Preface

Next-generation broadband wireless standards, e.g. IEEE 802.16e and Third Generation Partnership Project – Long Term Evolution (3GPP-LTE), use Orthogonal Frequency Division Multiple Access (OFDMA) as the preferred physical layer multiple access scheme, esp. for the downlink. Due to the limited resources available at the base station, e.g. bandwidth and power, intelligent allocation of these resources to the users is crucial for delivering the best possible quality of service (QoS) to the consumer with the least cost.

The problem of allocating time slots, subcarriers, rates, and power to the different users in an OFDMA system has been an area of active research in recent years. Previous research efforts in OFDMA resource allocation have typically focused on maximizing instantaneous performance, i.e. the allocation decisions are performed for the current time instant subject to the current resource constraints, which is unable to fully utilize the time-varying nature of the wireless channel to improve the communication performance of the system. This book focuses instead on maximizing time-averaged rates, allowing us to exploit the temporal dimension to improve performance.

Furthermore, due to the difficult combinatorial nature of the problem, many researchers in the past have focused on developing sub-optimal heuristic algorithms. This book proposes a unified algorithmic framework based on dual optimization techniques that have complexities that are linear in the number of subcarriers and users, and that achieve negligible optimality gaps in standards-based numerical simulations. Adaptive algorithms based on stochastic approximation techniques are also proposed, which are shown to achieve similar performance with even much lower complexity.

Finally, it was assumed in previous work that perfect channel state information (CSI) is available at the transmitter, which is quite unrealistic due to inevitable channel estimation errors and feedback delay. This book develops algorithms assuming that only imperfect CSI is available, such that allocation decisions are made while explicitly considering the error statistics of the CSI.

Austin, TX
June 2007

Ian Wong
Brian Evans

2

Background

2.1 Introduction

In this chapter, we begin by reviewing the seminal and recent work in the field of multi-user wireless communications in Sec. 2.2, with emphasis on physical layer transmit optimization algorithms for OFDMA. This is followed by an exposition of my proposed approach to the problem of OFDMA resource allocation in Sec. 2.3, and a description of the OFDMA system model and key assumptions used throughout this book in Sec. 2.4. Finally, we conclude this chapter in Sec. 2.5.

2.2 Review of Related Work

2.2.1 Scheduling in Wireless Networks

The idea of using channel information at the transmitter to improve the performance of communication systems have been around since at least 1968 [22]. The main concept is to utilize knowledge about the channel to adjust transmission parameters accordingly to maximize communications performance, which is known as *adaptive modulation and coding*. Adaptive modulation and coding in single-user wireless communication systems have been studied extensively (see [23] [24] and the references therein). The extension of the adaptive modulation concept to scheduling in multi-user wireless networks have also been very well studied since the introduction of the concepts of *multiuser diversity* [25] and *proportional fair scheduling* [26]. In these seminal papers, the fading wireless channel was seen as a vehicle to improve the overall system performance when multiple users are involved. The theoretical underpinnings behind this concept, and the fundamental limits of these multiuser channels are addressed by the field of *multiuser information theory*, which is the topic of the next subsection.

2.2.2 Multiuser Information Theory

The focus of this book is on the downlink transmission channel for OFDMA, since this is typically where the increased performance is needed for mobile broadband wireless access applications. This is called a *broadcast channel* [27] in information theory, which consists of a sender with a transmit power and bandwidth budget that is sending independent information simultaneously to multiple users. The capacity and optimal resource allocation for fading broadcast channels has been quite well studied. In [28] and [29], the ergodic and outage capacity , and the optimal resource allocation for a flat-fading broadcast channel was derived. In [30], the capacity region for a frequency-selective broadcast channel with colored Gaussian noise was derived. In [31], the capacity and optimal power allocation for a flat-fading broadcast channel was derived subject to minimum rate constraints. It was shown in the aforementioned publications that superposition coding, followed by successive interference cancellation, is required in order to achieve the capacity of the channel. If we use OFDM transmission with infinitesimally small subcarrier widths to approximate the superposition coding transmission over a frequency-selective channel, some subcarriers would need to be shared among different users, which makes decoding overly complex for practical implementations. Fortunately, the amount of subcarrier sharing is minimal even in the capacity-achieving case [30]. Thus, assigning only one user to each subcarrier could still achieve transmissions close to capacity, and is essentially the downlink OFDMA transmission scheme. However, near capacity performance can be achieved only when optimal allocation of subcarriers, rates, and power is performed.

