing function we can set $g(x)=\inf\{t: df/ds(t)=x\}$. Using the non-increasing property of df/ds we can see that $g(x)\wedge y = g\left(x\vee\frac{df}{ds}(y)\right)$. Note that we have assumed $df/ds(0)=1$ and $\lim_{t\to+\infty} df/ds(t)=0$ but we do not assume that $\lim_{s\to+\infty} f(s)=+\infty$ (e.g., see the self-noise example). In case $f(\cdot)$ is not differentiable, then we would define the function $g(\cdot)$ using the subgradients of $f(\cdot)$. In all cases, the key conclusion from (16) is that the optimal value of p_{ij}^* is always a linear function of x_{ij}.

Note that when $f=\log(1+\frac{1}{\beta+1/s})$, with $\beta\geq 0$, as given by (8), then

$$g(x)=q((1/x-1)_+),$$

where

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

Figure 4 shows p_{ij}^* in (16) as a function of e_{ij} for the specific choice of f from (8) with three different values of $\beta=0,0.01,0.1$. When $\beta=0$, (16) becomes a "water-filling" type of solution in which p_{ij}^* is non-decreasing in e_{ij}. For a fixed $\beta>0$, this is not necessarily true, i.e., due to self-noise, less power may be allocated to "better" subchannels. We also consider the case where $\beta=10/e$ to model the case where self-noise is due to channel estimation error.

Inserting the expression for p_{ij}^* into the Lagrangian yields

$$\begin{aligned} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\lambda}) = &\sum_i (w_i-\alpha_i)_+ R_i^{\max} - \sum_i (\alpha_i - w_i)_+ R_i^{\min} + \sum_j \mu_j + \sum_{m=1}^{M}\lambda_m P_m \\ &+\sum_{i,j} x_{ij}\left[\alpha_i f\left(g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right)\wedge s_{ij}\right) - \frac{\hat{\lambda}_i}{e_{ij}}\left(g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right)\wedge s_{ij}\right) - \mu_j\right], \end{aligned} \tag{17}$$

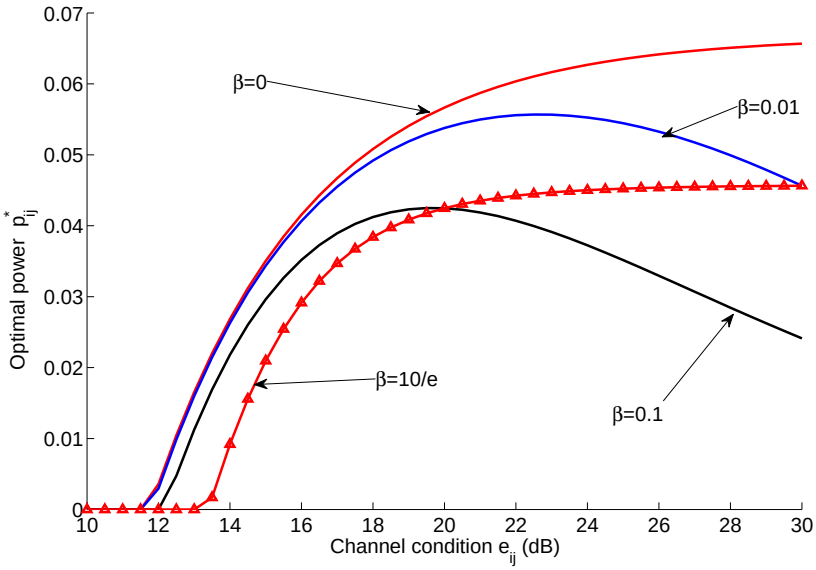

Figure 4: Optimal power p_{ij}^* as a function of the channel condition e_{ij}. Here $x_{ij} = 1$, $\alpha_i = 1$, $s_{ij} = +\infty$, and $\hat{\lambda}_i = 15$.

which is now a linear function of $\{x_{ij}\}$. Thus, optimizing over x_{ij} yields the dual function for (P1),

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_i (w_i - \alpha_i)_+ R_i^{\max} - \sum_i (\alpha_i - w_i)_+ R_i^{\min} + \sum_j \mu_j + \sum_m \lambda_m P_m \\ &\quad + \sum_{i,j} \left[\alpha_i f\left(g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}} \right) \wedge s_{ij} \right) - \frac{\hat{\lambda}_i}{e_{ij}} \left(g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}} \right) \wedge s_{ij} \right) - \mu_j \right]_+ \\ &= \sum_i \Big((w_i - \alpha_i)_+ R_i^{\max} - (\alpha_i - w_i)_+ R_i^{\min} \Big) + \sum_m \lambda_m P_m \\ &\quad + \sum_j \left(\sum_i \left[\mu_{ij}\Big(\alpha_i, \tfrac{\hat{\lambda}_i}{\alpha_i e_{ij}} \Big) - \mu_j \right]_+ + \mu_j \right), \end{aligned} \tag{18}$$

where

$$\mu_{ij}(a, b) := a\Big(f\Big(g(b) \wedge s_{ij} \Big) - b\Big(g(b) \wedge s_{ij} \Big) \Big).$$

Note that any choice such that

$$x_{ij}^* \in \begin{cases} \{1\}, & \text{if } \mu_{ij}\left(\alpha_i, \frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right) > \mu_j, \\ [0,1], & \text{if } \mu_{ij}\left(\alpha_i, \frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right) = \mu_j, \\ \{0\}, & \text{if } \mu_{ij}\left(\alpha_i, \frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right) < \mu_j \end{cases} \tag{19}$$

will optimize the Lagrangian in (17).

3.3.2 Optimizing the Dual Function over $\boldsymbol{\mu}$

From the duality theory of convex optimization [26, 27] the optimal solution to problem P1 is given by minimizing the dual function in (18) over all $(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \geq 0$. We do this coordinate-wise starting with the $\boldsymbol{\mu}$ variables. The following lemma characterizes this optimization.

Lemma 1 *For all $\boldsymbol{\alpha}, \boldsymbol{\lambda} \geq 0$,*

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\lambda}) &:= \min_{\boldsymbol{\mu} \geq \mathbf{0}} L(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= \sum_i \left((w_i - \alpha_i)_+ R_i^{\max} - (\alpha_i - w_i)_+ R_i^{\min} \right) + \sum_m \lambda_m P_m + \sum_j \mu_j^*(\boldsymbol{\alpha}, \boldsymbol{\lambda}), \end{aligned} \tag{20}$$

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

$$\mu_j^*(\boldsymbol{\alpha}, \boldsymbol{\lambda}) := \max_i \mu_{ij} \left(\alpha_i, \frac{\hat{\lambda}_i}{\alpha_i e_{ij}} \right). \tag{21}$$

The proof of Lemma 1 follows from a similar argument as in [9]. Note that (21) requires searching for the maximum value of the metrics μ_{ij} across all users for each tone j. Since $L(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ is the minimum of a convex function over a convex set, it is a convex function of $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$.

3.3.3 Optimizing the Dual Function over $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$

Now we are ready to optimize the remaining variables in the dual functions, namely, $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$. In the single cell downlink case with no rate constraints (and thus no $\boldsymbol{\alpha}$ variables), this reduces to a one dimensional problem in $\boldsymbol{\lambda}$ and hence, it can be minimized using an iterated one dimensional search (e.g., the Golden Section method [26]). Since there is no duality gap, at $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda} \geq 0} L(\boldsymbol{\lambda})$, $L(\boldsymbol{\lambda}^*)$ gives the optimal objective

value of problem (P1). Similarly, in the absence of rate constraints, the multiple sites/sectors problem with a partition of the users $\{\mathcal{K}_m\}_{m=1}^M$ also leads to a one dimensional problem within each partition.

In general, however, one would need to use subgradient methods [26, 27] to numerically solve for the optimal $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$. The following lemma characterizes the set of subgradients of $L(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ with respect to $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$.

