Optimal Downlink OFDMA Resource Allocation With Linear Complexity to Maximize Ergodic Rates

Ian C. Wong, *Student Member, IEEE*, and Brian L. Evans, *Senior Member, IEEE*

***Abstract*—OFDMA resource allocation assigns subcarriers and power, and possibly data rates, to each user. Previous research efforts to optimize OFDMA resource allocation with respect to communication performance have focused on formulations considering only instantaneous per-symbol rate maximization, and on solutions using suboptimal heuristic algorithms. This paper intends to fill gaps in the literature through two key contributions. First, we formulate continuous and discrete ergodic weighted sum rate maximization in OFDMA assuming the availability of perfect channel state information (CSI). Our formulations exploit time, frequency, and multi-user diversity, while enforcing various notions of fairness through weighting factors for each user. Second, we derive algorithms based on a dual optimization framework that solve the OFDMA ergodic rate maximization problem with $\mathcal{O}(MK)$ complexity per OFDMA symbol for M users and K subcarriers, while achieving data rates shown to be at least 99.9999% of the optimal rate in simulations based on realistic parameters. Hence, this paper attempts to demonstrate that OFDMA resource allocation problems are *not* computationally prohibitive to solve optimally, even when considering ergodic rates.**

***Index Terms*—Broadband communication, information rates, multiaccess communication, orthogonal frequency division multiple access (OFDMA), radio spectrum management.**

I. Introduction

NEXT-generation broadband wireless system standards, e.g. IEEE 802.16e [1] and 3GPP-Long Term Evolution (LTE) [2], consider Orthogonal Frequency Division Multiple Access (OFDMA) as the preferred physical layer multiple access scheme, esp. for the downlink. OFDMA allows multiple users to transmit simultaneously on the different subcarriers per OFDM symbol. In most scenarios, the channel response for each user can be considered to be statistically independent, esp. when there is considerable spatial separation among the users. Thus, we could potentially exploit this *multiuser diversity* through intelligent allocation of the subcarriers and power to each user, and increase the overall performance of the system.

The problem of assigning the subcarriers, bits, and powers to the different users in an OFDMA system has been an area of active research. In [3] and [4], the *margin-adaptive* resource allocation problem was investigated, whose objective is to minimize the total transmit power given a set of fixed user data rates and bit error rate (BER) requirements. This formulation is useful primarily for applications which require a fixed data rate, e.g. standard voice. Most of the subsequent research have been on the *rate-adaptive* problem, whose objective is to maximize the data rates subject to power and/or BER constraints. This formulation is more relevant for next-generation data-centric wireless networks. In [5], it was shown that in order to maximize the ergodic sum capacity, each subcarrier should be allocated to the user with the best channel gain on it, and the power should be allocated using the water-filling algorithm across time and frequency. However, no fairness among the users was considered. Thus, the users that have the best channel conditions will be assigned almost all the resources, which leaves many users without a chance to use the spectrum at all. Furthermore, [5] did not propose any efficient algorithm to solve this problem.

In [6], an alternative formulation that considered fairness by maximizing the minimum user's data rate was solved by using a subcarrier-exclusivity constraint relaxation method, similar to the approach in [3]. In [7], prioritization was enforced using a weighted sum rate maximization, and a similar constraint relaxation was used to derive the optimum subcarrier and power allocation. Different weights were assigned to different users, and a higher weight for a user would imply a higher priority of getting resources. In [8], the sum data rate was maximized under a *proportional rate constraint*; i.e., the rate of each user should adhere to a set of predetermined proportionality constants. The solution approach in [8] involves decoupling the subcarrier and power allocation by first running a greedy heuristic for subcarrier allocation, followed by solving a convex power allocation problem given the subcarrier allocations. In [9], tradeoffs between efficiency and fairness were realized by maximizing a concave utility function of the user's data rate. Time diversity was also exploited in [9] by maximizing the utility function of an *exponentially weighted* and *time-windowed* average data rate of each user.

In most of the aforementioned work, the formulation and algorithms only consider instantaneous performance metrics. Thus, the temporal dimension is not being exploited when the resource allocation is performed. Although [9] considered some form of temporal diversity, their approach focused more on the effect of the past channel information on the fairness, rather than exploiting the time variations directly to improve the overall data rate performance. Instead of considering only instantaneous data rate, we formulate the problem considering

Manuscript received September 17, 2006; revised February 3, 2007 and May 22, 2007; accepted June 17, 2007. The associate editor coordinating the review of this paper and approving it for publication was V. Lau. This paper was presented in part at the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, Hawaii USA 2007.

I. C. Wong was with the Electrical and Computer Engineering Dept., The University of Texas at Austin, Austin, TX 78712 USA. He is now with Freescale Semiconductor, Austin, TX 78729 USA (email: ian.wong@ieee.org).

B. L. Evans is with the Electrical and Computer Engineering Dept., The University of Texas at Austin, Austin, TX 78712 USA (email: bevans@ece.utexas.edu).

Digital Object Identifier 10.1109/TWC.2008.xxxxx.

user-weighted *ergodic* rates for both continuous (capacity-based) and discrete (adaptive modulation and coding) rates. This allows us to exploit the time dimension explicitly in the formulation, and utilize all three degrees of freedom in our system, namely frequency, time, and multiuser dimensions. At the same time, we can enforce certain notions of fairness through the user weights (e.g. proportional fairness can be attained by setting the user weights as the reciprocal of the user's average rate so far [9]). A similar ergodic formulation for continuous rates has been considered in [5], but their approach is limited to the case of maximizing the unweighted sum capacity, which is much simpler, but has limited practical value because fairness is not considered.

Furthermore, previous research have assumed that algorithms to find the optimal or near-optimal solution to the problem is too computationally complex for real-time implementation. A popular approach to attain near-optimality is *constraint relaxation* (see e.g. [3] [6] [7]). This approach performs a convex reformulation of the problem by relaxing the binary integer constraints $x_{m,k} \in \{0,1\}$ which indicate a subcarrier assignment of user m to subcarrier k; to interval constraints $0 \leq x_{m,k} \leq 1$, where $x_{m,k}$ is now a *sharing factor*. The solution to the reformulated convex problem is then projected back to the original constraint space by assigning each subcarrier to the user with the largest sharing factor. This heuristic approach is suboptimal, and more importantly, is also computationally prohibitive, because it involves solving a large constrained convex optimization problem with $2MK$ variables with interval constraints and $K+1$ linear inequality constraints. Hence, the main focus of previous research have been on developing heuristic approaches with typical complexities in the order of $\mathcal{O}(MK^2)$ (e.g. [4] [7]).

Our approach, on the other hand, is based on a *Lagrangian relaxation* of the power constraints, instead of the *constraint relaxation* proposed previously. This relaxation retains the subcarrier assignment exclusivity constraints, but "dualizes" the power constraint and incorporates it into the objective function, allowing us to solve the dual problem instead. This dual optimization framework is shown to be less complex ($\mathcal{O}(MK)$ per iteration, with less than 10 iterations) and achieves relative optimality gaps that are less than 10^{-4} (i.e. achieving 99.9999% of the optimal solution) in typical scenarios, and thus actually allowing us to claim *practical optimality*. Note that the dual optimization approach is also studied in [10], but their focus has been on instantaneous rate optimization in DSL systems. Also note that during the review of this paper, it was brought to our attention that [11] and [12] have independently discovered a similar dual optimization framework to solve a weighted-sum continuous instantaneous rate maximization problem, which is similar to the algorithm in Section III-G.

