Energy-Efficient Resource Allocation in OFDMA Systems with Large Numbers of Base Station Antennas

Derrick Wing Kwan Ng, *Student Member, IEEE*, Ernest S. Lo, *Member, IEEE*, and Robert Schober, *Fellow, IEEE*

***Abstract*—In this paper, resource allocation for energy-efficient communication in an orthogonal frequency division multiple access (OFDMA) downlink network with a large number of transmit antennas is studied. The considered problem is modeled as a non-convex optimization problem which takes into account the circuit power consumption, imperfect channel state information at the transmitter (CSIT), and different quality of service (QoS) requirements including a minimum required data rate and a maximum tolerable channel outage probability. The power allocation, data rate adaptation, antenna allocation, and subcarrier allocation policies are optimized for maximization of the energy efficiency of data transmission (bit/Joule delivered to the users). By exploiting the properties of fractional programming, the resulting non-convex optimization problem in fractional form is transformed into an equivalent optimization problem in subtractive form, which leads to an efficient iterative resource allocation algorithm. In each iteration, the objective function is lower bounded by a concave function which can be maximized by using dual decomposition. Simulation results illustrate that the proposed iterative resource allocation algorithm converges in a small number of iterations and demonstrate the trade-off between energy efficiency and the number of transmit antennas.**

***Index Terms*—Energy efficiency, green communication, multiuser MIMO, large numbers of antennas, resource allocation.**

I. Introduction

Multiple-input multiple-output (MIMO) technology provides extra degrees of freedom which facilitate multiplexing gains and diversity gains. It can be shown that the ergodic capacity of a MIMO fading channel increases practically linearly with the minimum of the number of transmit and receiver antennas [1], [2]. Hence, it is not surprising that MIMO has attracted a lot of research interest in the past decade since it enables significant performance enhancement without requiring additional transmit power and bandwidth resources. However, the complexity of MIMO receivers limits the gains that can be achieved in practice, especially for handheld devices. An alternative is multiuser MIMO [3], [4] where a transmitter with a large number of antennas serves multiple single antenna users. In [3], the authors investigated the uplink sum capacity (bit-per-second-per-Hertz) of cellular networks assuming unlimited numbers of antennas at both the base station (BS) and the users. In [4], high throughputs for both the uplink and the downlink were shown for a time-division duplex multi-cell system which employed multiple BSs equipped with large numbers of antennas. In [3], [4], substantial capacity gains and better interference management capabilities were observed for MIMO, compared to single antenna systems. On the other hand, due to its high spectral efficiency and resistance to multipath fading, orthogonal frequency division multiple access (OFDMA) is a promising candidate for high speed wireless multiuser communication networks, such as 3GPP Long Term Evolution Advanced (LTE-A), IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX), and IEEE 802.22 Wireless Regional Area Networks (WRAN). In an OFDMA system, the fading coefficients of different subcarriers are likely to be statistically independent for different users. With channel state information at the transmitter (CSIT), the maximum system capacity can be achieved by selecting the best user for each subcarrier and adapting the corresponding transmit power [5], [6].

Recently, an increasing interest in multi-media services such as video conferencing and online high definition (HD) video streaming has led to a tremendous demand for high data rate communications with certain guaranteed quality of service (QoS) properties. The combination of MIMO and OFDMA is considered a viable solution for achieving these high data rates [7], [8], [9]. In fact, the data rate improvement due to multiple antennas is unlimited if we allow the numbers of antennas employed at both the transmitter and the receiver to grow. Yet, the advantages of MIMO and OFDMA do not come for free. They have significant financial implications for service providers due to the rapidly increasing cost for energy consumption in circuitries, which is often overlooked in the literature [4]-[9]. As a result, energy-efficient system designs, which adopt *energy efficiency* (bit-per-Joule) as the performance metric, have recently drawn much attention in both industry and academia [10]-[16]. In [10], a power loading algorithm is designed to minimize the energy-per-goodbit of a MIMO system. In [11] and [12], power allocation algorithms for energy-efficient multi-carrier systems were studied for macro-cell and hybrid cell structures, respectively. In

Manuscript received October 12, 2011; revised March 15 and May 30, 2012; accepted June 3, 2012. The associate editor coordinating the review of this paper and approving it for publication was G. Wunder.

D. W. K. Ng and R. Schober are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada (e-mail: {wingn, rschober}@ece.ubc.ca).

E. S. Lo is with Centre Tecnològic de Telecomunicacions de Catalunya - Hong Kong (CTTC-HK) (e-mail: ernest.lo@ieee.org).

This paper has been presented in part at the 2012 IEEE International Conference on Communications, Ottawa, Canada. This work has been supported in part by the Natural Science and Engineering Council of Canada (NSERC) under Project STPGP 396545.

Digital Object Identifier 10.1109/TWC.2012.072512.111850

[14] and [15], energy-efficient link adaptation for a linear sum rate-dependent dynamic circuit power consumption was considered. In [16], a risk-return model was proposed as a performance metric for energy-efficient power allocation in multi-carrier systems. However, all of these works assume that perfect global channel state information (CSI) of all links is available at the base station (BS). Hence, the power allocation can be done optimally and channel outage [1] can be avoided by data rate adaptation. However, in practice, CSIT is hardly perfect due to the mobility of users and/or estimation errors. Thus, channel outages occur with a non-zero probability and maximum tolerable outage probability requirements should be taken into consideration. Furthermore, if user selection and link adaptation are jointly optimized in MIMO-OFDMA systems, the energy-efficient resource allocation algorithms proposed in [10]-[16], which were designed for perfect CSIT and a single user, are no longer applicable[1]. In addition, the number of active antennas used for transmission has been assumed to be fixed in the existing literature, e.g. [1]-[16]. In other words, the optimal number of active antennas used for transmission has not been investigated, at least not from an energy efficiency point of view.

Motivated by the aforementioned observations, we formulate the resource allocation problem for energy-efficient communication in OFDMA systems with a large number of antennas and imperfect CSIT as an optimization problem. In particular, we optimize the number of activated antennas jointly with power allocation, subcarrier allocation, and data rate adaption for energy efficiency maximization. By exploiting the properties of fractional programming, the considered non-convex optimization problem in fractional form is transformed into an equivalent optimization problem in subtractive form whose solution can be computed with an iterative algorithm. Because of the large numbers of antennas, the iterative algorithm requires only path loss and shadowing information. In other words, the BS updates the resource allocation policies based on the realizations of path loss and shadowing, which only change in the order of seconds. In each iteration, the transformed objective function is further lower bounded by a concave function which can be maximized by using dual decomposition. As a result, closed-form power, data rate, antenna, and subcarrier allocation policies are obtained for maximizing the energy efficiency in each iteration.

The remainder of the paper is organized as follows. In Section II, we outline the signalling model and circuit power consumption model for downlink OFDMA systems. In Section III, we define the performance metric and formulate the resource allocation with imperfect CSIT as an optimization problem. In Section IV, the non-convex optimization problem is solved via an iterative algorithm. Section V presents numerical performance results, and in Section VI, we conclude with a brief summary of our results.

II. OFDMA Downlink Network Model

In this section, after introducing the notation used in this paper, we present the adopted channel and signal models.