Lemma 2 *About any* $(\boldsymbol{\alpha}^0, \boldsymbol{\lambda}^0) \geq 0$,

$$L(\boldsymbol{\alpha}, \boldsymbol{\lambda}) \geq \sum_i d(\alpha_i^0)(\alpha_i - \alpha_i^0) + \sum_m d(\lambda_m^0)(\lambda_m - \lambda_m^0), \tag{22}$$

with

$$d(\lambda_m) = P_m - \sum_{i \in \mathcal{K}_m} p_{ij}^* = P_m - \sum_{i \in \mathcal{K}_m} \frac{x_{ij}^*}{e_{ij}} g\left(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\right) \wedge s_{ij} \tag{23}$$

$$d(\alpha_m) = \sum_j x_{ij}^* f\left(g\Big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\Big) \wedge s_{ij}\right) - r_i^* \tag{24}$$

where $x_{ij}^* s$ *satisfy*

$$\sum_i x_{ij}^* \leq 1 \text{ and } \mu_j(\boldsymbol{\alpha}, \boldsymbol{\lambda}) \left(1 - \sum_i x_{ij}^*\right) = 0; \qquad \forall j,$$

and satisfy the equation (19) with $\mu_j = \mu_j^*(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ *as given in equation (21), and* r_i^* *satisfy equation (13). Thus the subgradients* $d(\lambda_m)$ *and* $d(\alpha_i)$ *are parameterized by* $(\boldsymbol{r}^*, \boldsymbol{x}^*)$ *and are linear in these variables. Moreover, the permissible values of* $\boldsymbol{r}^*$ *lie in a hypercube and those of* $\boldsymbol{x}^*$ *in a simplex.*

Observe that the dual function at any point $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ is obtained by taking the maximum of the Lagrangian over $(\boldsymbol{r}^*, \boldsymbol{p}^*, \boldsymbol{x}^*)$ satisfying $\sum_i x_{ij} \leq 1, \forall j \in \mathcal{N}$, $(\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X}$. In case $(\boldsymbol{r}^*, \boldsymbol{p}^*, \boldsymbol{x}^*)$ is unique, then the resulting Lagrangian is a gradient to the dual function at $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$. In case there are multiple optimizers, the resulting Lagrangians are each a subgradient, and every subgradient can be obtained by a convex combination of these subgradients so that the set of subgradients is convex. The lemma follows easily by substituting for the optimal $(\boldsymbol{r}^*, \boldsymbol{p}^*, \boldsymbol{x}^*)$.

Having characterized the set of subgradients, a method similar to that used in [25] for the single cell uplink problem can be used to solve for the optimal dual variables $(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$ numerically. In each step of this method we change the dual variables along the direction given by a subgradient subject to non-negativity of the dual variables. The convergence of this procedure (for a proper step-size choice) is once again guaranteed by the convexity of $L(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ (see [26, Exer. 6.3.2], [25]).

3.3.4 Optimizing the dual function over $\boldsymbol{\alpha}$

Since the dimension of $\boldsymbol{\alpha}$ equals the number of users and the dimension of $\boldsymbol{\mu}$ equals the number of tones, it may be computationally better to optimize over $\boldsymbol{\alpha}$ instead of $\boldsymbol{\mu}$ if the number of users is greater, and then use numerical methods to solve the problem. Next we detail the means to optimize over $\boldsymbol{\alpha}$ before $\boldsymbol{\mu}$. The dual function contains many terms that have definitions with $(\cdot)_+$, and therefore we would need to identify exactly when these terms are non-zero. For this we need to solve a non-linear equation which is guaranteed to have a unique solution. We first discuss this and then apply it to optimizing the dual function over $\boldsymbol{\alpha}$.

Given $y, z \geq 0$, define by $v(y,z)$ the unique solution with $1 \leq x < +\infty$ to

$$xf\Big(g\Big(\frac{1}{x} \vee \frac{df}{ds}(z)\Big)\Big) - g\Big(\frac{1}{x} \vee \frac{df}{ds}(z)\Big) = y,$$

where it is easy to show that $xf\Big(g\Big(\frac{1}{x} \vee \frac{df}{ds}(z)\Big)\Big) - g\Big(\frac{1}{x} \vee \frac{df}{ds}(z)\Big)$ is a monotonically increasing function taking value 0 at $x = 1$ and increasing without bound as $x \to +\infty$. If $y \geq f(z)/(df/ds(z)) - z \left(\geq 0\right)$, then $v(y,z) = (y+z)/f(z)$ where it is easy to verify that $v(y,z) \geq z/f(z) \geq 1/(df/ds(z)) \geq 1/(df/ds(0)) = 1$ from the concavity of $f(\cdot)$ and from $f(0) = 0$. Otherwise we need to solve for the unique $1 \leq x \leq 1/(df/ds(z))$ such that

$$xf\Big(g\Big(\frac{1}{x}\Big)\Big) - g\Big(\frac{1}{x}\Big) = y.$$

For our results we will be interested in $v\left(\frac{\mu_j e_{ij}}{\hat{\lambda}_i}, s_{ij}\right)$, using which we also define

$$\nu_{ij} := \frac{\hat{\lambda}_i v\left(\frac{\mu_j e_{ij}}{\hat{\lambda}_i}, s_{ij}\right)}{e_{ij}} \text{ and } \zeta_{ij} := \frac{\mu_j + \frac{s_{ij}\hat{\lambda}_i}{e_{ij}}}{f(s_{ij})},$$

where $\nu_{ij} = \zeta_{ij}$ if $\frac{\mu_j e_{ij}}{\hat{\lambda}_i} \geq \frac{f(s_{ij})}{\frac{df(s_{ij})}{ds}} - s_{ij}$.

First note that we can rewrite the function in (18) as follows

$$L(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_j \mu_j + \sum_m \lambda_m P_m + \sum_i \tilde{L}_i,$$

where $\tilde{L}_i = (w_i - \alpha_i)_+ R_i^{\max} - (\alpha_i - w_i)_+ R_i^{\min}$

$$+ \sum_j \frac{\hat{\lambda}_i}{e_{ij}} \left[\frac{\alpha_i e_{ij}}{\hat{\lambda}_i} f\Big(g\Big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\Big) \wedge s_{ij}\Big) - \Big(g\Big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\Big) \wedge s_{ij}\Big) - \frac{\mu_j e_{ij}}{\hat{\lambda}_i} \right]_+ .$$

Now using the quantities defined earlier in this section, one can write $\tilde{L}_i$ as follows

$$\tilde{L}_i = \sum_j \frac{\hat{\lambda}_i}{e_{ij}} \Bigg[1_{\{0 \le \alpha_i \le \zeta_{ij}\}} \Big(\frac{\alpha_i e_{ij}}{\hat{\lambda}_i} f(s_{ij}) - s_{ij} - \frac{\mu_j e_{ij}}{\hat{\lambda}_i} \Big) + \\ 1_{\{\zeta_{ij} < \alpha_i \le \nu_{ij}\}} \left(\frac{\alpha_i e_{ij}}{\hat{\lambda}_i} f\left(g\Big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}} \Big) \right) - g\Big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}} \Big) - \frac{\mu_j e_{ij}}{\hat{\lambda}_i} \right) \Bigg] \\ + (w_i - \alpha_i)_+ R_i^{\max} - (\alpha_i - w_i)_+ R_i^{\min}.$$

Minimizing $\tilde{L}_i$ over $\alpha_i \ge 0$ can now be accomplished by a simple one dimensional search; we define the optimal vector of α_is to be $\boldsymbol{\alpha}^*(\boldsymbol{\lambda}, \boldsymbol{\mu})$. Thereafter one would need to use a subgradient method [25, 26] to numerically minimize over $(\boldsymbol{\mu}, \boldsymbol{\lambda})$. A subgradient of $\tilde{L}$ with respect to λ_m is given by $P_m - \sum_{i \in \mathcal{K}_m} p_{ij}^*$ where p_{ij}^* is taken from (16) where one substitutes x_{ij}^* from (19). A subgradient of $\tilde{L}$ with respect to μ_j is given by $1 - \sum_i x_{ij}^*$ where we substitute for x_{ij}^* from (19). Note, however, that it is important that we also meet the following constraints for all i, namely,

$$R_i^{\min} \le \sum_i x_{ij}^* f\left(\frac{p_{ij}^*}{x_{ij}^*} \right) \le R_i^{\max};$$

$$\text{if } \alpha_i^* < w_i, \text{ then } \sum_j x_{ij}^* f\left(\frac{p_{ij}^*}{x_{ij}^*} \right) = R_i^{\max}; \text{ and}$$

$$\text{if } \alpha_i^* > w_i, \text{ then } \sum_j x_{ij}^* f\left(\frac{p_{ij}^*}{x_{ij}^*} \right) = R_i^{\min}.$$

The proof of this follows by retracing the steps of the proof of Lemma 2 with the roles of $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ being switched.