2.2.3 Physical Layer (PHY) Transmit Optimization

The problem of assigning the subcarriers, rates, time slots, and power to the different users in an OFDMA system has been an area of active research over the past several years. The research in this area can be broadly categorized into two: *margin-adaptive* and *rate-adaptive*. *Margin adaptation* refers to minimizing the transmit power subject to minimum quality of service (QoS) parameters for each user, which could be a combination of data rate, bit error rates, delays, etc. *Rate adaptation* refers to maximizing the data rates subject to various QoS and/or resource constraints.

Margin-adaptive Resource Allocation

In [32], the *margin-adaptive* resource allocation problem was investigated, in which an iterative subcarrier and power allocation algorithm was proposed to minimize the total transmit power given a set of fixed user data rates and bit error rate (BER) requirements. They applied a *constraint relaxation* technique, which allowed the binary integer parameter of subcarrier assignment

to take on real values, which in turn implies a *time-sharing* of each subcarrier among users. This converted the problem into a convex minimization problem with a convex feasible region, and allowed the use of iterative convex optimization algorithms to find the global minimum transmit power. The user with the biggest time-sharing factor on each subcarrier is then assigned to that subcarrier, and a single-user OFDM bit-loading algorithm (see e.g. [33]) is then run for each user. Although an iterative solution is required in this algorithm, it is guaranteed to converge to a good solution. Unfortunately, the algorithm requires a large number of iterations to converge, and is too complex for cost-effective real-time implementation.

In [34], computationally inexpensive algorithms were proposed to solve the margin-adaptive problem. They decoupled the problem into a bandwidth allocation step, which determined the number of subcarriers to be assigned to each user; and a subcarrier allocation step, which determined the actual subcarrier assignments to each user. Greedy heuristics were developed for each of the two steps, and were shown to give comparable performance to the constraint relaxation technique of [32] with lower complexity.

In [35], an alternative integer programming (IP) formulation, and a linear programming (LP) relaxation algorithm were proposed for the margin-adaptive problem. It was shown that their methods outperform the constraint relaxation method in [32] at a lower complexity, but the complexity performance was not justified rigorously. In [36], iterative refinement is used to come close to the IP solution of [35].

Rate-adaptive Resource Allocation

In [37], the *rate-adaptive* problem was investigated, wherein the objective was to maximize the total sum continuous rate over all users subject to power and BER constraints. It was shown in [37] that in order to maximize the total capacity, each subcarrier should be allocated to the user with the best gain on it, and the power should be allocated using the water-filling algorithm across the subcarriers. However, no fairness among the users was considered in [37]. Thus, the users that have the best channel conditions will be assigned all the resources, which leaves many users without a chance to use the spectrum at all. The same authors extended the problem formulation to consider *ergodic rates* in [38], i.e. the expected value of the sum rate is maximized, which utilizes the temporal dimension when ergodicity of the channel gains is assumed to improve the data rate performance. However, [38] likewise suffers from the unfairness problem.

This problem was partially addressed in [39] and [40] by ensuring that each user would be able to transmit at a minimum rate. The authors of [39] approached it using two steps similar to [34], wherein the number of subcarriers and power is initially assigned to each user using a greedy algorithm; followed by the subcarrier assignment step using the Hungarian algorithm. In [40], the approach was a simple greedy algorithm that assumes equal power allocation

among subcarriers, and assigned the best subcarrier to each user until the rate requirements for all users are achieved. The remaining subcarriers are then assigned to the users with the best channel gains in them.