[1]Although the notion of goodbit was introduced in [10], data rate adaptation was not considered for maximization of the goodbit for a given outage probability requirement.

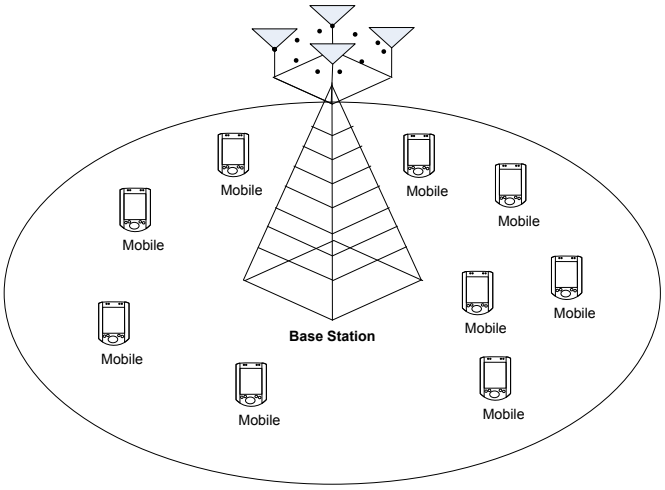

Fig. 1. Illustration of an OFDMA downlink network. There are one BS with a large number of antennas and $K = 9$ desired users equipped with a single antenna.

A. Notation

A complex Gaussian random variable with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$, and $\sim$ means "distributed as". In this paper, the following conventions are adopted. $\mathcal{O}(g(x))$ denotes an *asymptotic upper bound*. Specifically, $f(x) = \mathcal{O}(g(x))$ if $\lim_{x\to\infty} |\frac{f(x)}{g(x)}| \leq N$ for $0 < N < \infty$. $\left[x\right]^+ = \max\{0, x\}$. $\left[x\right]_b^a = a$, if $x > a$; $\left[x\right]_b^a = x$, if $b \leq x \leq a$; $\left[x\right]_b^a = b$, if $b > x$. $\Theta(g(x))$ denotes an *asymptotically tight bound*, i.e., $f(x) = \Theta(g(x))$ if $\lim_{x\to\infty} c|g(x)| \leq \lim_{x\to\infty} |f(x)| \leq \lim_{x\to\infty} d|g(x)|$ for some constants $c \leq d$. $\mathcal{E}\{\cdot\}$ denotes statistical expectation. $\mathbb{C}^{N\times M}$ is the space of all $N \times M$ matrices with complex entries. $\|\cdot\|$ and $|\cdot|$ denote the Euclidean norm of a matrix/vector and the absolute value of a complex-valued scalar, respectively. $[\cdot]^\dagger$, $[\cdot]^T$, and $\overline{[\cdot]}$ represent the conjugate transpose, transpose, and conjugate operations, respectively. $\mathrm{tr}(\mathbf{S})$ denotes the trace of matrix $\mathbf{S}$. $\Re(\cdot)$ denotes the real part of a complex number. $1(\cdot)$ denotes an indicator function which is 1 when the event is true and 0 otherwise.

B. Channel Model

We consider an OFDMA network which consists of a BS with multiple antennas and K mobile users equipped with a single antenna, cf. Figure 1. The impulse responses of all channels are assumed to be time-invariant (slow fading). There are n_F subcarriers in each orthogonal frequency division multiplexing (OFDM) symbol. The downlink received symbol at user $k \in \{1, \ldots, K\}$ on subcarrier $i \in \{1, \ldots, n_F\}$ is given by

$$y_{i,k} = \sqrt{P_{i,k} l_k g_k}\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,k} x_{i,k} + \underbrace{\sum_{j\neq k} \mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,j} x_{i,j} \sqrt{P_{i,j} l_k g_k s_{i,j}}}_{\text{Subcarrier reuse interference}} + z_{i,k}, \quad (1)$$

where $x_{i,k}$ and $\hat{\mathbf{f}}_{i,k} \in \mathbb{C}^{N_{T_{i,k}}\times 1}$ are the transmitted data symbol and the precoding vector used by the BS to transmit to user k on subcarrier i, respectively. $N_{T_{i,k}}$ is the number

of active antennas allocated to user k on subcarrier i for transmission. $P_{i,k}$ is the transmit power for the link from the BS to user k in subcarrier i. $s_{i,j} \in \{0,1\}$ is the subcarrier allocation indicator in subcarrier i for user j. $\mathbf{h}_{i,k} \in \mathbb{C}^{N_{T_{i,k}} \times 1}$ contains the small scale fading coefficients between the BS and user k on subcarrier i. l_k and g_k represent the path loss and the shadowing between the BS and user k, respectively. $z_{i,k}$ is the additive white Gaussian noise (AWGN) in subcarrier i at user k with distribution $\mathcal{CN}(0, N_0)$, where N_0 is the noise power spectral density.

C. Channel State Information

In the following, since path loss and shadowing are slowly varying random processes which both change on the order of seconds for low mobility users, we assume that the path loss and shadowing coefficients can be estimated perfectly. For the multipath fading, we assume that the users can obtain perfect estimates of the BS-to-user fading gains $\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,k} \in \mathbb{C}^{1\times 1}$, $i \in \{1,\ldots,n_F\}$, $k \in \{1,\ldots,K\}$ for signal detection purpose. However, the corresponding CSIT, i.e., $\mathbf{h}_{i,k} \in \mathbb{C}^{N_{T_{i,k}} \times 1}$ may be outdated/inaccurate at the BS because of the mobility of the users or errors in uplink channel estimation. To capture this effect, we model the multipath fading CSIT of the link between the BS and user k on subcarrier i as

$$\mathbf{h}_{i,k} = \hat{\mathbf{h}}_{i,k} + \Delta\mathbf{h}_{i,k}, \tag{2}$$

where $\hat{\mathbf{h}}_{i,k}$ and $\Delta\mathbf{h}_{i,k}$ denote the estimated CSIT vector and the CSIT error vector, respectively. $\hat{\mathbf{h}}_{i,k}$ and $\Delta\mathbf{h}_{i,k}$ are Gaussian random vectors and each vector has independent elements with respect to user index k. Besides, the elements of vectors $\mathbf{h}_{i,k}$, $\hat{\mathbf{h}}_{i,k}$, and $\Delta\mathbf{h}_{i,k}$ have zero means and normalized variances of 1, $1-\sigma_e^2$, and σ_e^2, respectively. Assuming a minimum mean square error (MMSE) estimator, the CSIT error vector and the actual CSIT vector are mutually uncorrelated. However, the fading gains of a given user may be correlated across different subcarriers.

III. Resource Allocation

In this section, we introduce the adopted system performance metric and formulate the corresponding resource allocation problem.