3.4 Primal optimal solution

For the general OFDMA problem we presented two methods to solve for V^*: in the first method we showed how to characterize the dual variables $\boldsymbol{\mu}(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ and then we proposed numerically solving for the optimal $(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$ using subgradient methods, while in the second method followed the same strategy after switching the roles of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$. However, we still need to solve for the values of the corresponding optimal primal variables. Concentrating on the first method, we know by duality theory [26] that given $(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*)$ we need to find one vector from the set of $(\boldsymbol{r}^*, \boldsymbol{x}^*, \boldsymbol{p}^*)$ that also satisfies primal feasibility and complementary slackness. These constraints can easily be seen to translate to the following:

$$d(\lambda_m^*) \ge 0, \qquad d(\lambda_m^*)\lambda_m^* = 0, \qquad \forall m; \tag{25}$$

$$d(\alpha_i^*) \ge 0, \qquad d(\alpha_i^*)\alpha_i^* = 0, \qquad \forall i. \tag{26}$$

From the linearity of $d(\lambda_m^*), d(\alpha_i^*)$ in $(\boldsymbol{r}^*, \boldsymbol{x}^*)$ it follows that the primal optimal $(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{p})$ are the solution of a linear program in $(\boldsymbol{r}^*, \boldsymbol{x}^*)$.

For the single cell downlink case with no rate constraints, as we have previously noted searching for the dual optimal is a one dimensional numerical search in λ. In that case, the search for primal optimal solution turns out to have additional structure as shown in [28].

3.5 OFDMA Feasibility

Next we turn to the corresponding feasibility problem, which can be stated as:

$$
\begin{aligned}
& V^* = \min \sigma && && (27)\\
\text{subject to: } & R_i \le \sum_j x_{ij} f\left(\frac{p_{ij} e_{ij}}{x_{ij}}\right), \quad \forall\, i && (\alpha_i)\\
& \sum_i x_{ij} \le 1 \quad \forall\, j && (\mu_j)\\
& \sum_{i \in \mathcal{K}_m} \sum_j \frac{p_{ij}}{P_m} \le \sigma \quad \forall\, m && (\lambda_m)\\
& (\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X}.
\end{aligned}
$$

The vector of rates (R_i) is feasible if $V^* \le 1$, i.e., all the power constraints will also be satisfied by a vector $(\mathbf{x}^*, \mathbf{p}^*)$. As mentioned earlier, we need to check that $(R_i) = (R_i^{\min})$ is indeed feasible; otherwise problems (P1) and (P2) are both infeasible as well. Moreover, if $(R_i) = (R_i^{\max})$ is also feasible, then $\boldsymbol{r} = (R_i^{\max})$ is the optimizer for problems (P1) and (P2). In which case, the optimal solution to the problem above with $(R_i) = (R_i^{\max})$ will also yield an optimal solution to the scheduling problem. Observe that problem (27) is convex and satisfies Slater's conditions. Finally, we also note that other alternate formulations of the feasibility problem are possible where one could either apply the σ constraint also on the subchannel utilization or switch the roles of subchannel and power utilization. All of these will yield the same conclusion about feasibility although the actual solutions, in terms of $(\boldsymbol{x}^*, \boldsymbol{p}^*)$, would possibly be different.

The Lagrangian considering the marked constraints is

$$
\begin{aligned}
L(\sigma, \boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sigma \left(1 - \sum_m \lambda_m\right) - \sum_j \mu_j + \sum_i \alpha_i R_i \\
+ \sum_{ij} \mu_j x_{ij} - \sum_{ij} \left(\alpha_i x_{ij} f\left(\frac{p_{ij} e_{ij}}{x_{ij}}\right) + p_{ij} \tilde{\lambda}_i\right)
\end{aligned}
$$

where $\tilde{\lambda}_i := \sum_{m:i\in\mathcal{K}_m} \frac{\lambda_m}{P_m}$. As before, minimizing over p_{ij} yields $\frac{p^*_{ij}e_{ij}}{x_{ij}} = g\left(\frac{\tilde{\lambda}_i}{\alpha_i e_{ij}}\right) \wedge s_{ij}$. Substituting this in the Lagrangian, we get

$$L(\sigma, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_i \alpha_i R_i - \sum_j \mu_j + \sigma\left(1 - \sum_m \lambda_m\right) - \sum_{i,j} x_{ij}\left[\alpha_i f(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \frac{\hat{\lambda}_i}{e_{ij}}(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \mu_j\right].$$

Minimizing over $0 \le x_{ij} \le 1$ yields

$$L(\sigma, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_i \tilde{L}_i - \sum_j \mu_j + \sigma\left(1 - \sum_m \lambda_m\right)$$

where

$$\tilde{L}_i = \alpha_i R_i - \sum_j \left[\alpha_i f(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \frac{\hat{\lambda}_i}{e_{ij}}(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \mu_j\right]_+.$$

Next we minimize L over all values of σ. Since there are no constraints on σ, it follows that the resulting L is finite only when $\sum_m \lambda_m = 1$; for all other values we would get $L = -\infty$. Hereafter we will assume that $\sum_m \lambda_m = 1$. Thus

$$L(\sigma, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_i \tilde{L}_i - \sum_j \mu_j.$$

Note that as before, as a function of α_i the problem is now separable. Therefore we only need to maximize $\tilde{L}_i$ over $\alpha_i \ge 0$.

Similarly we can write L as

$$L(\sigma, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_j \hat{L}_j + \sum_i \alpha_i R_i,$$

where we have

$$\hat{L}_j = -\left(\mu_j + \sum_i \left[\alpha_i f(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \frac{\hat{\lambda}_i}{e_{ij}}(g(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}) \wedge s_{ij}) - \mu_j\right]_+\right)$$

As a function of μ_j the problem is now separable, and we only need to maximize $\hat{L}_j$ over $\mu_i \ge 0$.

Thus, we could optimize first over either $\boldsymbol{\mu}$ or $\boldsymbol{\alpha}$, once again based upon whether the number of users or subchannels is smaller. In either case, the methodology and the functions that appear are very similar to the corresponding problem in the scheduling problem (P1), and due to space constraints we do not elaborate on this. Care must be take, however, while evaluating subgradients with respect to $\boldsymbol{\lambda}$. Here we propose using a projected gradient method [26, 27] based upon the constraint $\sum_m \lambda_m = 1$ to numerically solve for the optimal $\boldsymbol{\lambda}$.

3.6 Power allocation given subchannel allocation

In many of the suboptimal scheduling algorithms that we will discuss, a central feature will be a computationally simpler (but still close to optimal) method to provide a subchannel allocation. Once the subchannel allocation has been made, all that will remain is the power allocation problem, subject to the various constraints that we discussed earlier. Here we discuss how this can be solved in an optimal manner. A similar question can also be asked about the feasibility problem, hence we also discuss this here. In all cases, we assume that we are given a feasible subchannel allocation.