II. System Model

We consider the downlink of a single OFDMA base station with K-subcarriers and M-users indexed by the set $\mathcal{K} = \{1,\ldots,k,\ldots,K\}$ and $\mathcal{M} = \{1,\ldots,m,\ldots,M\}$ (typically $K \gg M$) respectively. We assume an average transmit power of $\bar{P} > 0$, bandwidth B, and noise density N_0. The received signal vector for the mth user at the nth OFDM symbol is

$$\boldsymbol{y}_m[n] = \mathbf{G}_m[n]\mathbf{H}_m[n]\boldsymbol{x}_m[n] + \boldsymbol{w}_m[n] \tag{1}$$

where $\boldsymbol{y}_m[n]$ and $\boldsymbol{x}_m[n]$ are the K-length received and transmitted complex-valued signal vectors; $\mathbf{G}_m[n] = \text{diag}\left\{\sqrt{p_{m,1}[n]},\ldots,\sqrt{p_{m,K}[n]}\right\}$ is the diagonal gain allocation matrix; $\boldsymbol{w}_m[n] \sim \mathcal{CN}(\mathbf{0},\sigma_w^2\boldsymbol{I}_K)$ with noise variance $\sigma_w^2 = N_0B/K$ is the white zero-mean, circular-symmetric, complex Gaussian (ZMCSCG) noise vector; and $\mathbf{H}_m[n] = \text{diag}\{h_{m,1}[n],\ldots,h_{m,K}[n]\}$ is the diagonal channel response matrix, where

$$h_{m,k}[n] = \sum_{i=1}^{N_t} g_{m,i}[n]e^{-j2\pi\tau_i k\Delta f}. \tag{2}$$

are the complex-valued frequency-domain wireless channel fading random processes, given as the discrete-time Fourier transform of the N_t time-domain multipath taps $g_{m,i}[n]$ with time-delay τ_i and subcarrier spacing Δf. These taps are modeled as stationary and ergodic discrete-time random processes with tap powers $\sigma_{m,i}^2$, which we assume to be independent across the fading paths i and across users m. Since $g_{m,i}[n]$ is stationary and ergodic, so is $h_{m,k}[n]$. Hence, the distribution of $\boldsymbol{h}_m[n]$ is independent of n through stationarity, and we can replace time averages with ensemble averages in the problem formulations through ergodicity. In the subsequent discussion, we shall drop the index n when the context is clear for notational brevity.

We assume[1] that the time domain channel taps are independent ZMCSCG random variables $g_{m,i} \sim \mathcal{CN}(0,\sigma_{m,i}^2)$ with total power $\sigma_m^2 = \sum_{i=1}^{N_t}\sigma_{m,i}^2$. Then from (2), we have

$$\begin{aligned}\boldsymbol{h}_m &\sim \mathcal{CN}(\mathbf{0}_K,\mathbf{R}_{\boldsymbol{h}_m})\\ \mathbf{R}_{\boldsymbol{h}_m} &= \mathbf{W}\boldsymbol{\Sigma}_m\mathbf{W}^H\end{aligned} \tag{3}$$

where $\mathbf{W}$ is the $K \times N_t$ DFT matrix with entries $[\mathbf{W}]_{k,i} = e^{-j2\pi\tau_i k\Delta f}, k = -K/2-1,\ldots,K/2; i = 1,\ldots,N_t$ and $\boldsymbol{\Sigma}_m = \text{diag}\{\sigma_{m,1}^2,\ldots,\sigma_{m,N_t}^2\}$ is an $N_t \times N_t$ diagonal matrix of the time-domain path powers. Since we also assume that the fading for each user is independent, then the joint distribution of the stacked fading vector for all users $\boldsymbol{h} = [\boldsymbol{h}_1^T,\ldots,\boldsymbol{h}_M^T]^T$ is likewise a ZMCSCG random vector with distribution $\boldsymbol{h} \sim \mathcal{CN}(\mathbf{0}_{KM},\mathbf{R}_{\boldsymbol{h}})$ where $[\cdot]^T$ is the vector transpose and $\mathbf{R}_{\boldsymbol{h}}$ is the $KM \times KM$ block diagonal covariance matrix with $\mathbf{R}_{\boldsymbol{h}_m}$ as the diagonal block elements.

We let $\boldsymbol{\gamma}_m = [\gamma_{m,1},\ldots,\gamma_{m,k}]^T$ where $\gamma_{m,k} = |h_{m,k}|^2/\sigma_w^2$ denote the instantaneous channel-to-noise ratio (CNR) with mean $\bar{\gamma}_{m,k} = \sigma_m^2/\sigma_w^2$. Note that $\gamma_{m,k}$ for a particular subcarrier k and different users m are independent but not necessarily identically distributed (INID) exponential random variables; and for a particular user m and different subcarriers k are not independent but identically distributed (NIID) exponential random variables. Throughout the paper, we assume that the transmitter has perfect knowledge of $\boldsymbol{\gamma}_m$ for all users (perfect CSI assumption), and that the resource allocation decisions are made known to the users through an error-free control

[1] Although the results of this paper are applicable to any fading distribution, we shall prescribe a particular distribution for the fading channels for illustration purposes.

channel. The case where only partial channel knowledge is available is investigated in [13].

III. Ergodic Rate Maximization in OFDMA - Continuous Rates

A. Problem Formulation

The capacity of user m on subcarrier k assuming independent and identically distributed (IID) Gaussian signalling under IID Gaussian noise is given as (see e.g. [7])

$$R(p_{m,k}\gamma_{m,k}) = \log_2(1 + p_{m,k}\gamma_{m,k}) \quad \text{bps/Hz} \tag{4}$$

Denote by $\boldsymbol{p} = [\boldsymbol{p}_1^T, \cdots, \boldsymbol{p}_K^T]^T$ the vector of powers to be determined, where $\boldsymbol{p}_k = [p_{1,k}, \cdots, p_{M,k}]^T$. The exclusive subcarrier assignment restriction in OFDMA can be written as $\boldsymbol{p}_k \in \boldsymbol{\mathcal{P}}_k \subset \mathbb{R}_+^M$,

$$\boldsymbol{\mathcal{P}}_k \equiv \{\boldsymbol{p}_k \in \mathbb{R}_+^M | p_{m,k}p_{m',k} = 0; \forall m \neq m'\} \tag{5}$$

For notational convenience, we let $\boldsymbol{p} \in \boldsymbol{\mathcal{P}} \equiv \boldsymbol{\mathcal{P}}_1 \times \cdots \times \boldsymbol{\mathcal{P}}_K \subset \mathbb{R}_+^{MK}$ denote the space of allowable power vectors for all subcarriers. Since we assumed perfect CSI, we can consider the power allocation vector $\boldsymbol{p}(\boldsymbol{\gamma})$ as a function of the realization of the fading CNR of all users $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_M^T]^T$. The ergodic weighted sum capacity maximization problem is then

$$\begin{aligned} f^* = \max_{\boldsymbol{p}(\boldsymbol{\gamma})\in\boldsymbol{\mathcal{P}}} \quad & \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}} w_m \sum_{k\in\mathcal{K}} R\left(p_{m,k}\gamma_{m,k}\right)\right\} \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} p_{m,k}\right\} \leq \bar{P} \end{aligned} \tag{6}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator, w_m are positive constants such that $\sum_{m\in\mathcal{M}} w_m = 1$. Theoretically, varying these weights allows us to trace out the ergodic capacity region [14]; algorithmically, varying the weights allows us to prioritize the different users in the system and enforce certain notions of fairness[2]. A caveat for this ergodic weighted sum capacity formulation, however, is that the w_ms need to be held constant for a time period that allows the ergodicity property of the channels gains to kick in, which may hurt the fairness of the system. Fortunately, in next generation OFDMA implementations (e.g. IEEE 802.16e [1] and 3GPP-LTE [2]), the MAC layer hands down user-weights to the physical layer on a per-frame (or longer) basis. This is because holding weights constant for a period is beneficial from a system implementation complexity perspective, requiring less signaling and feedback overhead, while still enforcing fairness, albeit on a larger timescale. Thus, depending on the frame lengths (which in IEEE 802.16e can reach up to 20ms [1]) and the mobile speed, ergodicity can be assumed in a lot of cases within the frame, and the ergodic weighted sum capacity formulation is ideal in these scenarios. A comparison of the fairness in ergodic and instantaneous rate formulations, and the effect of the w_m's on overall system performance, however, is beyond the scope of this paper.