A. Instantaneous Channel Capacity and Outage Capacity

In this subsection, we define the adopted system performance metric. Given perfect CSI at the receiver, the channel capacity between the BS and user k on subcarrier i with subcarrier bandwidth W is given by

$$C_{i,k} = W\log_2\left(1+\Gamma_{i,k}\right) \quad \text{with} \tag{3}$$

$$\Gamma_{i,k} = \frac{P_{i,k} l_k g_k |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,k}|^2}{WN_0 + \sum_{j\neq k} |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,j}|^2 P_{i,j} s_{i,j} l_k g_k}, \tag{4}$$

where $\Gamma_{i,k}$ is the received signal-to-interference-plus-noise ratio (SINR) at user k on subcarrier i. The beamforming vector adopted at the BS is chosen to be the eigenvector corresponding to the maximum eigenvalue of $\hat{\mathbf{h}}_{i,k}\hat{\mathbf{h}}_{i,k}^T$, i.e, $\hat{\mathbf{f}}_{i,k} = \frac{\hat{\mathbf{h}}_{i,k}}{\|\hat{\mathbf{h}}_{i,k}\|}$, which is known as maximum ratio transmission (MRT). Note that zero-forcing beamforming (ZFBF) is not considered in this paper since it requires the inversion of an $N_{T_{i,k}} \times N_{T_{i,k}}$ matrix on each subcarrier for each user, which is computational expensive for large $N_{T_{i,k}}$, n_F, and K.

On the other hand, we adopt the outage capacity [1] as performance metric to account for the packet decoding errors in slow fading. The *average weighted system outage capacity* is defined as the total average number of bit/s successfully delivered to the K mobile users and is given by

$$\begin{aligned} & U(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) \\ = & \sum_{k=1}^{K} w_k \sum_{i=1}^{n_F} s_{i,k} \mathcal{E}\Big\{ R_{i,k} \times \ 1\Big(R_{i,k} \le C_{i,k}\Big)\Big\} \\ = & \sum_{k=1}^{K} w_k \sum_{i=1}^{n_F} s_{i,k} R_{i,k} \Pr\Big[R_{i,k} \le C_{i,k}\Big], \end{aligned} \tag{5}$$

where $\mathcal{P}, \mathcal{A}, \mathcal{R}$, and $\mathcal{S}$ are the power, antenna, data rate, and subcarrier allocation policies, respectively. $R_{i,k}$ is the scheduled data rate for user k on subcarrier i. $0 \le w_k \le 1$ is a positive constant provided by the upper layers, which allows the resource allocator to give different priorities to different users and to enforce certain notions of fairness. On the other hand, for designing an energy-efficient resource allocation algorithm, the total power consumption has to be included in the optimization objective function. Thus, we model the power dissipation, $U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})$, of the system as the sum of two dynamic terms and one static term [10]:

$$\begin{aligned} & U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) \\ = & \underbrace{\max_{i,k}\{s_{i,k} \times N_{T_{i,k}}\} \times P_C}_{\text{Circuit power consumption of all antennas at the BS}} \\ & + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_F} \rho P_{i,k} s_{i,k}}_{\text{BS power amplifier}} + P_0, \end{aligned} \tag{6}$$

where P_C is the constant circuit *power consumption per antenna* which includes the power dissipations in the transmit filter, mixer, frequency synthesizer, and digital-to-analog converter which are independent of the actual transmitted power. In the considered system, we assume that there is a maximum number of antennas, $N_{\max}$, at the BS. However, we only activate some of them for the sake of energy-efficient communication[2]. The physical meaning of the term $\max_{i,k}\{s_{i,k} \times N_{T_{i,k}}\}$ is that an antenna consumes power whenever it is activated even if it is used only by some of the users on some of the subcarriers[3]. $\rho \ge 1$ is a constant which accounts for the inefficiency of the power amplifier. For example, if $\rho = 5$, for every 10 Watts of radiated power in the RF, 50 Watts are consumed in the power amplifier and the power efficiency is

[2]The optimized number of active antennas will be found in next section by solving an optimization problem.

[3]Note that multiplexing the data of different users over different antennas for a fixed number of users would not increase the scheduled data rate $R_{i,k}$ because of the large number of antennas and the use of MRT precoding.

$\frac{1}{\rho} = \frac{1}{5} = 20\%$. P_0 is the basic power consumed at the BS independent of the number of transmit antennas. Hence, the *energy efficiency* of the considered system is defined as the total average number of bit/Joule successfully delivered to the users which is given by

$$U_{eff}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) \quad = \quad \frac{U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})}{U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})}. \tag{7}$$

B. Optimization Problem Formulation

The optimal power allocation policy, $\mathcal{P}^*$, antenna allocation policy, $\mathcal{A}^*$, data rate adaption policy, $\mathcal{R}^*$, and subcarrier allocation policy, $\mathcal{S}^*$, can be obtained by solving

$$\max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S},} \; U_{eff}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})$$

$$\text{s.t.} \;\; \text{C1:} \sum_{k=1}^{K} \sum_{i=1}^{n_F} s_{i,k} R_{i,k} \geq r, \quad \text{C2:} \sum_{k=1}^{K} \sum_{i=1}^{n_F} P_{i,k} s_{i,k} \leq P_T,$$

$$\text{C3:} \Pr\left(C_{i,k} < R_{i,k}\right) \leq \varepsilon, \; \forall i, k,$$

$$\text{C4:} \; P_{i,k} \geq 0, \; \forall i, k, \qquad \text{C5:} \; s_{i,k} = \{0, 1\}, \; \forall i, k,$$

$$\text{C6:} \; N_{T_{i,k}} = \{1, 2, 3, \ldots, N_{\max}\}, N_{T_{i,k}} \in \mathbb{Z}^+ \quad \forall i, k,$$

where $\mathbb{Z}^+$ denotes the set of positive integers. C1 specifies the minimum system data rate requirement r. C2 is a transmit power constraint for the BS in the downlink. The value of P_T in C2 puts a limit on the amount of out-of-cell interference in the downlink. C3 specifies the channel outage probability requirement ε. Note that the number of active antennas is an optimization variable in this paper. Hence, the imperfect CSI of the multipath fading can only be acquired by the BS after the resource allocator has decided on the number of active antennas. Therefore, the outage probability conditioned on the multipath fading, which is commonly considered in the literature, cannot be adopted in C3. C5 is a combinatorial constraint on the subcarrier assignment. Furthermore, C5 implicitly imposes a fairness constraint, since no user can dominate the subcarrier reuse process. In other words, selected users are not allowed to multiplex different messages on the same subcarrier, since a sophisticated receiver would be required at each user, such as a successive interference cancellation receiver, to recover more than one message. Besides, the weaker users have a higher chance of being selected for reusing a subcarrier. C4 is the boundary constraint for the power allocation variables. C6 is the combinatorial constraint on the number of antennas.

Remark 1: We note that instead of the sum rate constraint in C1, an individual data rate requirement for each user could be imposed by applying a similar approach as in [17] or [18] on top of the adopted problem formulation. In this case, the individual data rate requirement of each user would act as a Lagrange multiplier, γ_k, $k \in \{1, \ldots, K\}$, which would appear in the resource allocation policy solution, in (18)-(21). The γ_k would have a similar effect as variable w_k in (18)-(21), except γ_k would be adjustable within a scheduling slot for satisfying the individual data rate requirement. However, it has been shown that if individual data rate requirements are imposed, a resource hungry user consumes almost all the system resources most of the time [17]. In other words, unveiling a trade-off between energy efficiency (EE) and spectral efficiency (SE) under such a problem formulation seems impossible, since the degrees of freedom in resource allocation are decreased significantly in this case.