Since we are given a feasible subchannel allocation $\boldsymbol{x}$, the Lagrangian of the new scheduling problem (power allocation only) can be easily derived by setting $\boldsymbol{\mu} = \mathbf{0}$. For this we once again use the formulation based upon Problem (P1). The optimal power allocation is then given by $p^*_{ij} = \frac{x_{ij}}{e_{ij}}\Big(g\big(\frac{\hat{\lambda}_i}{\alpha_i e_{ij}}\big) \wedge s_{ij}\Big)$. The Lagrangian that results from substituting this formula is

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_m \lambda_m P_m + \sum_i (w_i - \alpha_i)_+ R_i^{\max} - \sum_i (\alpha_i - w_i)_+ R_i^{\min} + \sum_i \sum_j \alpha_i x_{ij} f\left(g\Big(\frac{\hat{\lambda}_i}{e_{ij}\alpha_i}\Big) \wedge s_{ij}\right) - \frac{\hat{\lambda}_i x_{ij}}{e_{ij}}\left(g\Big(\frac{\hat{\lambda}_i}{e_{ij}\alpha_i}\Big) \wedge s_{ij}\right).$$

Now it is easy to argue that if $R_i^{\min} = 0$ and $R_i^{\max} = +\infty$ and if the $\mathcal{K}_m$s form a partition, then within each partition the λ_ms can be solved for as in Section 3.3. In any case, in this setting solving for the optimal $\alpha_i \geq 0$ is easier, but uses some of the functions described at the end of Section 3.3. However, after this step we would still need to solve for $\boldsymbol{\lambda}$ numerically; if the partitions assumption holds, then it would only need a single dimensional search within each partition. A finite-time algorithm for achieving the optimal $\boldsymbol{\lambda}$ has been given in [25, 28] under the assumption that $f(\cdot)$ represents the Shannon capacity as in (8) with $\beta = 0$.

3.6.1 Feasibility check

Under the assumption that a feasible subchannel allocation has already been provided, even the feasibility check problem becomes a lot easier. As before we can assume $\sum_m \lambda_m = 1$, and that the optimal power allocation is given by $p_{ij}^* = \frac{x_{ij}}{e_{ij}}\Big(g\big(\frac{\tilde{\lambda}_i}{e_{ij}\alpha_i}\big) \wedge s_{ij}\Big)$, and substituting this we get

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_i \alpha_i \hat{R}_i - \sum_j x_{ij} \left[\alpha_i f\Big(g\Big(\frac{\tilde{\lambda}_i}{e_{ij}\alpha_i}\Big) \wedge s_{ij}\Big) - \frac{\tilde{\lambda}_i}{e_{ij}} \Big(g\Big(\frac{\tilde{\lambda}_i}{e_{ij}\alpha_i}\Big) \wedge s_{ij}\Big) \right].$$

Again solving for the optimal α_i is simpler. Once again the $\boldsymbol{\lambda}$ vector would need to be computed numerically, subject to it being a probability vector, i.e., $\sum_m \lambda_m = 1$ and $\lambda_m \geq 0$ for each m.

4 Low Complexity Suboptimal Algorithms with Integer Channel Allocation

There are two shortcomings with using the optimal algorithm outlined in the previous section for scheduling and resource allocation: (*i*) the complexity of the algorithm in general is not computationally feasible for even moderate sized systems; (*ii*) the solution found may require a time-sharing channel allocation, while practical implementations typically require a single user per sub-channel. One way to address the second point is to first find the optimal primal solution as in the previous section and then project this onto a "nearby" integer solution. Such an approach is presented in [28] for the case of a single cell downlink system ($M = 1$) without any rate constraints. In that setting, after minimizing the dual function over $\boldsymbol{\mu}$, one optimizes the function $L(\lambda)$, which only depends on a single variable. This function will have scalar subgradients which can then be used to develop rules for implementing such an integer projection. Moreover, in this case since $L(\lambda)$ is a one-dimensional function the search for the optimal dual values is greatly simplified. However, in the general setting, this type of approach does not appear to be promising.[10]

In this section we discuss a family of sub-optimal algorithms (SOA's) for the general setting that try to reduce the complexity of the optimal algorithm, while sacrificing little in performance. These algorithms seek to exploit the problem structure revealed by the optimal algorithm. Furthermore, all of these sub-optimal algorithms enforce an integer tone allocation during each scheduling interval. In the following we consider the general model from Section 3.1 with the restriction that $\{\mathcal{K}_m\}$ forms

[10] See [25] for a more detailed discussion of this in the context of the uplink scenario.

a partition of the user groups (i.e. each user is in only one of these sets) and that $R_i^{min} = 0$ for all i. In a typical setting both of these assumptions will be true.

In the optimal algorithm, given the optimal $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, the optimal tone allocation up to any ties is determined by sorting the users on each tone according to the metric $\mu_{ij}(\alpha_i, \frac{\hat{\lambda}_i}{\alpha_i e_{ij}})$ (cf. (19)). Given an optimal tone allocation, the optimal power allocation is given by (16). In each SOA, we use the same two phases with some modifications to reduce the complexity of computing $(\boldsymbol{\lambda}, \boldsymbol{\alpha})$ and the optimal tone allocation. Specifically, we begin with a *subChannel Allocation (CA)* phase in which we assign each tone to at most one user. We consider two different SOAs that implement the CA phase differently. In SOA1, instead of using the metric given by the optimal $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, we consider metrics based on a constant power allocation over all tones assigned to a partition. In SOA2, we find the tone allocation, once again through a dual based approach, but here we first determine the number of tones assigned to each user and then match specific tones and users. In all cases we assign the tones to distinct partitions which will, in turn, yield an interference-free operation. After the tone allocation is done in both SOAs, we execute the *Power Allocation (PA)* phase in which each user's power is allocated across the assigned tones using the optimal power allocation in (16).

4.1 CA in SOA1: Progressive Subchannel Allocation Based on Metric Sorting

In this family of SOAs, tones are assigned sequentially in one pass based on a per user metric for each tone, i.e., we iterate N times, where each iteration corresponds to the assignment of one tone. Let $\mathcal{N}_i(n)$ denote the set of tones assigned to user i after the nth iteration. Let $g_i(n)$ denote user i's metric during the nth iteration and let $l_i(n)$ be the tone index that user i would like to be assigned if he/she is assigned the nth tone. The resulting CA algorithm is given in Algorithm 1. Note that all the user metrics are updated after each tone is assigned.

We consider several variations of Algorithm 1 which correspond to different choices for steps 4 and 5. The choices for step 4 are:

(4A): Sort the tones based on the best channel condition among *all* users. This involves two steps. First, for each tone j, find the best channel condition among all users and denote it by $\tilde{\mu}_j := \max_i e_{ij}$. Second, find a tone permutation $\{\alpha_j\}_{j\in\mathcal{N}}$ such that $\tilde{\mu}_{\alpha_1} \geq \tilde{\mu}_{\alpha_2} \geq \cdots \geq \tilde{\mu}_{\alpha_N}$, and set $l_i(n) = \alpha_n$ for each user i at the nth iteration. Each max operation has complexity of $O(K)$, and the sorting operation has a complexity of $O(N\log(N))$. The total complexity is $O(NK + N\log N)$. We note that this is a one-time "pre-processing" that needs to done before the CA phase starts. During the tone allocation iterations, the users just choose the tone index

Algorithm 1 CA Phase for SOA1

1: Initialization: set $n = 0$ and $\mathcal{N}_i(n) = \emptyset$ for each user i.
2: **while** $n < N$ **do**
3: $n + 1$.
4: Update tone index $l_i(n)$ for each user i.
5: Update metric $g_i(n)$ for each user i.
6: Find $i^*(n) = \arg\max_i g_i(n)$ (break ties arbitrarily).
7: **if** $g_{i^*(n)}(n) \geq 0$ **then**
8: Assign the nth tone to user $i^*(n)$:

$$\mathcal{N}_i(n) = \begin{cases} \mathcal{N}_i(n-1) \cup \{l_i(n)\}, & \text{if } i = i_n^*; \\ \mathcal{N}_i(n-1), & \text{otherwise.} \end{cases}$$

9: **else**
10: Do not assign the nth tone.
11: **end if**
12: **end while**

from the sorted list.