In [41], an alternative formulation that maximized the minimum user's data rate was solved by using subcarrier time-sharing methods as in [32]. This enforced a notion of max-min fairness, and thus the starvation of some users in the method of [37] can be avoided. A suboptimal greedy algorithm was also developed which was shown to be close to the relaxed convex problem. This method, though, assumes that all users have similar QoS requirements, which is not the case for practical systems.

In [42], prioritization was enforced using a weighted-sum rate maximization , and a subcarrier time-sharing convex relaxation similar to [32] was used to derive the optimum subcarrier and power allocation. Several greedy algorithms were also proposed to solve the problem with lower complexity. Different weights were assigned to different users, and a higher weight for a user would imply a higher priority of getting resources. By varying the weights for each user's rate, the boundary of the rate-region can also be traced out . In the special case of the weights being identically unity, it would reduce to the problem addressed in [37]. The authors, however, neglected to indicate how the weights are to be assigned in an actual system. More recently, [43] and [44] have discovered a dual optimization framework to solve a similar weighted-sum continuous rate maximization problem. Their work is similar to the approach we advocate in this book, and is one of the special cases that our unified framework can solve (see Section 3.2.6). Note that our contribution in Sec. 3.2.6 was developed independently of [43] and [44].

In [45], 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. This is a concrete way of assigning priorities to the users, instead of simply assigning arbitrary weights as in [42]. This method is also very useful for service level differentiation, which allows for flexible billing mechanisms for different classes of users. However, the power allocation algorithm proposed in [45] involves solving simultaneous non-linear equations , which requires computationally expensive iterative operations and is thus not suitable for a cost-effective real-time implementation. In cases there the signal-to-noise ratio is high, the algorithm in [45] is shown to reduce to a one-dimensional zero-finding routine, which is much less complex, but may suffer from stability problems. In [46], the strict proportional rate constraints are relaxed to hold approximately, which allowed the power allocation to be solved in closed-form, significantly reducing the complexity, while improving the achieved sum capacity.

Several other methods that use various heuristics have also been proposed. Examples of these include subcarrier partitioning to reduce complexity [47], and game-theoretic Nash bargaining solutions [48].

2.2.4 PHY-MAC Cross-layer Optimization

All of the aforementioned approaches focused on the physical layer transmission optimization for OFDMA. This section reviews several important papers on the PHY-MAC cross-layer approach to OFDMA resource allocation, where longer-term throughput optimality and queue state information is included in their optimization goals.

In [49], resource allocation that optimizes total packet throughput subject to the user's outage probability constraint was proposed. Their algorithm assumes a finite queue size for arrival packets, and dynamically allocates the resources every time-slot based on the users' average SNR, traffic patterns, and QoS requirements. In [50], throughput maximization coupled with queue load balancing was proposed for a simple ON/OFF channel model. Their approach reduced the allocation problem into a maximum weight matching of a bipartite graph, and was shown to stabilize the queues in the OFDMA system, whereas using instantaneous optimization approaches do not.

In [51], an *opportunistic cumulative distribution function (CDF)*-scheduling based subcarrier allocation, and a proportionally-fair power allocation was proposed. Their algorithm was shown to improve overall system capacity in terms of time-average throughput. In [52], a similar opportunistic scheduling algorithm based on [53] that exploits the time varying channel was proposed. In their work, a constant power allocation is assumed, and each user is assigned a *time-slot* for which it could transmit on the assigned subcarrier. Optimal scheduling policies for three QoS/fairness constraints–temporal fairness, utilitarian fairness, and minimum-performance guarantees, were derived to maximize the asymptotic best-case system performance. More recently, in [54] [55], a cross-layer approach that bridges the gap between the physical (PHY) layer and the media access control (MAC) layer was investigated. It was shown that tradeoffs between efficiency and fairness can be realized by maximizing a concave utility function of the user's data rate, instead of maximizing the data rates themselves. Time diversity was also exploited in [55] by maximizing the utility function of an *exponentially weighted* and *time-windowed* average data rate of each user. Prepublished work by the same authors [56] extend the utility based optimization to develop a max-delay-utility scheduling algorithm that utilizes both channel and queue state information.