[2]Note that the choice of w_m is dependent on the system designer, and is typically handed down to the physical layer from a higher layer, e.g. the MAC. A possible choice is $w_m(n) = 1/R_m(n)$ where $R_m(n)$ is the average rate for user m so far at time n, which was shown to approximate proportional fairness [9]

B. Dual Optimization Framework

Note that the objective function in (6) is concave, but the constraint space $\boldsymbol{\mathcal{P}}$ is highly non-convex (it is in fact a discrete space), and is in general very difficult to solve. Fortunately, (6) is separable across the subcarriers, and is tied together only by the power constraint. In these problems, it is useful to approach the problem using duality principles [10] [15]. Let us write the Lagrangian

$$\begin{aligned} L(\boldsymbol{p}(\boldsymbol{\gamma}), \lambda) = \mathbb{E}_{\boldsymbol{\gamma}}&\left\{\sum_{m\in\mathcal{M}} w_m \sum_{k\in\mathcal{K}} R\left(p_{m,k}\gamma_{m,k}\right)\right\} \\ &+ \lambda\left(\bar{P} - \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} p_{m,k}\right\}\right) \end{aligned} \tag{7}$$

The dual problem is defined as

$$g^* = \min_{\lambda\geq 0} \Theta(\lambda) \tag{8}$$

where the dual objective is given by

$$\Theta(\lambda) = \max_{\boldsymbol{p}(\boldsymbol{\gamma})\in\boldsymbol{\mathcal{P}}} L(\boldsymbol{p}(\boldsymbol{\gamma}), \lambda) \tag{9a}$$

$$= \lambda\bar{P} + \max_{\boldsymbol{p}(\boldsymbol{\gamma})\in\boldsymbol{\mathcal{P}}} \sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}} [w_m R\left(p_{m,k}\gamma_{m,k}\right) - \lambda p_{m,k}]\right\} \tag{9b}$$

$$= \lambda\bar{P} + \sum_{k\in\mathcal{K}} \max_{\boldsymbol{p}_k(\boldsymbol{\gamma})\in\boldsymbol{\mathcal{P}}_k} \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}} [w_m R\left(p_{m,k}\gamma_{m,k}\right) - \lambda p_{m,k}]\right\} \tag{9c}$$

$$= \lambda\bar{P} + \sum_{k\in\mathcal{K}} \mathbb{E}_{\boldsymbol{\gamma}}\left\{\max_{\boldsymbol{p}_k(\boldsymbol{\gamma})\in\boldsymbol{\mathcal{P}}_k} \sum_{m\in\mathcal{M}} [w_m R\left(p_{m,k}\gamma_{m,k}\right) - \lambda p_{m,k}]\right\} \tag{9d}$$

$$= \lambda\bar{P} + K\mathbb{E}_{\boldsymbol{\gamma}_k}\left\{\max_{m\in\mathcal{M}}\left[\max_{p_{m,k}\geq 0} w_m R(p_{m,k}\gamma_{m,k}) - \lambda p_{m,k}\right]\right\} \tag{9e}$$

where (9a) is the dual objective; (9b) follows from the linearity of the expected value; (9c) follows from the fact that the power variables are separable across the subcarriers[3]; (9d) follows from the fact that the power variables are a function of each realization of $\boldsymbol{\gamma}$, allowing us to interchange the order of maximization and expected value; and (9e) follows from the fact that the channel gains are NIID across subcarriers and the exclusive subcarrier assignment constraint. We have reduced the problem to a per-subcarrier optimization, and since $K \gg M$, we have significantly decreased the computational burden.

[3]The separability is due to the fact that the exclusive subcarrier allocation constraint is enforced on a per-subcarrier basis (see (5)), and that the average power constraint that ties the power variables across subcarriers has been "dualized " into the Lagrangian objective function

The innermost maximization between the square brackets in (9e) has a simple closed-form expression for the optimal powers given as

$$\tilde{p}_{m,k}(\lambda) = \left[\frac{1}{\gamma_{0,m}(\lambda)} - \frac{1}{\gamma_{m,k}}\right]^+ \tag{10}$$

where $[x]^+ = \max(0, x)$ and $\gamma_{0,m}(\lambda) = \frac{\lambda \ln 2}{w_m}$, which is a simple "multi-level water-filling" power allocation with cut-off CNR $\gamma_{0,m}(\lambda)$, below which we do not transmit any power, and above which we transmit more power when the CNR $\gamma_{m,k}$ is higher.

Using (10) in (9e), the dual problem (8) can be written as

$$g^* = \min_{\lambda \geq 0} \left[\lambda \bar{P} + K \mathbb{E}_{\boldsymbol{\gamma}_k} \left\{ g_k(\boldsymbol{\gamma}_k, \lambda) \right\}\right] \tag{11}$$

$$g_k(\boldsymbol{\gamma}_k, \lambda) = \max_{m \in \mathcal{M}} \left\{ g_{m,k}(\gamma_{m,k}, \lambda) \right\} \tag{12}$$

where (12) is a max function over the m per-subcarrier dual functions

$$\begin{aligned} g_{m,k}(\gamma_{m,k}, \lambda) &= w_m R\left(\tilde{p}_{m,k}(\lambda)\gamma_{m,k}\right) - \lambda \tilde{p}_{m,k}(\lambda) \\ &= \left(\frac{w_m}{\ln 2} \ln\left(\frac{\gamma_{m,k}}{\gamma_{0,m}(\lambda)}\right) - \frac{w_m}{\ln 2} + \frac{\lambda}{\gamma_{m,k}}\right) \\ &\quad \times u(\gamma_{m,k} - \gamma_{0,m}(\lambda)) \end{aligned} \tag{13}$$

and where $u(x)$ is the unit (Heaviside) step function with $u(0) = 1$. Note that (13) is non-negative and is not differentiable at $g_{m,k}(\gamma_{0,m}(\lambda), \lambda) = 0$.

C. Numerical Evaluation of the Expected Dual

Computing the expectation in (11) in a straightforward manner involves an $M-$dimensional integral over the joint pdf of the $M-$length fading vector $\boldsymbol{\gamma}_k$, which is typically too complex to solve using direct numerical integration techniques (e.g. Gaussian quadrature) except for small M, e.g. 2 or 3, since this requires $\mathcal{O}(N^M)$ computations where N is the number of function evaluations required for a one-dimensional integral with the same accuracy [16]. However, if we can somehow compute a closed-form expression for the pdf of (12), then we can reduce the expectation to just a one-dimensional integral that is solvable in $\mathcal{O}(MN)$. Since $\gamma_{m,k}$ for different ms are INID, then (13) is likewise INID for different ms. Thus, (12) is the largest order statistic of INID random variables $g_{m,k}(\gamma_{m,k}, \lambda)$ with pdf [17, Sec. 5.2]

$$f_{g_k}(g_k) = \prod_{m \in \mathcal{M}} F_{g_{m,k}}(g_k) \left(\sum_{m \in \mathcal{M}} \frac{f_{g_{m,k}}(g_k)}{F_{g_{m,k}}(g_k)} \right) \tag{14}$$

where $F_{g_{m,k}}(g_{m,k})$ and $f_{g_{m,k}}(g_{m,k})$ are the cumulative distribution function (CDF) and probability density function (PDF) of $g_{m,k}(\gamma_{m,k}, \lambda)$, respectively.

In order to derive these distribution functions given the distribution $F_{\gamma_{m,k}}(\gamma_{m,k})$ of $\gamma_{m,k}$, we need an expression for the inverse function of $g_{m,k}(\gamma_{m,k}, \lambda)$ with respect to $\gamma_{m,k}$, which is given as (see Appendix A)

$$\check{\gamma}_{m,k}(g_{m,k}) = \frac{-\gamma_{0,m}(\lambda)}{W\left(-e^{-\left(g_{m,k}\frac{\ln 2}{w_m}+1\right)}\right)} u(g_{m,k}) \tag{15}$$

where $W(x)$ is the *Lambert-W* function, which is the solution to the transcendental equation $W(x)e^{W(x)} = x$. This function is ubiquitous in the physical sciences, and efficient algorithms have been developed for its computation [18]. Note that $\check{\gamma}_{m,k}(0) = \gamma_{0,m}(\lambda)$ as expected.