IV. Solution of the Optimization Problem

The objective function in (8) is a non-convex function. In general, a brute force approach is required for obtaining a global optimal solution. However, such a method has exponential complexity with respect to (w.r.t.) the number of subcarriers which is computationally infeasible even for small size systems. In order to obtain an efficient resource allocation algorithm, we introduce the following transformation.

A. Problem Transformation

The fractional objective function in (7) can be classified as a nonlinear fractional program [19]. For the sake of notational simplicity, we define $\mathcal{F}$ as the set of feasible solutions of the optimization problem in (8). Without loss of generality, we define the maximum energy efficiency q^* of the considered system as

$$\begin{aligned} q^* \quad &= \quad \frac{U(\mathcal{P}^*, \mathcal{A}^*, \mathcal{R}^*, \mathcal{S}^*)}{U_{TP}(\mathcal{P}^*, \mathcal{A}^*, \mathcal{R}^*, \mathcal{S}^*)} \\ &= \quad \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \frac{U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})}{U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})}, \quad \forall \{\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}\} \in \mathcal{F}. \end{aligned} \tag{8}$$

We are now ready to introduce the following Theorem.

Theorem 1: The maximum energy efficiency q^* is achieved if and only if

$$\begin{aligned} \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \quad & U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) - q^* U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) \\ = \quad & U(\mathcal{P}^*, \mathcal{A}^*, \mathcal{R}^*, \mathcal{S}^*) - q^* U_{TP}(\mathcal{P}^*, \mathcal{A}^*, \mathcal{R}^*, \mathcal{S}^*) \\ = \quad & 0, \end{aligned} \tag{9}$$

for $U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) \geq 0$ and $U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) > 0$.

Proof: Since $U_{eff}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})$ is well defined, Theorem 1 can be proved by following a similar approach as in [19].

Theorem 1 reveals that for an optimization problem with an objective function in fractional form, there exists an equivalent[4] objective function in subtractive form, e.g. $U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) - q^* U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})$ in the considered case. As a result, we can focus on the equivalent objective function in the rest of the paper.

B. Iterative Algorithm for Energy Efficiency Maximization

In this section, we propose an iterative algorithm (known as the Dinkelbach method [19]) for solving (8) with an equivalent objective function. The proposed algorithm is summarized in Table I and the convergence to the optimal energy efficiency is guaranteed.

Proof: Please refer to Appendix A for the proof of convergence.

[4]Here, "equivalent" means that both problem formulations lead to the same resource allocation policies.

TABLE I
ITERATIVE RESOURCE ALLOCATION ALGORITHM.

Algorithm 1 Iterative Resource Allocation Algorithm

1: Initialize the maximum number of iterations L_{max} and the maximum tolerance ϵ
2: Set maximum energy efficiency $q = 0$ and iteration index $n = 0$
3: **repeat** {Main Loop}
4: Solve the inner loop problem in (10) for a given q and obtain resource allocation policies $\{\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}'\}$
5: **if** $U(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}') - qU_{TP}(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}') < \epsilon$ **then**
6: Convergence = **true**
7: **return** $\{\mathcal{P}^*, \mathcal{A}^*, \mathcal{R}^*, \mathcal{S}^*\} = \{\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}'\}$ and $q^* = \frac{U(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}{U_{TP}(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}$
8: **else**
9: Set $q = \frac{U(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}{U_{TP}(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}$ and $n = n + 1$
10: Convergence = **false**
11: **end if**
12: **until** Convergence = **true** or $n = L_{max}$

As shown in Table I, in each iteration in the main loop (line 4 in Table I), we solve the following optimization problem for a given parameter q:

$$\max_{\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}} \quad U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) - qU_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) \quad \text{s.t. C1, C2, C3, C4, C5, C6.} \tag{10}$$

In the following, dual decomposition is applied for deriving a tractable sub-optimal solution of the main loop problem after a series of approximations[5].

1) Sub-Optimal Solution of the Main Loop Problem: The transformed problem is a mixed combinatorial and non-convex optimization problem. In order to derive an efficient resource allocation algorithm, we introduce the following proposition by taking advantage of the large numbers of antennas.

Proposition 1 (Equivalent Data Rate): For a given outage probability $\varepsilon \ll 1$ in C3, the equivalent data rate which incorporates the outage probability on subcarrier i for user k is given by

$$R_{i,k} = (1-\varepsilon) W \log_2 \Big(1 + \frac{P_{i,k} l_k g_k N_{T_{i,k}} (1-\sigma_e^2)(1-\delta)}{WN_0 + \sum_{j \neq k} (\frac{2}{\varepsilon}) P_{i,j} s_{i,j} l_k g_k}\Big), \tag{11}$$

where $0 < \delta < 1$ is a constant backoff factor. Note that $N_{T_{i,k}} \geq \lceil N_{th} \rceil$ and $\lceil N_{th} \rceil$ is the solution of (33) in Appendix B and denotes the minimum number of antennas required for Proposition 1 to hold.

Proof: Please refer to Appendix B for a proof of Proposition 1 and the meaning of δ.

The next step in solving the considered problem is to handle the inter-user interference on each subcarrier. To this end, we introduce an additional constraint C7 to the original problem which is given by

$$\text{C7:} \sum_{j \neq k} (\frac{2}{\varepsilon}) P_{i,j} s_{i,j} l_k g_k \leq I, \quad \forall k, i. \tag{12}$$

C7 can be interpreted as the maximum inter-user interference temperature [20] (tolerable interference level) in each subcarrier. In general, adding an additional constraint to the optimization problem results in a performance lower bound of the original problem due to the smaller feasible set. By varying[6] the value of I, the resource allocator is able to control the amount of interference in each subcarrier to improve the system performance. Furthermore, by substituting $\sum_{j \neq k} (\frac{2}{\varepsilon}) P_{i,j} s_{i,j} l_k g_k$ in (12) by I, the inter-user interference can be decoupled from the objective function, which facilitates the design of an efficient resource allocation algorithm. Then, the scheduled data rate between the BS and user k on subcarrier i can be lower bounded by

$$\begin{aligned} R_{i,k} &= (1-\varepsilon) W \log_2 \Big(1 + \frac{P_{i,k} l_k g_k N_{T_{i,k}} (1-\sigma_e^2)(1-\delta)}{WN_0 + I}\Big) \\ &> (1-\varepsilon) W \log_2 \Big(\frac{P_{i,k} l_k g_k}{WN_0 + I}\Big) \\ &+ (1-\varepsilon) W \log_2 \Big(N_{T_{i,k}} (1-\sigma_e^2)(1-\delta)\Big). \end{aligned} \tag{13}$$

By substituting the lower bound on the outage equivalent data rate in (13) into (10), a modified objective function, which incorporates the channel outage requirement, can be obtained for the main loop problem in (10). Indeed, it can be observed that the scheduled data rate for user k on subcarrier i in (13) depends only on the *path loss* and *shadowing* information of user k due to the large number of antennas. In other words, the dervied resource allocation policy will be identical for all subcarriers of user k.