2.2.5 Comparison of Related Work

Table 2.1 presents a summary of the comparison among several relevant research efforts in OFDMA physical layer transmit optimization. We compare the various research publications in terms of how they formulated the problem, their proposed solution to the problem, and the channel knowledge assumptions that they made. The criteria we use is such that a "Yes" is more desirable in terms of achieving better performance, requiring less computational complexity, or making more realistic assumptions.

Table 2.1. Related work comparison

Method \ Criteria	Formulation			Solution		Assumption	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Max-min rate [41]	No	No	No	No	No	No	Yes
Sum rate [37][38]	Yes	No	No	Yes	No	No	No
Proportional rate [45][46]	No	No	Yes	No	No	No	Yes
Max-utility [54][55]	[a]No	Yes	Yes	No	No	No	Yes
Weighted rate [43][44]	No	No	Yes	[b]Yes	[c]Yes	No	Yes

[a] Considered some form of temporal diversity by maximizing an exponentially windowed running average of the rate
[b] Independently developed a similar instantaneous continuous rate maximization algorithm
[c] Only for instantaneous continuous rate case, but was not shown in their papers

Criteria

(1) Ergodic rates: The optimization problem is posed such that the *expected value* of the rate is being maximized instead of *instantaneous rate*, which allows the temporal dimension to be exploited when assuming ergodicity of channel gains.
(2) Discrete rates: The practical transmission scheme of only allowing a discrete set of possible data rates is considered rather than just the theoretical continuous rate.
(3) User prioritization: The problem formulation allows setting varying priorities among users to ensure fairness in the system.
(4) Practically optimal: The algorithm is shown in simulations using realistic parameters to have negligible optimality gaps.
(5) Linear complexity: The algorithm can be performed with complexity that is just linear in the number of users and subcarriers.
(6) Imperfect CSI: The algorithm assumes the more realistic scenario of the presence of errors in the available channel state information.
(7) Does not require CDI: The algorithm does not assume knowledge of the probability distribution function of the channel gains, which is difficult to obtain in practice.

In terms of the problem formulation, only [38] considered ergodic rates, and only [55] considered discrete rates. Under the proposed solutions, only [43] and [44] can be considered practically optimal with linear complexity. In terms of channel knowledge assumption, it should be noted that none of the surveyed papers considered imperfect CSI, and only [38] requires CDI since it is also the only work that considers ergodic rate maximization.

2.3 A New Approach to OFDMA Resource Allocation

This book primarily focuses on the physical layer transmit optimization in OFDMA, and assumes that the upper MAC layer performs the other necessary functions, including admission and congestion control, queue management, and user prioritization. This book can thus be seen as a complementary

work to the PHY-MAC cross-layer scheduling work, since it extracts further improvements to the physical layer data rate performance in order to benefit the overall system throughput performance.

We observe that in most of the aforementioned work in physical layer transmit optimization, the formulation and algorithms only consider instantaneous performance metrics. Thus, the temporal dimension is not being exploited when the resource allocation is performed. Although the PHY-MAC cross-layer studies performed in [51] and [55] considered time-averaged throughput performance, their channel-based adaptations are based on the average channel-to-noise ratio (CNR), and their approaches focused more on the effect of the past channel information on fairness, rather than exploiting the temporal variations of the wireless channel directly to improve the overall physical data rate performance. We formulate problems considering *ergodic* rates for both continuous (capacity-based) and discrete (Adaptive modulation and coding) rates assuming the availability of the distribution function of the CNR (this assumption is subsequently relaxed in Chapter 5). 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. Interestingly, when considering ergodic rates, we increase the complexity only slightly during an initialization step, e.g. during frame preamble processing in a frame-based transmission; but actually reduce the complexity when performing the actual resource allocation during data transmission versus instantaneous optimization.