Using this expression for the root, we can then derive the cdf of $\gamma_{m,k}$ as [19]

$$F_{g_{m,k}}(g_{m,k}) = F_{\gamma_{m,k}}\left(\check{\gamma}_{m,k}(g_{m,k})\right) u(g_{m,k}) \tag{16}$$

The pdf is then given as the derivative of (16) with respect to $g_{m,k}$

$$\begin{aligned} f_{g_{m,k}}(g_{m,k}) &= F_{\gamma_{m,k}}\left(\gamma_{0,m}(\lambda)\right) \delta(g_{m,k}) \\ &\quad + \frac{f_{\gamma_{m,k}}\left(\check{\gamma}_{m,k}(g_{m,k})\right) \check{\gamma}^2_{m,k}(g_{m,k})}{\check{\gamma}_{m,k}(g_{m,k}) \frac{w_m}{\ln 2} - \lambda} u(g_{m,k}) \end{aligned} \tag{17}$$

where $\delta(x)$ is the Dirac delta function[4]. Finally, using (16) and (17) in (14) and then in (11), our dual problem can now be written as

$$g^* = \min_{\lambda \geq 0} \left[\lambda \bar{P} + K \int_0^\infty g_k f_{g_k}(g_k) dg_k\right] \tag{18}$$

D. Optimal Subcarrier and Power Allocation

Using standard duality arguments (see e.g. [15, Prop. 5.1.2]), the dual objective function in (18) can be shown to be convex and continuously differentiable in the single variable λ. Thus, we could simply take its derivative with respect to λ and set it to zero to find the optimum geometric multiplier λ^*. However, the derivative function requires $\mathcal{O}(M^2)$ computations due to the product terms in the pdf. Thus, it is more efficient to resort to derivative-free line search procedures that only need function evaluations, e.g. Golden-section or Fibonacci search [16].

Once we determine λ^*, we plug it back into the optimal power allocation function and arrive at the following simple user assignment and power allocation for each subcarrier k given as

$$m_k^* = \arg \max_{m \in \mathcal{M}} \left\{ w_m R\left(\tilde{p}_{m,k}(\lambda^*)\gamma_{m,k}\right) - \lambda^* \tilde{p}_{m,k}(\lambda^*) \right\} \tag{19}$$

$$p_{m,k}^* = \tilde{p}_{m,k}(\lambda^*) \mathbf{1}(m = m_k^*) \tag{20}$$

where $\mathbf{1}(x)$ is the indicator function, which evaluates to 1 if x is true and 0 if false.

Note that it is possible that the dual optimal powers do not satisfy the total power constraint. Hence, our final power allocation values should be multiplied by $\eta = \bar{P} / \mathbb{E}_{\boldsymbol{\gamma}} \left\{ \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} p_{m,k}^* \right\}$ which we plug back into the objective in (6) to arrive at our computed primal optimal value

$$\hat{f}^* = \mathbb{E}_{\boldsymbol{\gamma}} \left\{ \sum_{m \in \mathcal{M}} w_m \sum_{k \in \mathcal{K}}^{K} \log_2(1 + \eta \gamma_{m,k} p_{m,k}^*) \right\} \tag{21}$$

[4]Note that $F_{g_{m,k}}(g_{m,k})$ is discontinuous at $g_{m,k} = 0$ with $F_{g_{m,k}}(0^-) = 0$ and $F_{g_{m,k}}(0^+) = F_{\gamma_{m,k}}\left(\gamma_{0,m}(\lambda)\right)$.

E. Bound on the Relative Duality Gap

The following theorem provides a bound on the relative optimality gap which we can compute in order to assess how far we are from the optimal value.

Theorem 1: Let $f^* > 0$ and $g^* > 0$ given in (6) and (18) be the optimal values of the primal and dual problems, and let $\hat{f}^* > 0$ given in (21) be the computed feasible primal value. Then the relative duality (optimality) gap can be bounded as

$$0 \leq \frac{g^* - f^*}{f^*} \leq \frac{g^* - \hat{f}^*}{\hat{f}^*} \tag{22}$$

Proof: The left inequality follows directly from the positivity of f^* and the weak duality theorem [15, Prop. 5.1.3. p. 495], which states that $g^* \geq f^*$. The right inequality is because $\hat{f}^* \leq f^*$, since $\hat{f}^*$ is a feasible primal value and f^* is the optimal feasible primal value. ∎

When the power constraints are met tightly, i.e. $\mathbb{E}\left\{\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} p^*_{m,k}\right\} = \bar{P}$, then the duality gap is zero, and thus solving the dual problem also solves the primal problem. In our numerical results, the power constraints are met almost exactly, resulting in relative optimality gaps that are practically zero ($< 10^{-4}$). Thus, our approach can, for all practical purposes, be considered an optimal solution to the problem. This fortuitous phenomenon is brought about primarily by the separability of the problem, and furthermore by the fact that we have K separable terms (which is typically large) and only a single constraint (average power constraint). This problem structure has been shown to be particularly suitable to dual optimization approaches, and has been noted in [10], and more generally treated with a theoretical justification in [15, Sec. 5.1.6].

F. Complexity Analysis

Once we determine λ^* by solving (18), we do not need to update it as long as the statistics of the fading channel vector $\boldsymbol{\gamma}$ remain the same. Thus, the complexity of resource allocation requires an initial $\mathcal{O}(INM)$ computations to determine λ^*, where I is the number of iterations for the line search procedure to converge, and N is the number of function evaluations to compute the dual objective integral. The allocation in (19)-(20) needs $\mathcal{O}(MK)$ computations per symbol.

G. Instantaneous Weighted Sum Rate Maximization

Although we have focused on the ergodic rate maximization problem, our duality framework can be simplified to solve the instantaneous rate maximization problem given as

$$\begin{aligned} f^*_{\text{inst}} = \max_{\boldsymbol{p}\in\boldsymbol{\mathcal{P}}} \quad & \sum_{m\in\mathcal{M}} w_m \sum_{k\in\mathcal{K}} R\left(p_{m,k}\gamma_{m,k}\right) \\ \text{s.t.} \quad & \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} p_{m,k} \leq \bar{P} \end{aligned} \tag{23}$$

and is essentially identical to the problem considered in [7]. We use the dual optimization approach[5], where the dual problem can be derived similarly as the ergodic case

$$g^*_{\text{inst}} = \min_{\lambda\geq 0}\left[\lambda\bar{P} + \sum_{k\in\mathcal{K}} \max_{m\in\mathcal{M}} \left\{w_m R\left(\tilde{p}_{m,k}(\lambda)\gamma_{m,k}\right) - \lambda\tilde{p}_{m,k}(\lambda)\right\}\right] \tag{24}$$

where $\tilde{p}_{m,k}(\lambda)$ is the same power allocation function given in (10). Note the similarity to the ergodic case (c.f. (8)-(9e)), where the primary difference is that the expected values are no longer present. Using a similar line search procedure, we can find the optimal λ^*_{inst} and end up with the same optimal subcarrier and power allocation functions as in (19)-(20). One subtle, albeit important difference, is that in instantaneous maximization, the optimal λ^*_{inst} is dependent on each channel realization $\boldsymbol{\gamma}$, and thus needs to be computed every time the channel changes. This is in contrast to the ergodic maximization case where the λ^* depends on the *distribution function* of the channel $f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$, and thus needs to be computed only when the statistics of the channel has changed. Thus, although the initialization for the ergodic maximization is more complex, the per-symbol resource allocation complexity ends up to be less complex than the instantaneous optimization case. Furthermore, because the total power in each time instant is constrained to be less than or equal to $\bar{P}$ in the instantaneous case, there is no flexibility of allowing the total power in each time instant to vary (while still maintaining the average power constraint across time) unlike the ergodic maximization case.

H. Constant Power Allocation

It has been established in previous research that constant power allocation actually performs as well as optimal water-filling, esp. in high SNR cases [20]. Under the constant power allocation assumption, the power is set to $\bar{P}/K$, and the subcarrier allocation is simplified to

$$m^*_k = \arg\max_{m\in\mathcal{M}}\left\{w_m R\left(\frac{\bar{P}}{K}\gamma_{m,k}\right)\right\} \tag{25}$$

IV. Ergodic Rate Maximization in OFDMA - Discrete Rates

A. Problem Formulation

In this section, we derive resource allocation algorithms for the practically relevant case of when only a discrete number of modulation and coding levels are available (i.e. adaptive modulation and coding). In the discrete rate case, the data rate of the kth subcarrier for the mth user can be given by the staircase function

$$R^d_{m,k}(p_{m,k}\gamma_{m,k}) = \begin{cases} r_0, & \eta_0 \leq p_{m,k}\gamma_{m,k} < \eta_1 \\ r_1, & \eta_1 \leq p_{m,k}\gamma_{m,k} < \eta_2 \\ \vdots, & \vdots \\ r_{L-1}, & \eta_{L-1} \leq p_{m,k}\gamma_{m,k} < \eta_L \end{cases} \tag{26}$$

where $\{\eta_l\}_{l\in\mathcal{L}}$, $\mathcal{L} = \{0,\ldots,L-1\}$, are the SNR boundaries which define a particular code-rate and constellation pair combination that result in r_l data bits per transmission with a predefined target bit error rate (BER), and where $r_l \geq 0$, $r_{l+1} > r_l$, $r_0 = 0$, $\eta_0 = 0$, and $\eta_L = \infty$.