(4B): Sort the tones based on the channel conditions for each *individual* user. For each user i at the nth iteration, set $l_i(n)$ to be the tone index with the largest gain among all unassigned tones, i.e., $l_i(n) = \arg\max_{j \in \mathcal{N} \setminus \cup_i \mathcal{N}_i(n-1)} e_{ij}$. This requires K sorts (one per user); these also need to be performed only once (since each tone assignment does not change a user's ordering of the remaining tones) and can be done in parallel. The total complexity of the K sorting operations is $O(KN \log N)$, which is higher than that in *(4A)*.

During the nth iteration, let $k_i(n) = |\cup_{j \in \mathcal{K}_m(i)} \mathcal{N}_j(n)|$ denote the number of tones assigned to users in the group to which user i belongs, i.e., $m(i)$. The choices for Line 5 are:

(5A): Set $g_i(n)$ to be the total increase in user i's utility if assigned tone $l_i(n)$, assuming the power for each user group is allocated uniformly over the tones assigned

to that group, i.e.,

$$g_i(n) = \begin{cases} w_i\left[\left(\sum_{j\in\mathcal{N}_i(n-1)\cup\{l_i(n)\}} f\left(\frac{P_i e_{ij}}{k_i(n-1)+1}\wedge s_{ij}\right)\right)\wedge R_i^{max}\right. \\ \qquad \left. -\left(\sum_{j\in\mathcal{N}_i(n-1)} f\left(\frac{P_i e_{ij}}{k_i(n-1)}\wedge s_{ij}\right)\right)\wedge R_i^{max}\right] & \text{,if } k_i(n-1)>0; \\ w_i\left[\left(\sum_{j\in\mathcal{N}_i(n-1)\cup\{l_i(n)\}} f\left(\frac{P_i e_{ij}}{k_i(n-1)+1}\wedge s_{ij}\right)\right)\wedge R_i^{max}\right] & \text{,otherwise.} \end{cases} \tag{28}$$

(5B): Set $g_i(n)$ to be user i's gain from only tone $l_i(n)$, again assuming constant power allocation within each group, i.e.

$$g_i(n) = w_i\left[f\left(\frac{P_i e_{i,l_i(n)}}{k_i(n-1)+1}\wedge s_{ij}\right)\wedge R_i^{max}\right].$$

Compared with $(5A)$, this metric is simpler to calculate but ignores the change in user i's utility due to the decrease in power allocated to any tones in $\mathcal{N}_i(n-1)$. It also does not accurately enforce the maximum rate constraint, since it only considers one tone at a time.

The complexity of either of these choices over N iterations is $O(NK)$, and so the total complexity for the CA phase is $O(NK + N\log N)$ (if $(4A)$ is chosen) or $O(KN\log N)$ (if $(4B)$ is chosen). Algorithms similar to SOA1 with $(4B)$ and $(5B)$ have been proposed in the literature for both the single cell downlink setting [12][11] and the uplink [36] without rate or SNR constraints. In the single cell downlink case, the algorithm in [12] is shown via numerical examples to have near optimal performance. In the uplink case, this also performs reasonably well in simulations [36], but [25] shows that better performance can be obtained using $(4B)$ and $(5A)$ instead.

4.2 CA in SOA2: tone Number Assignment & tone User Matching

SOA2 implements the CA phase through two steps: tone number assignment (CNA) and tone user matching (CUM). The algorithm is summarized in Algorithm 2.

[11]The main difference with the algorithm in [12] is that after each iteration n, it then checks to see if $\sum_i w_i r_i$ is increasing and if not it stops at iteration $n-1$. Such a step can be added to Algorithm 1; however, unless the system is lightly loaded it is unlikely to have a large impact on the performance.

Algorithm 2 CA Phase of SOA2

1: subChannel Number Assignment (CNA) step: determine the number of tones n_i allocated to each user i such that $\sum_{i\in\mathcal{K}} n_i \leq N$.
2: subChannel User Matching (CUM) step: determine the tone assignment $x_{ij} \in \{0,1\}$ for all users i and tones j, such that $\sum_{j\in\mathcal{N}} x_{ij} = n_i$.

4.2.1 subChannel Number Assignment (CNA)

In the CNA step, we determine the number of tones n_i assigned to each user $i \in \mathcal{K}$. The assignment is calculated based on the approximation that each user sees a flat wide-band fading tone. Notice that here we do not specify which tone is allocated to which user; such a mapping will be determined in the CUM step. The CNA step is further divided into two stages: a basic assignment stage and an assignment improvement stage.

Stage 1, Basic Assignment: Here, the assignment is based on the normalized SNR averaged over all tones. Specifically, we model each user i as having a normalized SNR $\overline{e}_i = \frac{1}{N}\sum_{j\in\mathcal{N}} e_{ij}$, and then determine a tone number assignment n_i for all i by solving:

$$\begin{aligned} \max_{\{n_i\geq 0, i\in\mathcal{K}\}} & \sum_{i\in\mathcal{K}} w_i n_i f\left(\frac{P_{m(i)}\overline{e}_i}{\sum_{j\in\mathcal{K}_{m(i)}} n_j} \wedge s_i\right) \\ \text{subject to:} & \sum_{i\in\mathcal{K}} n_i \leq N \\ & n_i f\left(\frac{P_{m(i)}\overline{e}_i}{\sum_{j\in\mathcal{K}_{m(i)}} n_j} \wedge s_i\right) \leq R_i^{max}. \end{aligned} \tag{SOA2-CNA}$$

Here, we are again assuming that power is allocated uniformly over all the channels assigned to a given user group.

Unfortunately, in general the objective in Problem SOA2-CNA is not concave. However, in the special case of the uplink ($\mathcal{K}_{m(i)} = \{i\}$) it will be.[12] In the case of the single cell downlink, if $nf(a/n)$ is increasing for all $a > 0$ (as in our general formulation), then the problem can be re-formulated to have a concave objective by noting that in this case it must be that $\sum_{i\in\mathcal{K}} n_i = N$ at any optimal solution. Additionally, due to the maximum rate constraint, the constraint set may not be convex; this can be accommodated by considering a higher dimensional problem as in Section 3.3.

[12]Some care is required at the point where the SNR constraint becomes active as the objective is not differentiable there; nevertheless, by evaluating left and right derivatives the concavity can be shown.