To handle the combinatorial constraints C5 and C6, cf. (8), we follow the approach in [21] and relax constraints C5 and C6. In particular, we allow $s_{i,k}$ to be a real value between zero and one instead of a Boolean, while $N_{T_{i,k}}$ can be a positive real value. Then, $s_{i,k}$ can be interpreted as a time sharing factor for the K users for utilizing subcarrier i. Although the relaxations of $N_{T_{i,k}}$ and $s_{i,k}$ are generally sub-optimal, they facilitate the design of an efficient resource allocation algorithm. Therefore, using the equivalent data rate in Proposition 1, the auxiliary time-shared powers $\tilde{P}_{i,k} = P_{i,k} s_{i,k}$, the auxiliary time-shared number of antennas, $\tilde{N}_{T_{i,k}} = N_{T_{i,k}} s_{i,k}$, and the continuous relaxation of both C5 and C6, we can

[5]The tightness of the proposed approximations will be verified in the simulation section.

[6]The maximum inter-user interference temperature variable I is not an optimization variable in the proposed framework. However, a suitable value of I can be found via simulation in an off-line manner.

rewrite the problem in (10) for a given parameter q as

$$\begin{aligned}
&\max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \quad \tilde{U}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) - q\tilde{U}_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) \\
&\qquad \text{s.t. C4,} \\
&\text{C1:} \sum_{k=1}^{K}\sum_{i=1}^{n_F} s_{i,k}\tilde{R}_{i,k} \geq r, \qquad \text{C2:} \sum_{k=1}^{K}\sum_{i=1}^{n_F} \tilde{P}_{i,k} \leq P_T, \\
&\text{C5: } 0 \leq s_{i,k} \leq 1, \ \forall i,k, \\
&\text{C6: } N_{\max} \geq \tilde{N}_{T_{i,k}} \geq \left\lceil \tilde{N}_{th} \right\rceil, \forall i,k, \\
&\text{C7:} \sum_{j\neq k} (\frac{2}{\varepsilon})\tilde{P}_{i,j}s_{i,j}l_k g_k \leq I, \ \forall i,k, \qquad (14)
\end{aligned}$$

where $\tilde{U}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) = U(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\Big|_{P_{i,k}=\frac{\tilde{P}_{i,k}}{s_{i,k}}, N_{T_{i,k}}=\frac{\tilde{N}_{T_{i,k}}}{s_{i,k}}}$, $\tilde{U}_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) = U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\Big|_{P_{i,k}=\frac{\tilde{P}_{i,k}}{s_{i,k}}, N_{T_{i,k}}=\frac{\tilde{N}_{T_{i,k}}}{s_{i,k}}}$, and $\tilde{R}_{i,k} = R_{i,k}\Big|_{P_{i,k}=\frac{\tilde{P}_{i,k}}{s_{i,k}}, N_{T_{i,k}}=\frac{\tilde{N}_{T_{i,k}}}{s_{i,k}}}$.

The transformed problem in (14) is now jointly concave w.r.t. all optimization variables, cf. Appendix C. Thus, under some mild conditions [22], it can be shown that strong duality holds and the duality gap is equal to zero. In other words, solving the dual problem is equivalent to solving the primal problem[7].

2) Dual Problem: In this subsection, we solve the main loop problem in (14) by solving its dual. For this purpose, we first need the Lagrangian function of the primal problem. Upon rearranging terms, the Lagrangian can be written as

$$\begin{aligned}
&\mathcal{L}(\mu,\gamma,\boldsymbol{\theta},\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) \\
=\ & \sum_{k=1}^{K}(w_k+\gamma)\sum_{i=1}^{n_F} s_{i,k}\tilde{R}_{i,k} - \mu\sum_{k=1}^{K}\sum_{i=1}^{n_F}\tilde{P}_{i,k} + \mu P_T - \gamma r \\
-\ & q\Big(\max_{i,k}\{\tilde{N}_{T_{i,k}}\} \times P_C + \sum_{k=1}^{K}\sum_{i=1}^{n_F}\rho\tilde{P}_{i,k} + P_0\Big) \\
-\ & \sum_{k=1}^{K}\sum_{i=1}^{n_F}\theta_{i,k}\Big(\sum_{j\neq k}(\frac{2}{\varepsilon})\tilde{P}_{i,j}l_k g_k - I\Big), \qquad (15)
\end{aligned}$$

where $\mu \geq 0$ and $\gamma \geq 0$ are the Lagrange multipliers corresponding to the power constraint and the required minimum outage capacity constraint, respectively. $\boldsymbol{\theta}$ is the Lagrange multiplier vector associated with the inter-user interference temperature constraint C7 with elements $\theta_{i,k} \geq 0$. The boundary constraints C4, C5, and C6 will be absorbed into the Karush-Kuhn-Tucker (KKT) conditions when deriving the resource allocation policy in the following. Thus, the dual problem of (14) is given by

$$\min_{\mu,\gamma,\boldsymbol{\theta}\geq 0}\ \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \quad \mathcal{L}(\mu,\gamma,\boldsymbol{\theta},\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}). \qquad (16)$$

In the following, we solve the above dual problem iteratively by decomposing it into two layers: Layer 1 consists of n_F subproblems with identical structure; Layer 2 is the master dual problem to be solved with the gradient method.

[7]Note that by solving (14) instead of (10) in each main loop iteration of Algorithm 1, cf. Table 1, the algorithm converges to a lower bound for the maximum energy efficiency of (8).

Dual Decomposition and Layer 1 Solution: By dual decomposition, the BS first solves the following Layer 1 subproblem

$$\max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \quad \mathcal{L}(\mu,\gamma,\boldsymbol{\theta},\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) \qquad (17)$$

for a fixed set of Lagrange multipliers and a given parameter q. Using standard optimization techniques and the KKT conditions, the power allocation for user k on subcarrier i is obtained as

$$\begin{aligned}
\tilde{P}^*_{i,k} &= s_{i,k}P^*_{i,k} = s_{i,k}\left[\frac{(1-\varepsilon)W(w_k+\gamma)}{(\ln(2))(\mu+q\rho+\Omega_{i,k})}\right], \quad \text{where} \\
\Omega_{i,k} &= \sum_{j\neq k}\theta_{i,k}(\frac{2}{\varepsilon})l_j g_j \qquad (18)
\end{aligned}$$

represents the interference to the other users created by this power allocation. The power allocation has the form of *multi-level* water-filling. It can be observed that the energy efficiency variable $q \geq 0$ prevents energy inefficient transmission by truncating the water-levels. On the contrary, a large value of $\Omega_{i,k}$ results in a lower water-level in the power allocation to reduce the interference caused to the other users such that constraint C7 in (14) is satisfied.

Similarly, the close-to-optimal[8] number of activated antennas for user k on subcarrier i is given by

$$\begin{aligned}
&\tilde{N}^*_{T_{i,k}} \qquad (19) \\
=\ & N^*_{T_{i,k}}s_{i,k} = \left[\left\lceil \frac{(1-\varepsilon)W(\max_{k\in\Psi_i} w_k+\gamma)}{P_C(\frac{q}{\Phi_i})\ln(2)} \right\rceil\right]^{N_{\max}}_{\lfloor N_{th}\rfloor} s_{i,k},
\end{aligned}$$

where Ψ_i denotes a selected user set for using subcarrier i and $\Phi_i = \sum_{b\in\Psi_i} 1(\max_{k\in\Psi_i} w_k = w_b)$ counts the number of w_k which have a value equal to $\max_{k\in\Psi_i} w_k$ for all selected users. If the data rate constraint C1 in (8) is stringent, the dual variable γ is large and forces the resource allocator to assign more antennas to all selected users, cf. (19), such that constraint C1 can be satisfied. Besides, (19) reveals that all users will eventually use the same number of antennas. This behavior can be explained by the following example: Suppose user 1 and user 2 are using N_1 and N_2 antennas such that $N_1 > N_2$. Yet, from user 2's point of view, the cost for $N_1 - N_2$ extra antennas has been paid by user 1 already. Therefore, since no extra cost has to be paid, user 2 is willing to use extra antennas until $N_2 = N_1$, since this will benefit the system performance.