[5]Note that [11] and [12] have independently come up with a similar dual approach to this instantaneous rate maximization problem.

The average discrete weighted-sum rate maximization can then be formulated as

$$\begin{aligned} f_d^* = \max_{\boldsymbol{p}\in\boldsymbol{\mathcal{P}}} \quad & \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}} w_m \sum_{k\in\mathcal{K}} R_{m,k}^d(p_{m,k}\gamma_{m,k})\right\} \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{\gamma}}\left\{\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} p_{m,k}\right\} \le \bar{P} \end{aligned} \tag{27}$$

B. Dual Optimization Framework

Following a similar dual framework in Section III-B, we arrive at the dual objective (c.f. (9e))

$$\Theta_d(\lambda) = \lambda\bar{P} + K\mathbb{E}_{\boldsymbol{\gamma}}\left\{\max_{m\in\mathcal{M}}\left[\max_{p_{m,k}\ge 0}\left(w_m R_{m,k}^d(p_{m,k}\gamma_{m,k}) - \lambda p_{m,k}\right)\right]\right\} \tag{28}$$

The main difference of the inner maximization in this case with the continuous rate case in (9e) is that $R_{m,k}^d(p_{m,k}\gamma_{m,k})$ is a discontinuous function; hence, simple differentiation to arrive at the optimal solution is not feasible. However, note that we can divide the feasible region for $p_{m,k}$ (i.e. the non-negative real line) into L segments $\mathcal{R}_+^l = \left[\frac{\eta_l}{\gamma_{m,k}}, \frac{\eta_{l+1}}{\gamma_{m,k}}\right), l \in \mathcal{L}$. Since λ and $p_{m,k}$ are both non-negative, we have

$$\begin{aligned} w_m R_{m,k}^d(p_{m,k}\gamma_{m,k}) - \lambda p_{m,k} &= w_m r_l - \lambda p_{m,k} \\ &\le w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}}, \forall p_{m,k} \in \mathcal{R}_+^l \end{aligned}$$

Thus, there are only L candidate power allocation functions $\tilde{p}_{m,k}^d \in \left\{\frac{\eta_0}{\gamma_{m,k}}, \ldots, \frac{\eta_{L-1}}{\gamma_{m,k}}\right\}$ from which we need to choose the one that maximizes $w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}}$, i.e. $\tilde{p}_{m,k}^d = \eta_{l_{m,k}^*}/\gamma_{m,k}$ where

$$l_{m,k}^* \in \arg\max_{l\in\mathcal{L}} w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}} \tag{29}$$

This in turn also gives us the rate allocation $\tilde{R}_{m,k}^d = r_{l_{m,k}^*}$.

A straightforward computation of (29) would require $\mathcal{O}(L)$ complexity. However, if we assume that the discrete rate function $R^d(p_{m,k}\gamma_{m,k})$ is concave[6], we can reduce the complexity of finding the power allocation function by noticing that (29) is equivalent to (see Appendix B for a proof)

$$l_{m,k}^* = \left\{l \in \mathcal{L} : \frac{\lambda}{w_m\gamma_{m,k}} \in \left[\frac{r_{l+1}-r_l}{\eta_{l+1}-\eta_l}, \frac{r_l - r_{l-1}}{\eta_l - \eta_{l-1}}\right)\right\} \tag{30}$$

where with slight abuse of notation, we define $(r_0 - r_{-1})/(\eta_0 - \eta_{-1}) \equiv \infty$. This can be interpreted geometrically by treating $\frac{\lambda}{w_m\gamma_{m,k}}$ as a slope value for which we are looking for an interval of consecutive slope values for which it belongs (see [21] for a similar interpretation for single-user discrete multitone systems). Since the set of rates and SNR region boundaries r_l and η_l are predefined in a communications system, we can store the set of slopes into a lookup table, thereby reducing the complexity of finding the optimal powers to a single table lookup operation.

[6]Concavity for this discontinuous staircase function simply means that the slopes when "connecting the dots" of the edges of the staircase are non-increasing. This assumption is quite reasonable, and is applicable to uncoded square QAM constellations [20], DMT bit allocation using SNR gap approximation [21], and in coded modulation using empirical BER values [4, Fig. 1].

Finally, we can write the discrete rate maximization dual problem as

$$g_d^* = \min_{\lambda\ge 0} \lambda\bar{P} + K\mathbb{E}_{\boldsymbol{\gamma}_k}\left\{g_k^d(\boldsymbol{\gamma}_k, \lambda)\right\} \tag{31}$$

$$g_k^d(\boldsymbol{\gamma}_k, \lambda) = \max_{m\in\mathcal{M}}\left\{g_{m,k}^d(\gamma_{m,k}, \lambda)\right\} \tag{32}$$

where (32) is a max function over the m per-subcarrier dual functions given as

$$g_{m,k}^d(\gamma_{m,k}, \lambda) = \max_{l\in\mathcal{L}}\left\{w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}}\right\} \tag{33}$$

Note that despite the negative term in (33), $g_{m,k}^d(\gamma_{m,k}, \lambda)$ is always non-negative. This is because both r_0 and η_0 are equal to zero, hence the lowest possible value for the objective is zero.

C. Numerical Evaluation of the Expected Dual

Similar to the continuous rate case (cf. III-C), we require an M-dimensional integral to compute the expectation in (31) in a straightforward manner. Thus, we proceed similarly as the continuous rate case to derive a closed-form expression for the pdf of g_k^d in (32) and reduce the computation to just a single integral. The key to the derivation is to derive the CDF and PDF of (33), and use the same formula used in the continuous rate case for the maximum order statistic given in (14). Making the same assumption that the discrete rate function (26) is concave, the CDF and PDF are given as (see Appendix C for a derivation)

$$\begin{aligned} F_{g_{m,k}^d}(g_{m,k}^d) &= u(g_{m,k}^d)F_{\gamma_{m,k}}(s_1) \\ &\quad + \sum_{l\in\mathcal{L}\setminus 0}\left[F_{\gamma_{m,k}}\left(\min\left(h_l(g_{m,k}^d), s_{l+1}\right)\right) - F_{\gamma_{m,k}}(s_l)\right]^+ \end{aligned} \tag{34}$$

$$\begin{aligned} f_{g_{m,k}^d}(g_{m,k}^d) &= \delta(g_{m,k}^d)F_{\gamma_{m,k}}(s_1) \\ &\quad + \sum_{l\in\mathcal{L}\setminus 0}\mathbf{1}\left(h_l(g_{m,k}^d)\in\mathcal{S}\right) f_{\gamma_{m,k}}\left(h_l(g_{m,k}^d)\right)\frac{h_l^2(g_{m,k}^d)}{\lambda\eta_l} \end{aligned} \tag{35}$$

where

$$h_l(g_{m,k}^d) \equiv \frac{\lambda\eta_l}{[w_m r_l - g_{m,k}^d]^+} \tag{36}$$

$$s_l \equiv \frac{\lambda(\eta_l - \eta_{l-1})}{w_m(r_l - r_{l-1})} \tag{37}$$

Figs. 1-2 shows an example of the cdf and pdf for $w_m = 1$, $\lambda = 1$, $\bar{\gamma} = 20$ dB, and discrete rate function given in Section V. We also plot the L individual terms that sum to the functions, giving us better insight into how these functions are derived. We also superimposed empirical curves generated using Monte-Carlo generation for verification. It can be seen that for the pdf, only certain ranges of $g_{m,k}^d$ actually "activate" a particular component of a rate level l, which is analytically given as the range in the indicator function in (35). Furthermore, once the level l is activated for a given range of $g_{m,k}^d$, we simply take the derivative of $F_{\gamma_{m,k}}(h_l(g_{m,k}^d))$ with respect to $g_{m,k}^d$ which is given as the terms multiplied to the indicator function in (35).