Next, we focus on solving Problem SOA2-CNA in the uplink setting without maximum rate constraints. In this case, the problem will have a unique and possibly non-integer solution, which we can again use a dual relaxation to find. Consider the Lagrangian

$$L(\mathbf{n},\lambda) := \sum_{i\in\mathcal{K}} w_i n_i f\left(\frac{P_i\overline{e}_i}{n_i} \wedge s_i\right) - \lambda\left(\sum_{i\in\mathcal{K}} n_i - N\right).$$

Optimizing $L(\mathbf{n},\lambda)$ over $\boldsymbol{n} \geq \mathbf{0}$ for a given λ is equivalent to solving the following K subproblems,

$$n_i^*(\lambda) = \arg\max_{n_i\geq 0} w_i n_i f\left(\frac{P_i\overline{e}_i}{n_i} \wedge s_i\right) - \lambda n_i, \forall i. \tag{29}$$

Problem (29) can be solved by a simple line search over the range of $(0, N]$. Substituting the corresponding results into the Lagrangian yields

$$L(\lambda) := \sum_{i\in\mathcal{K}} w_i n_i^*(\lambda) f\left(\frac{P_i\overline{e}_i}{n_i^*(\lambda)} \wedge s_i\right) - \lambda\left(\sum_{i\in\mathcal{K}} n_i^*(\lambda) - N\right),$$

which is a convex function of λ [26]. The optimal value

$$\lambda^* = \arg\min_{\lambda\geq 0} L(\lambda) \tag{30}$$

can be found by a line section search over: $[0, \max_i w_i f(\frac{P_i\bar{e}_i}{N/K})]$[13]. For a given search precision, the maximum number of iterations needed to solve either (29) or (30) is fixed.[14]. Hence, the worst case complexity of the solving each subproblem is independent of K or N. Since there are K subproblems in (29), it follows that the complexity of the basic assignment step is $O(K)$. If the resultant channel allocations contain non-integer values, we will approximate with an integer solution that satisfies $\sum_{i\in\mathcal{K}} n_i = N$.[15] Since each user is allocated only a subset of the tones, the normalized SNR $\overline{e}_i = \frac{1}{N}\sum_{j\in\mathcal{N}} e_{ij}$ is typically a pessimistic estimate of the averaged

[13]The upperbound of the search interval can be obtained by examining the first order optimality condition of (29).

[14]For example, if we use bi-section search to solve (29) and stop when the relative error of the solution is less than $N/2^{10}$, then we only need a maximum of ten search iterations.

[15]One possible integer approximation is the following. Assume n_i^* is the unique optimal solution of Problem SOA2-CNA. First, sort users in the descending order of the mantissa of n_i^*, $fr(n_i^*) = n_i^* - \lfloor n_i^* \rfloor$. That is, find a user permutation subset $\{\alpha_k, 1 \leq k \leq N\}$ such that $fr(n_{\alpha_1}^*) \geq fr(n_{\alpha_2}^*) \geq \cdots \geq fr(n_{\alpha_M}^*)$. Second, for each user i, let $\tilde{n}_i^* = \lfloor n_i^* \rfloor$. Third, calculate the number of unallocated tones, $N^A = N - \sum_i \tilde{n}_i^*$. Finally, adjust users with large mantissas such that all the tones are allocated, i.e., $\tilde{n}_{\alpha_i}^* = \tilde{n}_{\alpha_i}^* + 1$ for all $1 \leq i \leq N^A$. The resulting $\{\tilde{n}_i^*\}_{i\in\mathcal{K}}$ give the integer approximation.

tone conditions over the allocated subset. This motivates us to consider the following assignment improvement stage of CNA.

Stage 2, Assignment Improvement: Here, assignment is performed by means of iterative calculations using the normalized SNR averaged over the best tone subset. Specifically, we iteratively solve the following variation of Problem SOA2-CNA (stated here for the uplink without maximum rate constraints):

$$\begin{aligned} \max_{\mathbf{n}(t)\geq \mathbf{0}} & \sum_{i\in\mathcal{K}} w_i n_i(t) f\left(\frac{P_i \overline{e}_i(t)}{n_i(t)} \wedge s_i\right) \\ \text{subject to: } & \sum_{i\in\mathcal{K}} n_i(t) \leq N \\ & n_i f\left(\frac{P_{m(i)} \overline{e}_i(t)}{\sum_{j\in\mathcal{K}_{m(i)}} n_j} \wedge s_i\right) \leq R_i^{max}, \end{aligned} \qquad \text{(SOA2-CNA-t)}$$

for $t = 1, 2, ...$ During the t-th iteration, $\overline{e}_i(t)$ is a refined estimate of the normalized SNR based on the best $\lfloor n_i(t-1) \rfloor$ (or $\lceil n_i(t-1) \rceil$) tones of user i; additionally, $n_i(0) := N$ for all i. The iteration stops when the tone allocation converges or the maximum number of iterations allowed is reached. An integer approximation will be performed if needed.

The complete algorithm for the CNA phase of SOA2 is given in Algorithm 3. In order to perform the assignment improvement, we need to perform K sorting operations, with a total complexity $O(KN\log(N))$. Note that this only needs to be done once. Step 4 of each iteration has complexity of $O(K)$ due to solving K subproblems for a fixed dual variable. The maximum number of iterations is fixed and thus is independent of N or K. The integer approximation stage requires a sorting with the complexity of $O(K\log(K))$. So the total complexity for the CNA phase of SOA2 is $O(KN\log(N) + K\log(K))$.

Algorithm 3 CNA Phase of SOA2

1: Initialization: integer MaxIte> 0, $t = 0$, $n_i(0) = N$ and $n_i(1) = N/2$ for each user i.
2: **while** ($n_i(t+1) \neq n_i(t)$ for some i) & ($t <$MaxIte) **do**
3: $t = t + 1$.
4: For each user i, $\overline{e}_i(t)$ = average gain of user i's best $n_i(t-1)$ tones.
5: Solve Problem (SOA2-CNA-t) to determine the optimal $n_i(t)$ for each user i.
6: **end while**
7: let $n_i^* = n_i(t)$ for each user i.

4.2.2 subChannel User Matching (CUM) Step

After the CNA step, we know how many tones are to be allocated to each user. However, we still need to determine which specific tones are assigned to which user. This is accomplished in the CUM step by finding a tone assignment that maximizes the weighted-sum rate assuming each user employs a flat power allocation, i.e. we solve the problem:

$$\begin{aligned} \max_{x_{ij}\in\{0,1\}} & \sum_{i\in\mathcal{K}}\sum_{j\in\mathcal{N}} x_{ij} w_i f\left(\frac{P_i e_{ij}}{n_i^*} \wedge s_i\right) \\ \text{subject to: } & \sum_{j\in\mathcal{N}} x_{ij} = n_i^*, \forall i \in \mathcal{K}, \\ & \sum_{i\in\mathcal{K}} x_{ij} = 1, \forall j \in \mathcal{N}, \end{aligned} \tag{SOA2-CUM}$$

where $\boldsymbol{n}^* = (n_i^*, i \in \mathcal{K})$ is the integer tone allocation obtained in the CNA step. Since we solved Problem (SOA2-CNA-t) using the average of the best $\mathbf{n}^*$, then concavity of $f(\cdot)$ ensures that any feasible tone allocation for Problem (SOA2-CUM) will satisfy the maximum rate constraint.

Problem SOA2-CUM is an integer *Assignment Problem* whose *optimal* solution can be found by using the *Hungarian Algorithm* [30].[16] To use the Hungarian algorithm here, we need to perform "virtual user splitting" as explained next. For user i, let $r_{ij} = w_i f\left(\frac{P_i e_{ij}}{n_i^*} \wedge s_{ij}\right)$, and let

$$\boldsymbol{r}_i = [r_{i1}, r_{i2}, \cdots, r_{iN}]$$

be user i's achievable rates over all possible tones. We can then form a $K \times N$ matrix $\boldsymbol{R} = \left[\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, \cdots, \boldsymbol{r}_M^T\right]^T$. Next, we split each user i into n_i^* virtual users by adding $n_i^* - 1$ copies of the row vector $\boldsymbol{r}_i$ to the matrix $\boldsymbol{R}$. This expands $\boldsymbol{R}$ into a $N \times N$ square matrix. Solving Problem SOA2-CUM is then equivalent to finding a *permutation matrix* $\boldsymbol{C}^* = [c_{ij}]_{N\times N}$ such that

$$\boldsymbol{C}^* = \arg\min_{\boldsymbol{C}\in\mathcal{C}} -\boldsymbol{C}\cdot\boldsymbol{R} := \arg\min_{\boldsymbol{C}\in\mathcal{C}} -\sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij} r_{ij}. \tag{31}$$

Here $\mathcal{C}$ is the set of permutation matrices, i.e., for any $\boldsymbol{C} \in \mathcal{C}$, we have $c_{ij} \in \{0, 1\}$, $\sum_i c_{ij} = 1$ and $\sum_j c_{ij} = 1$ for all i and j. This problem can be solved by the standard

[16] A similar idea has been used to solve various single cell downlink OFDMA resource allocation problems (e.g., [18]) as well as to find user coalitions for Nash Bargaining in an uplink OFDMA system in [32].