In order to obtain the subcarrier allocation, we take the derivative of the subproblem objective function w.r.t. $s_{i,k}$, which yields $\frac{\partial\mathcal{L}(\mu,\gamma,\boldsymbol{\theta},\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})}{\partial s_{i,k}}\Big|_{P_{i,k}=P^*_{i,k}, N_{T_{i,k}}=N^*_{T_{i,k}}} = M_{i,k}$, where $M_{i,k} \geq 0$ can be interpreted as the marginal benefit [23] for allocating subcarrier i to user k and is given by

$$\begin{aligned}
M_{i,k} = (1-\varepsilon)W(w_k+\gamma)\Bigg(&\log_2\Big(\frac{P^*_{i,k}l_k g_k}{WN_0+I}\Big) \\
&+ \log_2\Big(N^*_{T_{i,k}}(1-\sigma_e^2)(1-\delta)\Big) - 2/\ln(2)\Bigg). \qquad (20)
\end{aligned}$$

[8]Here, the sub-optimality is due to the floor and ceiling functions in (19) which are required for fulfilling the combinatorial constraint in practice.

$M_{i,k} \geq 0$ has the physical meaning that users with negative scheduled data rate on subcarrier i are not selected as they can only provide a negative marginal benefit to the system.

On the contrary, if a user has a larger weight w_k and enjoys good channel conditions with positive data rate on subcarrier i, he/she can provide a higher marginal benefit to the system. Thus, the allocation of subcarrier i at the BS is based on the following criterion:

$$s_{i,k}^* = 1 \text{ if } M_{i,k} \geq 0 \text{ and } s_{i,k}^* = 0 \quad \text{otherwise.} \qquad (21)$$

As explained earlier, since the multipath fading has vanished because of the beamforming with a large number of antennas, all the subcarriers of user k experience the same channel gain. Hence, the resource allocation policy for user k on subcarrier i, i.e., (18)-(21), is identical to that of the other $n_F - 1$ subcarriers of user k. Indeed, (21) can be interpreted as a chunk-based subcarrier allocation. In other words, if subcarrier i is allocated to user k, the other $n_F - 1$ subcarriers are also be allocated to user k since they provide the same marginal benefit. As a result, the complexity of solving the Layer 1 problem is reduced by a factor of n_F.

Finally, the data rate allocation $R_{i,k}^*$ is obtained by substituting (18) and (19) into the lower bound of the equivalent data rate in (13) for the subcarriers with $s_{i,k} = 1$.

Solution of Layer 2 Master Problem: The dual function is differentiable and, hence, the gradient method can be used to solve the Layer 2 master problem in (16) which leads to

$$\mu(m+1) = \Big[\mu(m) - \xi_1(m) \times (P_T - \sum_{k=1}^{K}\sum_{i=1}^{n_F} \tilde{P}_{i,k})\Big]^+ \qquad (22)$$

$$\gamma(m+1) = \Big[\gamma(m) - \xi_2(m) \times (\sum_{k=1}^{K}\sum_{i=1}^{n_F} s_{i,k}\tilde{R}_{i,k} - r)\Big]^+ \qquad (23)$$

$$\theta_{i,k}(m+1) = \Big[\theta_{i,k}(m) - \xi_3(m) \times (I - \sum_{j\neq k}(\frac{2}{\varepsilon})P_{i,j}s_{i,j}l_k g_k)\Big]^+ \quad \forall i,k, \qquad (24)$$

where index $m \geq 0$ is the iteration index and $\xi_u(m)$, $u \in \{1,2,3\}$, are positive step sizes. Since the transformed problem for a given parameter q is concave in nature, it is guaranteed that the iteration between Layer 1 and Layer 2 converges to the optimal solution of (14) in the main loop, if the chosen step sizes satisfy the infinite travel condition [22], [24]. Then, the updated Lagrange multipliers in (22)-(24) are used for solving the subproblems in (17) via updating the resource allocation policies.

Although equations (18)-(24) provide a solution for solving the main loop problem (line 4, Table I), (19) involves non-causal knowledge of the subcarrier allocation process for all users. This can be easily resolved by the coordinate ascent method [25], [26]. Due to page limitation, we only provide a sketch of this method. For each set of Lagrange multipliers and a given parameter q, we first keep $s_{i,k}$ fixed and find the optimized number of antennas $N_{T_{i,k}}$ and the optimized power allocation $P_{i,k}$ by using (18) and (19), respectively. Then, we solve for $s_{i,k}$ by using (21) while keeping both $N_{T_{i,k}}$ and $P_{i,k}$ fixed. The process is repeated iteratively. Once convergence is achieved, we can use (22)-(24) to update the Lagrange multipliers.

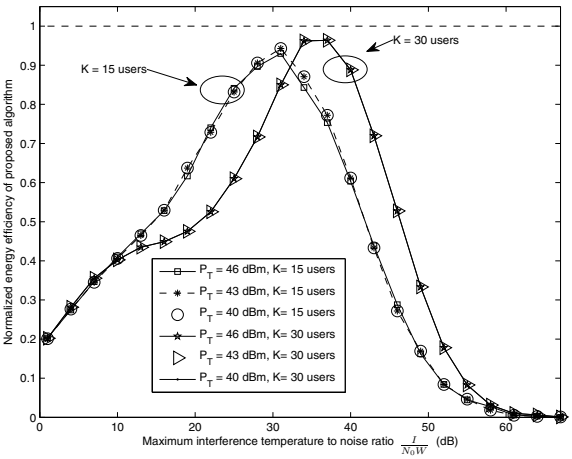

Fig. 2. The normalized performance of the proposed algorithm versus the maximum interference-temperature-to-noise ratio $\frac{I}{N_0 W}$ for different values of P_T and different numbers of users. The y-axis is normalized by a performance upper bound.

A summary of the overall algorithm is given in Table I. In each iteration of the main loop (line 3 in Table I), we solve the main loop problem in (14) for a given parameter q by dual decomposition and the coordinate ascent method, cf. (14)-(24). After obtaining the solution in the main loop, we update parameter q and use it for solving the main loop problem in the next iteration. This procedure is repeated until the proposed algorithm converges.

Note that because of the large number of antennas, the algorithm in Table I requires only path loss and shadowing information[9]. In other words, we have to execute the algorithm only in accordance with the coherence time of shadowing and path loss which is in the order of seconds for low mobility users.

V. Results

In this section, we evaluate the system performance through simulations. A single cell with a radius of 1 km is considered, cf. Figure 1. The simulation parameters can be found in Table II. In practice, the values of P_C and P_0 depend on the application-specific integrated circuits (ASIC) and the implementation. The values of P_C and P_0 adopted in this paper are for illustration purpose and are based on [27] and [28], respectively. Note that if the resource allocator is unable to guarantee the minimum data rate in a time slot, we set the energy efficiency and outage capacity in that particular time slot to zero to account for the corresponding failure. On the other hand, in the following results, the "number of iterations" is referring to the number of iterations of Algorithm 1 in Table I. Besides, we use (3) and (7) directly for computing the channel capacity and energy efficiency, respectively.