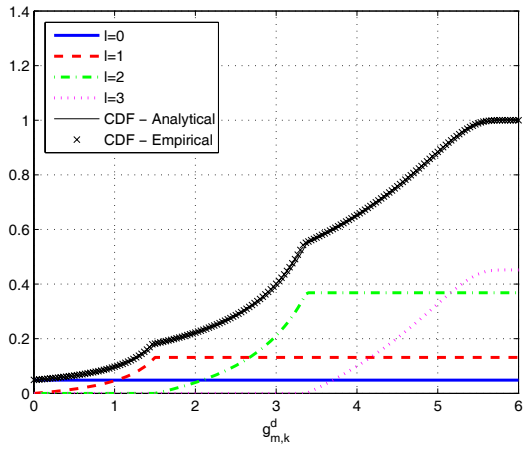

Fig. 1. Analytical and empirical CDF (34) of the discrete rate marginal dual function $g^d_{m,k}$ (33) with the L individual terms that sum to the CDF, corresponding to each discrete rate level.

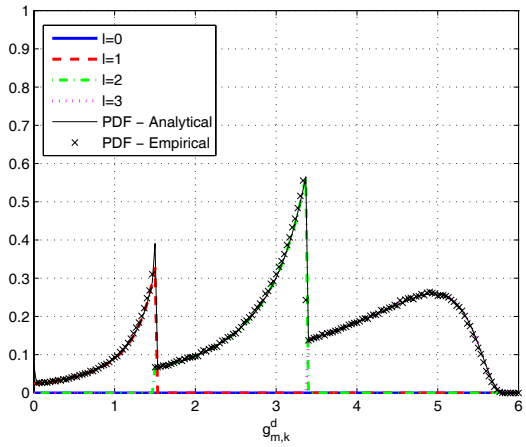

Fig. 2. Analytical and empirical PDF (34) of the discrete rate marginal dual function $g^d_{m,k}$ (33) with the L individual terms that sum to the PDF, corresponding to each discrete rate level.

Finally, using (34)-(35) in (14), then in (31), the dual problem can be written as

$$g^*_d = \min_{\lambda \geq 0} \left[\lambda \bar{P} + K \int_0^\infty g^d_k f_{g^d_k}(g^d_k) dg^d_k \right] \tag{38}$$

D. Optimal Discrete Rate, Subcarrier, and Power Allocation

The optimum solution to (38) denoted by λ^* can be found using similar line search techniques. The optimal subcarrier, rate, and power allocation is then determined using λ^* as

$$m^*_k = \arg \max_{m \in \mathcal{M}} w_m r_{l^*_{m,k}} - \lambda^* \frac{\eta_{l^*_{m,k}}}{\gamma_{m,k}} \tag{39}$$

$$R^*_{m,k} = r_{l^*_{m,k}} \mathbf{1}(m = m^*_k) \tag{40}$$

$$p^*_{m,k} = \frac{\eta_{l^*_{m,k}}}{\gamma_{m,k}} \mathbf{1}(m = m^*_k) \tag{41}$$

where $l^*_{m,k}$ is given by (30) with $\lambda = \lambda^*$. An upper bound on the relative duality gap of this algorithm can be derived similar to Section III-E. The complexity analysis is also similar to III-F, except for the additional $\mathcal{O}(L)$ factor to compute the dual objective (38), giving an initialization complexity of $\mathcal{O}(INML)$; and the additional $\mathcal{O}(\log(L))$ for the table lookup operation in (30), giving a resource allocation complexity of $\mathcal{O}(MK \log(L))$.

The instantaneous discrete rate maximization algorithm can also be derived by solving for the optimal instantaneous geometric multiplier λ^*_{inst} using (31) without the expectation and using the actual CNR vector $\boldsymbol{\gamma}$. The allocation rules are also given by (19)-(20) using the multiplier λ^*_{inst}. A further simplification is to assume constant power allocation, where the user selection is

$$m^*_k = \arg \max_{m \in \mathcal{M}} w_m R^d_{m,k} \left(\frac{\bar{P}}{K} \gamma_{m,k} \right) \tag{42}$$

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Subcarriers	128
Used Subcarriers	76
Bandwidth	1.25 MHz
Sampling Freq.	1.92 MHz
CP Length	6 samples

V. NUMERICAL RESULTS

We consider an OFDMA system based on a 3GPP-LTE downlink [2] with parameters given in Table I. We simulate the frequency-selective Rayleigh fading channel using the ITU-Vehicular A power delay profile (PDP) [22] which has a root-mean-square delay spread of $\sigma_\tau = 0.37\mu s$ and 50% coherence bandwidth of $B_c = 1/(5\sigma_\tau) \approx 540$ kHz. Each user's complex Gaussian channel vector realization $\boldsymbol{h}_m$ is then generated independently using (3) according to the PDP.

In Fig. 3, we compare the capacity regions for the continuous rate allocation case with 2 users using $10,000$ channel realizations and varying w_1 between 0 and 1, and setting $w_2 = 1 - w_1$. We see that ergodic rates maximization has better performance than the other methods due to its ability to exploit the temporal dimension. The gain is also more pronounced for low SNRs and more disparate user weights, which is analogous to previous studies in adaptive modulation, e.g. [5] [20], which concluded that the exploitation of the additional temporal dimension through the ergodic formulation is most useful when other degrees of freedom have been significantly curtailed. Fig. 4 shows the sum capacity as the number of users M is increased. We ran 500 frames with 1000 symbols per frame, where we draw a random realization of the normalized user weights w_m and hold it constant for each frame. We see the effect of *multiuser diversity* in that the capacity is actually increasing as the number of users increase. The gain of ergodic rates over the other methods diminish as we increase M, which is consistent with [5].

In Table II, we present other relevent metrics for the continuous rate maximization algorithms. For the ergodic rate maximization, the first main column indicates the average number of function evaluations required to numerically compute the integration of (18) with a tolerance of 10^{-6}, and the second main column indicates the average number of Golden-section search iterations to solve for λ^* in the dual

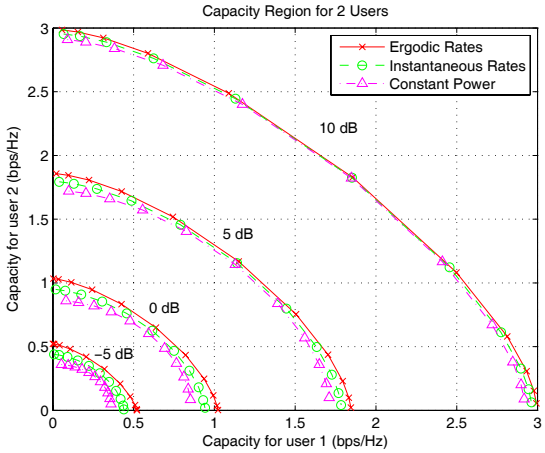

Fig. 3. 2-user capacity region for ergodic and instantaneous continuous rate maximization and constant power allocation.

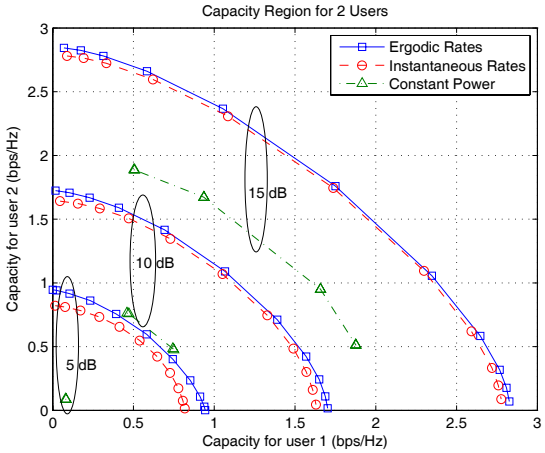

Fig. 5. 2-User Capacity Region for ergodic and instantaneous discrete rate maximization.

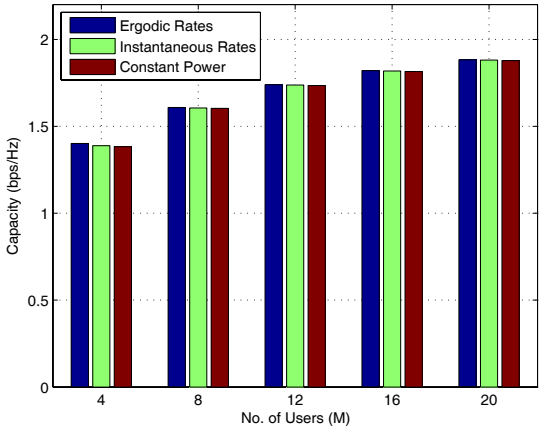

Fig. 4. Sum continuous rates for different numbers of users.