Hungarian algorithm which has a computational complexity of $O(N^3)$, where N is the total number of tones. The detailed algorithm can be found in [30]. After obtaining $\boldsymbol{C}^*$, we can calculate the corresponding tone allocation $\boldsymbol{x}^*$. For example, if $c^*_{kj} = 1$ and virtual user k corresponds to the actual user i, then we know $x^*_{ij} = 1$, i.e., tone j is allocated only to user i.

4.3 Power Allocation (PA) phase

We can follow the tone allocation (CA) phase in either SOA1 and SOA2 with a power allocation phase in which power is optimally allocated among the tones assigned to the users in each partition.[17] After this optimization it is possible that some tone is allocated zero power due to its poor tone gain. Alternatively, one can simply use a uniform power allocation as was assumed in the CA phase. For certain single cell downlink scenarios, such a uniform allocation has been shown to be nearly optimal in [12, 28].

Since the tone allocation is given, optimizing the power allocation for each group is equivalent to the problem considered in Section 3.6 and can be addressed in a similar way, i.e. by considering the dual formulation and numerically searching for the optimal dual variables. We note that in the uplink scenario without any maximum rate constraint, we need to solve one such problem for each user and for each problem only a single dual variable needs to be introduced (corresponding to the user's power constraint). Hence, the optimal dual value can be found through a simple line search, with a constant worst-case complexity given a fixed search precision as in our discussion of (29).

4.4 Complexity and performance of Suboptimal Algorithms for the Uplink Scenario

In this section we discuss the complexity and performance of the suboptimal algorithms in an uplink scenario without any maximum rate constraints [18]. The worst case computational complexities of the variations of SOA1 and SOA2 for this setting are summarized in Table 1.

Next we briefly discuss the performance of this algorithms with a realistic OFDMA simulator assuming parameters and assumptions commonly found in the IEEE 802.16 standards [10]. These results are for a single cell with 40 users. All users are infinitely back-logged and assigned a throughput-based utility as in (3) with parameter $c_i = 1$

[17] In this section, we again consider the case where $\{\mathcal{K}_m\}$ forms a partition of the users and allow for maximum rate constraints.

[18] It can be argued that this will also be the worst-case setting for the general problem assuming partitions and no rate constraints.

Table 1: Worst Case Computational Complexity of Suboptimal Algorithms

Suboptimal Algorithm			Worst Case Complexity
SOA1	subChannel Allocation (CA)	4A & 5A	$O(NK + N\log N)$
		4A & 5B	$O(NK + N\log N)$
		4B & 5A	$O(KN\log N)$
		4B & 5B	$O(KN\log N)$
	Power Allocation (PA)		$O(KN)$
	Total (CA + PA)		$\mathbf{O}(KN\log N)$
SOA2	subChannel Allocation (CA)	CNA	$O(KN\log N + K\log K)$
		CUM	$O(N^3)$
	Power Allocation (PA)		$O(KN)$
	Total (CA+PA)		$\mathbf{O}(N^3 + KN\log N + K\log K)$

and $\alpha = 0.5$. Each user i has a total transmission power constraint $P_i = 2$W. We calculate the achievable rate of user i on tone j as

$$r_{ij} = Bx_{ij}\log\left(1 + \frac{p_{ij}e_{ij}}{x_{ij}}\right),$$

where B is the tone bandwidth and e_{ij} is generated according to a product of a fixed location-based term and a frequency-selective fast fading term. A detailed description of the simulation set-up can be found in [25] with further results. Scheduling decisions are made every 20 OFDM symbols, which corresponds to one fading block.

Table 2 shows simulation results for the following four algorithms:

1. Integer-Dual: integer tone allocation (with tie breaking) based on optimal dual-based algorithm and optimal power control. To reduce computational complexity in the case of too many ties, we randomly inspect up to 128 ways of breaking the ties with an integer allocation and select the allocation among these with the largest weighted sum rate (before reallocating the power).

2. SOA1: tone allocation as in Section 4.1 and power control as in Section 4.3. There are four versions of SOA1, depending on how steps 4 and 5 in Algorithm 1 are implemented; we present results for each.

3. SOA2: tone allocation as in Section 4.2 (with up to 10 iterations) and power control as in Section 4.3.

Table 2: Example Uplink resource allocation performance

Algorithms		Utility	Log U	Rate	Scheduled Users
Integer-Dual		53922	514.0	21.56	37.5
SOA 1	4A & 5A	52494	510.7	22.86	34.6
	4A & 5B	51697	509.2	20.22	28.1
	4B & 5A	54165	513.3	22.25	35.0
	4B & 5B	53156	511.4	21.43	28.6
SOA 2		54316	513.6	22.33	35.1
Base Line		21406	-1960.5	16.13	2.66

4. Base-line: each tone j is allocated to the user i with the highest e_{ij}, without considering the weights w_i's and the power constraints. Each user's power is then allocated as in Section 4.3.

In this table it can be seen that SOA1 (with 4B & 5A) and SOA2 achieve the best performance in terms of total utility. Their performance is even better than the Integer-Dual approach, which was obtained based on the optimal value of the relaxed problem. This is likely because only 128 ways to break ties are considered which is typically not sufficient. Since the Integer-Dual algorithm achieves an optimality ratio of 0.9412, this suggests that SOA1 and SOA2 achieve very close to optimal performance as well. The base-line algorithm always has poor performance.

Here, and in other uplink simulation reported in [25], all of the SOAs have good performance with SOA1 (with 4B & 5A) and SOA2 consistently achieving the best performance in terms of total utility. From Table 1, we note that these have slightly higher complexity than some of the other SOAs. Hence if lower complexity is desired, this can be provided with only a slight loss in performance. We also note that in each case the SOAs and the integer-dual algorithm schedule a large number of users on average in each time-slot. A potential cost from this is that it may increase the needed signaling overhead. One way to reduce this cost is to add a penalty term to our objective which increases with the number of users scheduled.

5 Conclusions and Open Problems

In this chapter, we have considered a general model of gradient-based scheduling and resource allocation for OFDMA systems. This model includes single cell downlink, uplink, and multi-cell downlink with frequency sharing, and incorporates various

practical constraints such as per carrier SNR constraints, self-noise due to imperfect channel estimates or phase noise, and minimum and maximum per user rate constraints. Essentially the problem can be reduced to solving a weighted rate maximization problem in each time-slot. We address this problem with a Lagrangian dual relaxation method. By exploiting the structure of the OFDMA rate region, we can express the dual function in terms of a small subset of dual variables. The optimal values of these variables can be found through standard numerical search methods. An interesting observation is that recovering the optimal primal solutions given optimal dual variables is rather straightforward in most cases, since the optimal channel allocations often turn out to be integer "automatically". In the case when this is not true, we need to calculate the channel allocation by either allowing time-sharing or picking a good integer solution, and optimize the power allocation accordingly. Based on the intuition derived from the optimal algorithms, we demonstrate that it is possible to design a class of heuristic algorithms that are low in complexity but perform very well in simulation studies.

All algorithms presented in this chapter are centralized. This is not an issue for the single cell downlink case or even for a multi-sectored site, where the resource allocation decisions are made by the base station. In the uplink and multi-cell downlink cases, however, a distributed algorithm is more desirable since the decisions are made by the multiple network entities (either multiple mobile users or multiple base stations). Some preliminary results towards a fully distributed algorithm have been reported in [40, 41] and more work is needed along this line. Another open issue regarding the multi-cell downlink case is to consider models which allow dynamic frequency re-use. A challenge in such settings is that the resulting optimization problem may not longer be convex even when the integer constraints are relaxed.