[9]The calculation of the power, data rate, antenna, and subcarrier allocations are based on the path loss and shadowing information. However, the computation of precoding vector $\hat{\mathbf{f}}_{i,k} = \frac{\hat{\mathbf{h}}_{i,k}}{\|\hat{\mathbf{h}}_{i,k}\|}$ requires multipath information.

TABLE II
SYSTEM PARAMETERS

Cell radius	1 km
Reference distance d_0	35 m
Users distribution	Uniformly distributed between d_0 and cell boundary
Small scale fading distribution	Rayleigh fading with unit variance
Carrier center frequency	2.5 GHz
Number of subcarriers n_F	256
Total bandwidth	5 MHz
Subcarrier bandwidth	19.5 kHz
Noise power per subcarrier N_0W	-131 dBm
Channel path loss model	3GPP- Urban Micro
Lognormal shadowing	Standard deviation of 8 dB
Circuit power per antenna P_C	30 dBm [27]
Static circuit power consumption P_0	40 dBm [28]
Minimum data rate requirement r	7 bit/s/Hz
Power amplifier (PA) power efficiency	$1/\rho = 0.2$
Constant back-off factor δ	0.3
CSIT error variance σ_e^2 (unless specified)	0.1
Outage probability requirement ε	0.1
N_{th}	33
$N_{\max}$	100

A. Energy Efficiency versus Maximum Inter-user Interference Temperature I

In this section, we focus on the impact of the value of I on the system energy efficiency. As can be seen from (12) and (13), the multi-user interference temperature I, which is the key for transforming the main loop problem in (14) into a convex optimization problem, plays an important role in the proposed resource allocation algorithm. The value of I puts a limit on the subcarrier reuse by controlling the amount of interference temperature[10]. For instance, by setting $I = 0$, each subcarrier can be used by one user only. On the contrary, $I \gg 1$ allows all users to transmit simultaneously on the same subcarrier. Figure 2 shows the energy efficiency of the proposed algorithm versus the value of I for different P_T and different numbers of users K. The y-axis is normalized by an upper bound on the energy efficiency of the considered system [11], such that it illustrates the achievable percentage of the energy efficiency of the reference scheme. The x-axis is the interference temperature-to-noise ratio, i.e., $\frac{I}{N_0W}$. It can be seen that for a wide range of $\frac{I}{N_0W}$ values, we can achieve more than 90% of the upper bound performance while benefiting from the convexity of the transformed problem. Furthermore, the choice of I is dependent on the number of users. This is because a higher value of $\frac{I}{N_0W}$ can be tolerated for a larger number of users as the selected users can better cope with the co-channel interference in each subcarrier due to multiuser diversity (MUD). On the other hand, as expected, the optimal value of I is not sensitive to P_T when P_T is large, since the resource allocator clips the total transmit power for energy efficiency maximization, cf. (18).

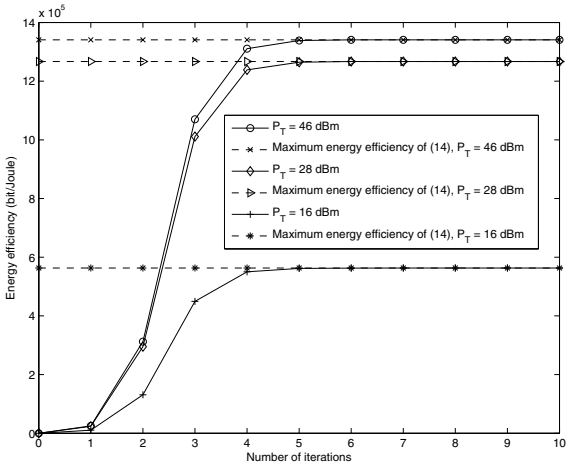

Fig. 3. Energy efficiency versus the number of iterations with $K = 15$ users for different maximum transmit powers, P_T, and channel estimation error variance $\sigma_e^2 = 0.1$.

In the following simulations, a fixed value of I is chosen for the proposed algorithm in each simulation point, such that we always achieve more than 90% of the average energy efficiency of the upper bound performance.

B. Convergence of Iterative Algorithm

Figure 3 illustrates the evolution of the proposed iterative algorithm for different values of the maximum transmit power, P_T, at the BS and $K = 15$ users. The results in Figure 3 were averaged over 100000 independent adaptation processes where each adaptation process involves different realizations of path loss, shadowing, and multipath fading. It can be observed that the iterative algorithm converges to 90% of the upper bound performance within 10 iterations for all considered numbers of transmit antennas.

[10]In practice, suitable values for I for implementing the proposed algorithm can be found in an off-line manner.

[11]The upper bound is obtained by assuming perfect channel state information is available at the base station. In addition, we remove constraints C3, C5, C6, and C7 from the optimization problem in (8) for obtaining the upper bound performance. The resulting optimization problem can be solved by using the Dinkelbach method and the spectrum balancing algorithm from [25]. Note that the spectrum balancing algorithm is a close-to-optimal numerical method for solving non-convex optimization problems in multicarrier systems. However, it converges slowly and is computationally infeasible for large size systems.

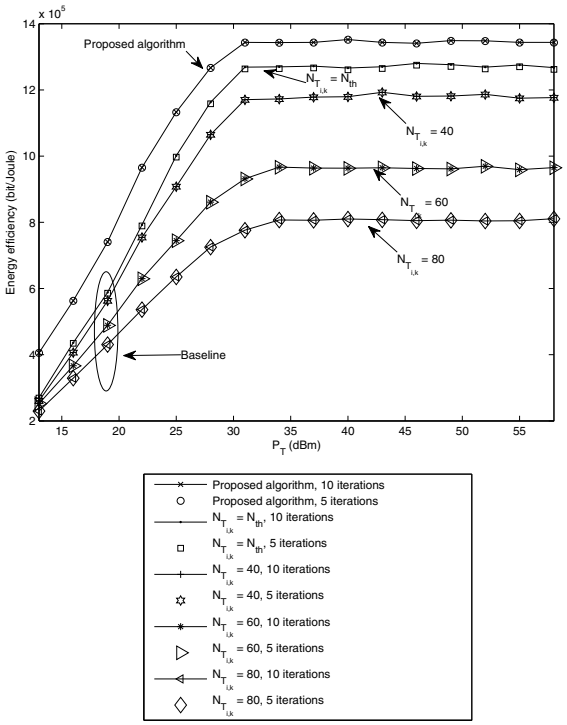

Fig. 4. Energy efficiency versus maximum transmit power, P_T, for different resource allocation algorithms with channel estimation error variance $\sigma_e^2 = 0.1$. The minimum required number of antennas is $N_{th} = 33$.