Furthermore, previous research efforts 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. [32] [41] [42]). 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 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, requiring $\mathcal{O}((2MK)^3)$ operations per iteration when using Newton-type projected gradient methods [57]. 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. [34] [42]).

Our approach, on the other hand, is based on a *Lagrangian relaxation* of the power constraints and (possibly) rate constraints, instead of the *constraint relaxation* proposed previously. This relaxation retains the subcarrier assignment exclusivity constraints, but "dualizes" the power/rate constraints and incorporate them into the objective function, thereby allowing us to solve the

dual problem instead. This dual optimization framework is much less complex, with complexity order $\mathcal{O}(MK)$; and achieves relative optimality gaps that are less than 10^{-4} (i.e. achieving 99.9999% of the optimal solution) in simulations based on realistic parameters. We also provide adaptive algorithms based on stochastic approximation methods that are shown to converge to the dual optimal solutions w.p.1 with linear complexity *without* the need for iterations. Note that the dual optimization approach is also studied in [43] [44] [58], but their focus has been on instantaneous continuous rate optimization only.

2.4 System Model

In this section, we elaborate on the system model and assumptions considered in this book. Table 2.2 is a notation glossary of the most commonly used terms in this book.

2.4.1 OFDMA Signal Model

We consider a single-cell OFDMA base station, where we ignore the effect of inter-cell interference, which we assume to be either absent (sufficient cell separation given the power budget) or simply modeled as additive white Gaussian noise which increases the noise variance of the signal model. The OFDMA base station has K_{fft} subcarriers with L_{cp} cyclic-prefix, wherein there are K used subcarriers and M active 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 base station transmit power of $\bar{P} > 0$, sampling frequency F_s, bandwidth B, and flat noise power spectral density N_0. The received signal vector for the mth user at the nth OFDM symbol assuming perfect sample and symbol synchronization, and sufficient cyclic prefix length, is given as

$$\boldsymbol{y}_m[n] = \boldsymbol{\Gamma}_m[n]\mathbf{H}_m[n]\boldsymbol{x}_m[n] + \boldsymbol{\nu}_m[n] \tag{2.1}$$

where $\boldsymbol{y}_m[n]$ and $\boldsymbol{x}_m[n]$ are the K-length received and transmitted complex-valued signal vectors; $\boldsymbol{\Gamma}_m[n] = \text{diag}\left\{\sqrt{p_{m,1}[n]}, \ldots, \sqrt{p_{m,K}[n]}\right\}$ is the diagonal gain allocation matrix with $p_{m,k}[n]$ as the power allocated to user m in subcarrier k at time n; $\boldsymbol{\nu}_m[n] \sim \mathcal{CN}(\mathbf{0}, \sigma_\nu^2 \boldsymbol{I}_K)$ with noise variance $\sigma_\nu^2 = N_0 B/K$ is the white zero-mean, circular-symmetric, complex Gaussian (ZMCSCG) noise vector; and

$$\mathbf{H}_m[n] = \text{diag}\left\{h_{m,1}[n], \ldots, h_{m,K}[n]\right\} \tag{2.2}$$

is the diagonal channel response matrix.