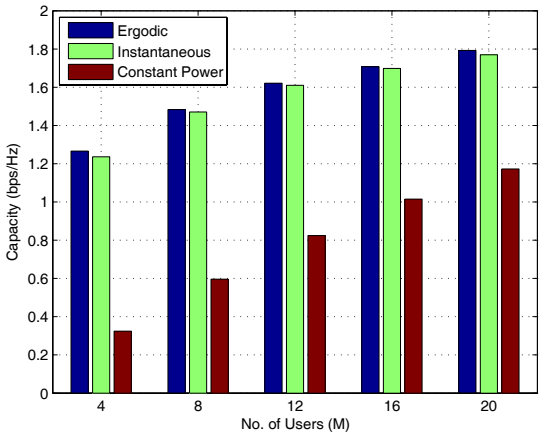

Fig. 6. Sum discrete rates for different numbers of users.

problem (11) with a tolerance of 10^{-4}. The second column for instantaneous rate maximization is the average number of iterations for each channel realization. The third column for both cases is the relative duality gap upper bound computed by (22). Note that the duality gaps are negligible, and thus both algorithms can be considered optimal.

For the discrete rate allocation results, we assume a Grey-coded square 2^{r_l}-QAM modulation scheme, where the BER in AWGN can be approximated to within 1-dB for $r_l \geq 2$ and BER $\leq 10^{-3}$ by BER $\approx 0.2e^{\frac{-1.6p_{m,k}\gamma_{m,k}}{2^{r_l}-1}}$ [20]. Fig. 5 shows the results of resource allocation, with rate set $r_l \in \{0, 2, 4, 6\}$ bits with SNR thresholds $\eta_l \in \{-\infty, 9.97, 16.96, 23.19\}$ dB computed using a BER constraint of 10^{-3}. Note that we assume that channel coding is not present in this case for simplicity, but since the framework merely needs the SNR thresholds and rate values, our results also apply to the coded case as long as the discrete rate function is concave.

Note that the general trends are similar to the continuous rate case, except that the advantage for ergodic rates is much more pronounced, and a large loss is incurred by the constant power allocation case. This is due to the big loss of freedom in the rate allocation (limited to just 4 rates in contrast to an infinite number of rates in the continuous case), which when coupled with constant power allocation results in a huge loss in performance. Note that increasing degrees of freedom in other dimensions, e.g. having more users or subcarriers will decrease this performance loss [5]. Fig. 6 shows the sum rates as the number of users is increased for the three different methods using a similar simulation setup as in the continuous rates case. We see similar trends as in the continuous rates case, but also with more pronounced gains for the ergodic rates case. Table III shows the average number of iterations and the relative optimality gaps for the discrete rate allocation algorithms. Note that the number of function evaluations are higher, due primarily to the discontinuities in the cdf and pdf functions (see Figs. 1-2).

Table IV shows the resource allocation complexity order of the different algorithms. If we use the average numbers given in Tables II and III, the ergodic rate algorithms are less complex than the instantaneous rate algorithms per symbol on average, as long as the rate of change of the channel fading statistics (roughly at the rate of change of slow fading, e.g. Log-normal shadowing) is much lower than the rate of change of the actual channel realizations (roughly at the rate of fast fading, e.g. Rayleigh fading), such that the initialization

TABLE II
RELEVANT METRICS FOR THE CONTINUOUS RATE RESOURCE ALLOCATION ALGORITHMS

Metric	No. of Func. Eval. (N)			No. of Iterations (I)			Relative Gap ($\times 10^{-6}$)		
SNR	5 dB	10 dB	15 dB	5 dB	10 dB	15 dB	5 dB	10 dB	15 dB
Ergodic Rates	47.91	50.09	53.73	8.091	7.727	7.727	7.936	5.462	5.444
Instantaneous Rates	–	–	–	8.344	8.333	8.539	.0251	.0226	.0159

TABLE III
RELEVANT METRICS FOR THE DISCRETE RATE RESOURCE ALLOCATION ALGORITHMS

Metric	No. of Func. Eval. (N)			No. of Iterations (I)			Relative Gap ($\times 10^{-4}$)		
SNR	5 dB	10 dB	15 dB	5 dB	10 dB	15 dB	5 dB	10 dB	15 dB
Ergodic Rates	62.09	91.55	133.0	9.818	10.55	9.909	.8711	.9507	.5322
Instantaneous Rates	–	–	–	17.24	17.20	17.30	3.602	1.038	.3996

TABLE IV
COMPLEXITY FOR THE RESOURCE ALLOCATION ALGORITHMS. M-NO. OF USERS, K-NO. OF SUBCARRIERS, L-NO. OF DISCRETE RATES, N-NO. OF FUNCTION EVALUATIONS FOR INTEGRATION, I-NO. OF LINE SEARCH ITERATIONS.

Algorithm	Initialization	Runtime
Cont. Ergodic Rates	$\mathcal{O}(INM)$	$\mathcal{O}(MK)$
Cont. Instantaneous Rates	–	$\mathcal{O}(IMK)$
Cont. Constant Power	–	$\mathcal{O}(MK)$
Disc. Ergodic Rates	$\mathcal{O}(INML)$	$\mathcal{O}(MK\log(L))$
Disc. Instantaneous Rates	–	$\mathcal{O}(IMK\log(L))$
Disc. Constant Power	–	$\mathcal{O}(MK\log(L))$

is performed less often. One caveat, however, is that the ergodic rate algorithms require information on the channel fading distribution function, which need an additional level of complexity and feedback overhead. Furthermore, the peak-to-average power ratio of the power allocation in the ergodic rates case is typically higher than for instantaneous rates, and even more so for constant power allocation.

VI. CONCLUSION

We derived optimal resource allocation algorithms for continuous and discrete ergodic weighted sum rate maximization in OFDMA systems based on a dual optimization framework with per-symbol complexity of $\mathcal{O}(MK)$ per iteration. The algorithms are shown to achieve relative optimality gaps of less than 10^{-4} in practical scenarios. We have also shown that ergodic rate maximization is actually less complex per symbol than instantaneous rate maximization, and thus presents an attractive communication performance vs. complexity tradeoff. The most gains in ergodic maximization occur at low SNRs and for discrete rate cases, primarily because of decreased degrees of freedom in these scenarios.

APPENDIX A
DERIVATION OF THE INVERSE FUNCTION (15) OF $g_{m,k}$ (13)

Since $g_{m,k}$ for $\gamma_{m,k} \geq \gamma_{0,m}$ is monotonically increasing and non-negative, there exists a unique inverse function. After some algebraic manipulation, we have

$$-\frac{\gamma_{0,m}(\lambda)}{\gamma_{m,k}} e^{-\frac{\gamma_{0,m}(\lambda)}{\gamma_{m,k}}} = -e^{-\left(g_{m,k}\frac{\ln 2}{w_m}+1\right)} \tag{43}$$

Observe that this is in the form of the Lambert-W function $W(x)$ [18], which is the solution to $W(x)\exp(W(x)) = x$. Thus, we can write

$$W\left(-e^{-\left(g_{m,k}\frac{\ln 2}{w_m}+1\right)}\right) = -\frac{\gamma_{0,m}(\lambda)}{\gamma_{m,k}} \tag{44}$$

which when solved for $\gamma_{m,k}$ gives us (15).