The algorithms presented here require assume that the scheduler has accurate channel quality information (though some inaccuracy may be accounted for via the self-noise terms). In OFDMA systems with many users and tones, the resulting feedback overhead can become significant. This overhead can be partially reduced by proper subchannelization methods (e.g., [28]) or by not reporting the channel quality on every subchannel as in [21]. However, there is little understanding of the interplay between these or other limited feedback schemes and the resulting scheduling performance. This is another area in which additional work is warranted.

References

[1] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," *Proc. of 2002 Allerton Conference on Communication, Control and Computing*, 2002.

[2] 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.

[3] H. Kushner and P. Whiting, "Asymptotic properties of proportional-fair sharing algorithms," *40th Annual Allerton Conference on Communication, Control, and Computing*, 2002.

[4] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *Proc. of IEEE Vehicular Technology Conference*, Spring, 2000.

[5] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556 – 567, October 2000.

[6] A. L. Stolyar, "MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[7] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queue with randomly varying connectivity," *IEEE Trans. on Inform. Th.*, vol. 39, pp. 466–478, 1993.

[8] 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.

[9] R. Agrawal, V. Subramanian, and R. Berry, "Joint scheduling and resource allocation in CDMA systems," in *Proc. WiOpt*, Cambridge, UK, March 2004. Journal version under submission.

[10] "IEEE 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005," `http://www.ieee802.org/16/`.

[11] H. Jin, R. Laroia, and T. Richardson, "Superposition by position," *2006 IEEE Information Theory Workshop*, March 2006.

[12] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms," *IEEE Transactions on Communications*, vol. 52, no. 6, pp. 922–930, 2004.

[13] 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.

[14] Y. Zhang and K. Letaief, "Adaptive resource allocation and scheduling for multiuser packet-based OFDM networks," *2004 IEEE ICC*, vol. 5, 2004.

[15] J. Jang and K. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, 2003.

[16] Y. Zhang and K. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *IEEE Trans. on Wireless Communications*, vol. 3, no. 5, pp. 1566–1575, 2004.

[17] T. Chee, C. Lim, and J. Choi, "Adaptive power allocation with user prioritization for downlink orthogonal frequency division multiple access systems," *ICCS 2004*, pp. 210–214, 2004.

[18] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," *IEEE Globecom*, 2000.

[19] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA Downlink Systems," in *IEEE ISIT*, pp. 1394–1398, 2006.

[20] M. Tao, Y. C. Liang and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. on Wireless Communications*, vol. 7, no. 6, pp. 2190-2201, June 2008

[21] W. Jiao, L. Cai and M. Tao, "Competitive scheduling for OFDMA systems with guaranteed transmission rate," *Elsevier Computer Communications*, special issue on Adaptive Multicarrier Communications and Networks. Avaliable online 29 August, 2008.

[22] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels. I. Ergodic capacity," *IEEE Trans. on Information Theory*, vol. 47, no. 3, pp. 1083–1102, 2001.

[23] M. Medard, "The Effect Upon Channel Capacity in Wireless Communications of Perfect and Imperfect Knowledge of the Channel," *IEEE Trans. on Information Theory*, vol. 46, no. 3, pp. 935-946, May 2000.

[24] J. Lee, H. Lou and D. Toumpakaris, "Analysis of Phase Noise Effects on Time-Direction Differential OFDM Receivers," *IEEE GLOBECOM*, 2005

[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 Journal on Selected Areas in Communications*, accepted

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

[27] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge University Press, 2004.

[28] J. Huang, V. G. Subramanian, R. Agrawal and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. on Wireless Communications*, accepted.

[29] A. L. Stolyar, "Maximizing Queueing Network Utility subject to Stability: Greedy Primal-Dual Algorithm," *Queueing Systems*, vol. 50, pp. 401–457, 2005.

[30] E. D. Nering and A. W. Tucker, *Linear Programs and Related Problems*. Academic Press Inc., 1993.

[31] T. M. Cover and J. A. Thomas, "Elements of information theory," Second edition, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[32] Z. Han, Z. Ji, and K. Liu, "Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions," *IEEE Trans. Comm.*, vol. 53, no. 8, pp. 1366–1376, 2005.

[33] S. Pfletschinger, G. Muenz, and J. Speidel, "Efficient subcarrier allocation for multiple access in OFDM systems," in *7th International OFDM-Workshop 2002 (InOWo'02)*, 2002.

[34] Y. Ma, "Constrained Rate-Maximization Scheduling for Uplink OFDMA," in *Proc. IEEE MILCOM*, pp. 1–7, Oct. 2007.

[35] B. Da and C. Ko, "Dynamic subcarrier sharing algorithms for uplink OFDMA resource allocation," in *Proc. 6th ICICS*, pp. 1–5, 2007.

[36] K. Kim, Y. Han, and S. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Comms. Letters*, vol. 9, no. 6, pp. 526–528, 2005.

[37] C. Ng and C. Sung, "Low complexity subcarrier and power allocation for utility maximization in uplink OFDMA systems," *IEEE Trans. Wireless Comm.*, vol. 7, no. 5 Part 1, pp. 1667–1675, 2008.

[38] K. Kwon, Y. Han, and S. Kim, "Efficient Subcarrier and Power Allocation Algorithm in OFDMA Uplink System," *IEICE Trans. Comm.*, vol. 90, no. 2, pp. 368–371, 2007.

[39] L. Gao and S. Cui, "Efficient subcarrier, power, and rate allocation with fairness consideration for OFDMA uplink," *IEEE Trans. Wireless Comm.*, vol. 7, no. 5 Part 1, pp. 1507–1511, 2008.

[40] M. Chen and J. Huang, "Optimal resource allocation for OFDM uplink communication: A primal-dual approach," in *Conference on Information Sciences and Systems (CISS)*, Princeton University, March 2008

[41] X. Zhang, L. Chen, J. Huang, M. Chen and Y. Zhao, "Distributed and optimal reduced primal-dual algorithm for uplink OFDM resource allocation", submitted.

[42] C.-S. Chiu and C.-C. Huang;, "Combined partial reuse and soft handover in OFDMA downlink transmission," *IEEE VTC*, pp. 1707 – 1711, Apr 2008.

[43] N. Damji and T. Le-Ngoc, "Dynamic resource allocation for delay-tolerant services in downlink OFDM wireless cellular systems," *IEEE ICC*, vol. 5, pp. 3095 – 3099 Vol. 5, Apr 2005.

[44] G. Li and H. Liu;, "Downlink dynamic resource allocation for multi-cell OFDMA system," *IEEE VTC*, vol. 3, pp. 1698 – 1702 Vol.3, Sep 2003.

[45] H. Lei, X. Zhang, and Y. Wang;, "Real-time traffic scheduling algorithm for MIMO-OFDMA systems," *IEEE ICC*, pp. 4511 – 4515, Apr 2008.

[46] H. Xiaoben and K. Valkealahti, "On distributed and self-organized inter-cell interference mitigation for OFDMA downlink in imt-advanced radio systems," *IEEE VTC*, pp. 1736 – 1740, Jan 2007.

[47] I.-K. Fu and W.-H. Sheen;, "An analysis on downlink capacity of multi-cell OFDMA systems under randomized inter-cell/sector interference," *IEEE VTC*, pp. 2736 – 2740, Mar 2007.

[48] L. Shao and S. Roy, "Downlink multicell MIMO-OFDM: an architecture for next generation wireless networks," *IEEE WCNC*, vol. 2, pp. 1120 – 1125 Vol. 2, Feb 2005.

[49] T. Thanabalasingham, S. Hanly, L. Andrew, and J. Papandriopoulos, "Joint allocation of subcarriers and transmit powers in a multiuser ofdm cellular network," *IEEE ICC*, vol. 1, pp. 269 – 274, Jun 2006.