2.4.2 Multiuser Statistical Fading Channel Model

The diagonal elements $h_{m,k}[n]$ of (2.2) are the complex-valued frequency-domain wireless channel fading random processes for the mth user at the kth

Table 2.2. Notation Glossary

Notation	Description
B	Bandwidth
N_0	Noise power spectral density
F_s	Sampling frequency
N_t	No. of time domain multipath taps
L_{cp}	Length of cyclic prefix
K_{fft}	Number of subcarriers
n	OFDMA symbol index
$h_{m,k}[n]$	Frequency domain complex channel gain
$g_{m,k}[n]$	Time domain complex channel gain
$\mathcal{K}$	Set of used subcarrier indices
K	Number of used subcarriers
k	Subcarrier index
$\mathcal{M}$	Set of active users
M	Number of active users
m	User index
$\mathcal{L}$	Set of discrete rate level indices
L	Number of discrete rate levels
l	Rate level index
r_l	Rate for level l
η_l	SNR upper boundary for rate level l
$\boldsymbol{\mathcal{L}}$	Space of allowable rate vectors
$l_{m,k}$	Rate allocation for user m and subcarrier k
BER_l	Bit error rate for rate level l
$\overline{\text{BER}}$	Average BER constraint
$\boldsymbol{\mathcal{P}}$	Space of allowable power vectors
$\overline{P}$	Total power constraint
$p_{m,k}$	Power allocated to user m and subcarrier k
$\gamma_{m,k}$	CNR of user m and subcarrier k
$\hat{\gamma}_{m,k}$	Predicted CNR of user m and subcarrier k
$\gamma_{0,m}$	Cut-off CNR for user m in multi-level waterfilling
σ_ν^2	Ambient noise variance
$\hat{\sigma}_{m,k}^2$	Prediction error variance for user m and subcarrier k
$\rho_{m,k}$	Prediction error to ambient noise ratio
λ	Geometric multiplier
w_m	User weights
$\widehat{\ }$	Superscript for estimated/predicted terms
*	Superscript for optimal terms
$^d/_d$	Superscript/subscript for discrete rate related terms

subcarrier, 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 $\Delta f = F_s/K_{\text{fft}}$

$$h_{m,k}[n] = \sum_{i=1}^{N_t} g_{m,i}[n] e^{-j2\pi\tau_i k\Delta f}. \tag{2.3}$$

The time-domain multipath taps $g_{m,i}[n]$ are modeled as stationary and ergodic discrete-time random processes with normalized temporal autocorrelation function

$$r_{m,i}[\Delta] = \frac{1}{\sigma_{m,i}^2}\mathbb{E}\{g_{m,i}[n]g_{m,i}^*[n+\Delta]\}, \quad i = 1,\ldots,N_t \tag{2.4}$$

with tap power $\sigma_{m,i}^2$, 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.

Although the results in this book are applicable to any stationary fading distribution, we shall prescribe a particular distribution for the fading channels for illustration purposes. We assume 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 (2.3), we have

$$\begin{aligned} \boldsymbol{h}_m &\sim \mathcal{CN}(\mathbf{0}_K, \mathbf{R}_{\boldsymbol{h}_m}) \\ \mathbf{R}_{\boldsymbol{h}_m} &= \mathbf{W}\boldsymbol{\Sigma}_m\mathbf{W}^H \end{aligned} \tag{2.5}$$

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 power[1]. 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 $\mathbf{R}_{\boldsymbol{h}}$ 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 $\gamma_{m,k} = |h_{m,k}|^2/\sigma_\nu^2$ denote the instantaneous channel-to-noise ratio (CNR) with mean $\bar{\gamma}_{m,k} = \sigma_m^2/\sigma_\nu^2$. 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.

[1] Following the convention in [17] and [19], we assume that the number of used subcarriers K is odd by including the null subcarrier at index 0 as part of the used subcarriers.

2.4.3 Optimization Variables

Denote by $\boldsymbol{p} = [\boldsymbol{p}_1^T, \cdots, \boldsymbol{p}_K^T]^T$ the length MK vector of power allocation values to be determined, where $\boldsymbol{p}_k = [p_{1,k}, \cdots, p_{M,k}]^T$ is the M-length vector of power allocation values with $p_{m,k}$ as the assigned power for user m in a subcarrier k. Although subcarrier, rate, and time slot allocation is required, in addition to determining the power values, it can be seen that the power vector can essentially capture these other resource assignments as well.