APPENDIX B
PROOF OF (30)

Observe that (29) implies

$$w_m r_{l^*_{m,k}} - \frac{\lambda\eta_{l^*_{m,k}}}{\gamma_{m,k}} \geq w_m r_l - \frac{\lambda\eta_l}{\gamma_{m,k}}, \quad \forall l \in \mathcal{L} \setminus l^*_{m,k} \tag{45}$$

After some algebraic manipulation, (45) can be written as

$$\frac{r_{\overline{l}} - r_{l^*_{m,k}}}{\eta_{\overline{l}} - \eta_{l^*_{m,k}}} \leq \frac{\lambda}{w_m\gamma_{m,k}} < \frac{r_{l^*_{m,k}} - r_{\underline{l}}}{\eta_{l^*_{m,k}} - \eta_{\underline{l}}}$$

for all $\overline{l} > l^*_{m,k}$ and for all $\underline{l} < l^*_{m,k}$, which is equivalent to

$$\max_{l>l^*_{m,k}} \frac{r_l - r_{l^*_{m,k}}}{\eta_l - \eta_{l^*_{m,k}}} \leq \frac{\lambda}{w_m\gamma_{m,k}} < \min_{l<l^*_{m,k}} \frac{r_{l^*_{m,k}} - r_l}{\eta_{l^*_{m,k}} - \eta_l} \tag{46}$$

Since the slope $\Delta r/\Delta\eta$ is non-increasing for a concave function, we arrive at (30).

APPENDIX C
DERIVATION OF THE CDF (34) AND PDF (35) OF $g^d_{m,k}$ (33)

First, we use (30) to get

$$\begin{aligned} P(l^*_{m,k} = l) &= P\left(\frac{r_{l+1} - r_l}{\eta_{l+1} - \eta_l} \leq \frac{\lambda}{w_m\gamma_{m,k}} < \frac{r_l - r_{l-1}}{\eta_l - \eta_{l-1}}\right) \\ &= P\left(s_l < \gamma_{m,k} \leq s_{l+1}\right) \end{aligned} \tag{47}$$

where s_l is defined in (37). Then, using the law of total probability [19], we have

$$\begin{aligned} F_{g^d_{m,k}}(g^d_{m,k}) &= \sum_{l\in\mathcal{L}} P(l^*_{m,k} = l) \\ &\quad \times P\left(\max_{l'\in\mathcal{L}} w_m r_{l'} - \lambda\frac{\eta_{l'}}{\gamma_{m,k}} \leq g^d_{m,k} \middle| l^*_{m,k} = l\right) \\ &= \sum_{l\in\mathcal{L}} P(l^*_{m,k} = l) \\ &\quad \times P\left(w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}} \leq g^d_{m,k} \middle| s_l < \gamma_{m,k} \leq s_{l+1}\right) \end{aligned} \tag{48}$$

$$
\begin{aligned}
F_{g^d_{m,k}}(g^d_{m,k}) &= u(g^d_{m,k})P(l^*_{m,k}=0) + \sum_{l\in\mathcal{L}\setminus 0} P(l^*_{m,k}=l)\frac{P\left(\gamma_{m,k}\le h_l(g^d_{m,k}), s_l\le\gamma_{m,k}\le s_{l+1}\right)}{P(l^*_{m,k}=l)} \\
&= u(g^d_{m,k})P(\gamma_{m,k}\le s_1) + \sum_{l\in\mathcal{L}\setminus 0} P\left(s_l\le\gamma_{m,k}\le \min\left(h_l(g^d_{m,k}), s_{l+1}\right)\right) \\
&= u(g^d_{m,k})F_{\gamma_{m,k}}(s_1) + \sum_{l\in\mathcal{L}\setminus 0}\left[F_{\gamma_{m,k}}\left(\min\left(h_l(g^d_{m,k}), s_{l+1}\right)\right) - F_{\gamma_{m,k}}(s_l)\right]^+
\end{aligned} \tag{49}
$$

Note that since $\lambda\frac{\eta_l}{\gamma_{m,k}}$ is non-negative, then if $w_m r_l - g^d_{m,k}$ is negative, $P\left(w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}} \le g^d_{m,k}\right) = 1$. Hence, we can write $P\left(w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}} \le g^d_{m,k}\right) = P\left(\gamma_{m,k}\le h_l(g^d_{m,k})\right)$ where $h_l(g^d_{m,k})$ is defined in (36), and where we safely defined $\frac{x}{0}=\infty, \forall x>0$. However, for $l=0$, we have $r_l=0$ and $\eta_l=0$, and $P\left(w_m r_l - \lambda\frac{\eta_l}{\gamma_{m,k}} \le g^d_{m,k}\right)$ is always one since $g^d_{m,k}\ge 0$. We can now simplify (48) as shown in (49) at the top of the page. Finally, the pdf (35) is the derivative of (49) with respect to $g^d_{m,k}$.

REFERENCES

[1] *Air Interface for Fixed and Mobile Broadband Wireless Access Systems*, IEEE Std. 802.16e-2005, Feb. 2006.

[2] *3rd Generation Partnership Project, Technical Specification Group Radio Access Network; Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)*, 3GPP Std. TR 25.814 v. 7.0.0, 2006.

[3] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[4] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1150–1158, Nov. 2003.

[5] J. Jang, K. B. Lee, and Y.-H. Lee, "Frequency-time domain transmit power adaptation for OFDM systems in multiuser environment," *IEE Electron. Lett.*, vol. 38, no. 25, pp. 1754–1756, 2002.

[6] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE Veh. Technol. Conf.*, May 2000, pp. 1085–1089.

[7] L. Hoo, B. Halder, J. Tellado, and J. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: Asymptotic FDMA capacity region and algorithms," *IEEE Trans. Commun.*, vol. 52, no. 6, pp. 922–930, June 2004.

[8] Z. Shen, J. Andrews, and B. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[9] G. Song and Y. Li, "Cross-Layer optimization for OFDM wireless networks part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[10] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, July 2006.

[11] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. IEEE International Symposium Inform. Theory*, July 2006, pp. 1394–1398.

[12] Y. Yu, X. Wang, and G. B. Giannakis, "Channel-adaptive congestion control and OFDMA scheduling for hybrid wireline-wireless networks," *IEEE Trans. Wireless Commun.*, to be published.

[13] I. C. Wong and B. L. Evans, "Optimal resource allocation in OFDMA systems with imperfect channel knowledge," *IEEE Trans. Commun.*, to be published.

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

[15] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Boston: Athena Scientific, 1999.

[16] W. H. Press, *Numerical Recipes in C*. Cambridge, UK: Cambridge University Press Cambridge, 1992.

[17] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. New York: John Wiley, 2003.

[18] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert-W function," *Advances Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[19] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 2002.

[20] S. T. Chung and A. Goldsmith, "Degrees of freedom in adaptive modulation: A unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, Sep. 2001.

[21] B. S. Krongold, K. Ramchandran, and D. L. Jones, "Computationally efficient optimal power allocation algorithms for multicarrier communication systems," *IEEE Trans. Commun.*, vol. 48, no. 1, pp. 23–27, 2000.

[22] *Selection procedures for the choice of radio transmission technologies for the UMTS*, ETSI Std. TR 101 112 v. 3.2.0, 1998.

Ian C. Wong received the BS degree in electronics and communications engineering (*magna cum laude*) from the University of the Philippines, Diliman, Quezon City, Philippines in 2000, and the MS and PhD degrees in electrical engineering from the University of Texas at Austin, Austin, TX USA, in 2004 and 2007, respectively.

From 2000-2002, he was a research scientist at the Advanced Science and Technology Institute in Diliman, Quezon City, Philippines where he was awarded the best male employee of the year for 2001. He has held summer internship positions at National Instruments and Freescale Semiconductor, both in Austin, TX USA from 2003-2006. He is currently a systems engineer in the advanced technology division of Freescale Semiconductor, Austin, TX USA.

Dr. Wong has 1 patent pending, 2 standards contributions, and 16 accepted/published peer-reviewed journal and conference papers. He was awarded the Texas Telecommunications Engineering Consortium Fellowship in 2003-2004, and the Wireless Networking and Communications Group student leadership award in 2007. His research interests include statistical signal processing and optimization for wired and wireless broadband communication systems.

Brian L. Evans (M'87-SM'97) received the BS degree in electrical engineering and computer science from the Rose-Hulman Institute of Technology, Terre Haute, IN USA, in 1987, and the MS and PhD degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA USA, in 1988 and 1993, respectively.

From 1993 to 1996, he was a post-doctoral researcher at the University of California, Berkeley. In 1996, he joined the faculty at The University of Texas at Austin. He currently holds the rank of Professor.

Prof. Evans has published more than 170 peer-reviewed conference and journal papers. He is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a member of the Design and Implementation of the Signal Processing Systems Technical Committee of the IEEE Signal Processing Society. He is the recipient of a 1997 US National Science Foundation CAREER Award.