Subcarrier Allocation

The exclusive subcarrier allocation restriction in OFDMA can be captured by constraining the power vector as $\boldsymbol{p}_k \in \boldsymbol{\mathcal{P}}_k \subset \mathbb{R}_+^M$, where the space of allowable power vectors is

$$\boldsymbol{\mathcal{P}}_k \equiv \{\boldsymbol{p}_k \in \mathbb{R}_+^M | p_{m,k} p_{m',k} = 0; \forall m \neq m'; m, m' \in \mathcal{M}\} \tag{2.6}$$

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.

Continuous Rate Allocation

The continuous rate or capacity for user m and subcarrier k is given as

$$R_{m,k}(p_{m,k}\gamma_{m,k}) = \log_2(1 + p_{m,k}\gamma_{m,k}) \quad \text{bps/Hz} \tag{2.7}$$

Thus, the power allocation value $p_{m,k}$ determines a unique rate allocation, and $p_{m,k} = 0$ also results in zero rate allocation, which of course also means that the subcarrier k is not assigned to user m.

Discrete Rate Allocation

In the discrete rate allocation case, the data rate of the kth subcarrier for the mth user can be given by the staircase function

$$R_{m,k}^d(p_{m,k}\gamma_{m,k}) = \begin{cases} r_0, & \eta_0 \leq p_{m,k}\gamma_{m,k} < \eta_1 \\ r_1, & \eta_1 \leq p_{m,k}\gamma_{m,k} < \eta_2 \\ \vdots, & \vdots \\ r_{L-1}, & \eta_{L-1} \leq p_{m,k}\gamma_{m,k} < \eta_L \end{cases} \tag{2.8}$$

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 $\eta_L = \infty$. Thus, similar to the continuous rate case, the power allocation value $p_{m,k}$ determines a unique rate allocation

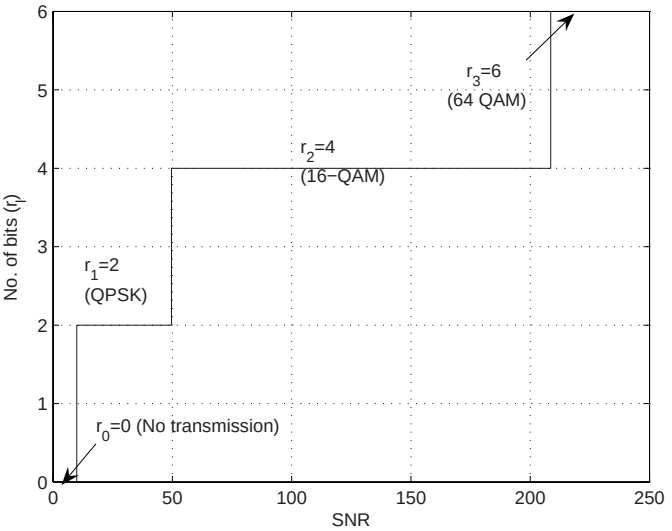

Fig. 2.1. Example discrete-rate function for an uncoded system with BER=10^{-3}. Note that the SNR is plotted in linear and not dB scale.

for a particular target BER, and $p_{m,k} = 0$ also results in zero rate allocation. We assume a Grey-coded square 2^{r_l}-QAM modulation scheme, where the BER without channel coding in AWGN can be approximated to within 1-dB for $r_l \geq 2$ and BER $\leq 10^{-3}$ by BER $\approx 0.2e^{\left[\frac{-1.6p_{m,k}\gamma_{m,k}}{2^{r_l}-1}\right]}$ [24]. Fig. 2.1 shows an example of a discrete rate function for rate levels $r_l = \{0, 2, 4, 6\}$ corresponding to no transmission, QPSK, 16-QAM, and 64-QAM transmission, and SNR boundaries $\eta_l \in \{0, 9.93, 49.66, 208.45\}$ with a BER constraint of 10^{-3}.

Time slot allocation

In the context of OFDMA, a time slot can be considered as a single OFDMA symbol (or several OFDMA symbols), and time slot allocation in this case is more granular than conventional TDMA time slot allocation since each OFDMA symbol may be shared by more than a single user. Hence, time slot allocation fundamentally entails performing the OFDMA resource allocation algorithms across time for each OFDMA symbol. In the previous work that considered instantaneous rate allocation only, the OFDMA algorithms were simply re-run every symbol (or several symbols). In this book, we can capture the idea of "time slot allocation" by using the ergodicity assumption , and determine *power allocation functions* that are parameterized by the channel knowledge. For example, if we assume perfect channel knowledge, then our optimization variable is essentially

$$\boldsymbol{p}(\cdot) \in \boldsymbol{\mathcal{P}} \equiv \left\{ \mathbb{R}_{+}^{MK} \rightarrow \mathbb{R}_{+}^{MK} : p_{m,k}(\cdot) p_{m',k}(\cdot) = 0 \text{ w.p.1}, \forall m' \neq m \right\} \quad (2.9)$$

whose search space includes all $\mathbb{R}_{+}^{MK}$-measurable functions with exclusive subcarrier allocation restriction imposed w.p.1. In the case of the adaptive algorithms discussed in Chapter 5, the power allocation is indexed by the time index n, i.e. $\boldsymbol{p}[n]$ and the exclusive subcarrier allocation restriction is simply imposed as $p_{m,k}[n]p_{m',k}[n] = 0, \forall m' \neq m, \forall n$.

2.4.4 PHY-MAC Interaction

The resource allocation problems considered in this book include assigning the power, subcarriers, rates, and time slots to the different users such that weighted-sum rate (Chapters 3-4) or sum rate subject to proportional rate constraints (Chapter 5) of the users are maximized. Although the focus of this book is primarily on the physical layer transmit optimization, it is important to discuss our assumptions on the cross-layer PHY-MAC interactions in order to see how one can apply the results in PHY-MAC cross-layer optimization discussed in Sec. 2.2.4. Specifically, we assume that the upper MAC layer passes the following information to the physical layer optimization routine:

- Set of active users $\mathcal{M}$: The MAC layer performs the necessary admission and congestion control to determine which are the active users at a particular time
- Priority for the active users w_m or ϕ_m for all $m \in \mathcal{M}$: Depending on queue back-logs and information on the average data rate for each user, the MAC layer sets the appropriate user weights w_m in the weighted-sum rate maximization formulations, or the user proportionality values ϕ_m in the proportional rate formulations.

There are numerous ways in which the MAC layer can determine these parameters, but are beyond the intended scope of this book. Admission and congestion control to determine the active user set depending on the utility of the network and availability of the resources are studied in [55] [59]. User prioritization by setting the weights w_m as the reciprocal of the user's average rate so far has been shown to approximate proportional fairness [55]. Another possibility is to set the weights as a directly proportional function of the queue-back log of the user, which can be shown to minimize the delay and ensure network stability [56].

2.5 Conclusion

In this chapter, we surveyed several important papers in OFDMA resource allocation, and showed the relative strengths and weaknesses of each of these. We then presented the general idea of our new approach to OFDMA resource

allocation based on dual optimization techniques. We also presented the system model and key assumptions used in this book.

Chapters 3-4 shall elaborate on the dual optimization framework for solving the weighted-sum rate maximization problem in OFDMA with channel distribution information, where we assume perfect and partial channel state information, respectively. Chapter 5 presents an extension of the framework to formulations that have proportional rate constraints with or without channel distribution information. Chapter 6 then concludes this book.