C. Energy Efficiency and Average Outage Capacity versus Transmit Power

Figure 4 illustrates the energy efficiency versus the total transmit power for $K = 15$ users. The number of iterations for the proposed iterative resource allocation algorithm is 5 and 10. The performance difference between 5 iterations and 10 iterations is negligible which confirms the practicality of our proposed iterative resource allocation algorithm. It can be observed that when the maximum transmit power at the power amplifier is large enough, e.g., $P_T \geq 40$ dBm, the energy efficiency of the proposed algorithm approaches a constant value since the resource allocator is not willing to consume more power or activate more antennas, when the maximum energy efficiency is achieved. For comparison, Figure 4 also contains the energy efficiency of a baseline resource allocation scheme in which resource allocation is performed in the same manner as in the proposed scheme, except that the number of transmit antennas is fixed to $N_{T_{i,k}} = N_{th}, 40, 60, 80, \forall i, k$, respectively. In other words, the baseline scheme optimizes energy efficiency only in terms of resource allocation policies $\{\mathcal{P}, \mathcal{R}, \mathcal{S}\}$, while the proposed algorithm optimizes energy efficiency in terms of resource allocation policies $\{\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}\}$. It can be observed that activating a fixed number of transmit antennas $N_{T_{i,k}}$ degrades the system performance in terms of energy efficiency. This is because in the baseline scheme, either more power is consumed by the circuitries for operating the antennas or the number of antennas is not large enough for satisfying the minimum data rate requirement. On the other hand, in the high transmit power regime, the performance

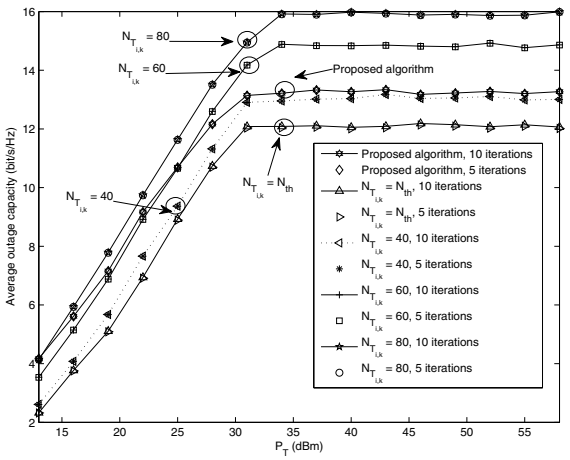

Fig. 5. Average outage capacity (bit/s/Hz) versus maximum transmit power, P_T, for different resource allocation algorithms, channel estimation error variance $\sigma_e^2 = 0.1$, and $K = 15$ users.

gain of the proposed algorithm over the baseline scheme with small $N_{T_{i,k}}$ is reduced. This is due to the fact that in the high transmit power regime, the data rate requirement is satisfied because of the high transmit power and the proposed algorithm tends to use the minimum number of antennas. In fact, the circuit power required for activating an extra antenna is relatively high, compared to the power consumed in the RF. Therefore, the proposed algorithm activates a relatively small number of antennas in the high transmit power regime and thus the performance gain due to antenna allocation becomes less significant.

Figure 5 shows the average outage capacity versus maximum transmit power P_T for $K = 15$ users. We compare the system performance of the proposed algorithm again with the baseline resource allocator. The number of iterations in the proposed algorithm is set to 5 and 10. It can be observed that the average outage capacity of the proposed algorithm approaches a constant in the high transmit power regime. This is because the proposed algorithm clips the transmit power at the BS in order to maximize the system energy efficiency. We note that, as expected, the baseline scheme resource allocator achieves a higher average outage capacity than the proposed algorithm in the high transmit power regime for most cases (except for $N_{T_{i,k}} = N_{th}$), since the proposed algorithm tends to use a smaller number of antennas. However, the superior average outage capacity of the baseline scheme comes at the expense of low energy efficiencies. On the contrary, in the low transmit power regime, i.e., $P_T \leq 25$ dBm, the proposed algorithm has a higher average outage capacity than the baseline scheme with $N_{T_{i,k}} \leq 60$ since the baseline scheme is not able to meet the data rate constraint due to insufficient numbers of antennas. On the other hand, an increasing number of antennas in the baseline scheme benefits the average outage capacity due to an improved beamforming gain. However, there is a diminishing return when $N_{T_{i,k}}$ is large due to the *channel hardening* effect [1] in the desired channels.

Figure 6 depicts the average total power consumption, i.e.,

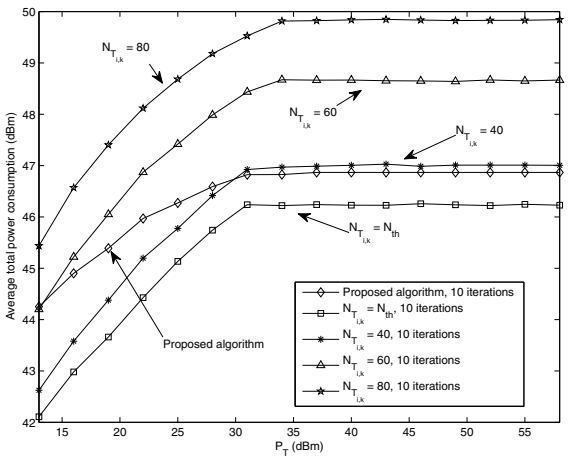

Fig. 6. Average total power consumption, $\mathcal{E}\{U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\}$, versus maximum transmit power, P_T, for different resource allocation algorithms, channel estimation error variance $\sigma_e^2 = 0.1$, 10 iterations, and $K = 15$ users.

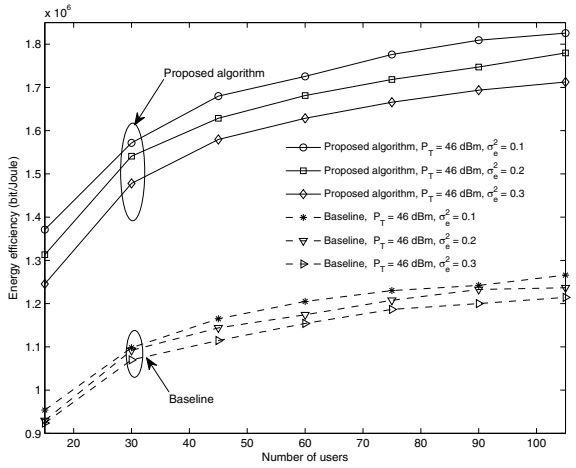

Fig. 7. Energy efficiency (bit-per-Joule) versus the number of users K for different resource allocation algorithms, different channel estimation error variances σ_e^2, and a maximum transmit power of $P_T = 46$ dBm.

$\mathcal{E}\{U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\}$, versus maximum transmit power P_T for the proposed algorithm and the baseline scheme for 10 iterations. In the regime of $P_T \leq 30$ dBm, the proposed algorithm consumes more power than the baseline scheme with $N_{T_{i,k}} \leq 40$. This is because more antennas have to be activated for satisfying the data rate requirement. However, as the maximum transmit power allowance P_T increases, the proposed algorithm gradually approaches a constant power consumption since neither further increasing the transmit power nor activating more antennas benefits the system energy efficiency.

D. Energy Efficiency versus Number of Users

Figure 7 depicts the energy efficiency versus the number of users. Different CSIT error variances σ_e^2, $P_T = 46$ dBm, and 10 iterations of the proposed algorithm are considered. It can be observed that the energy efficiency grows with the number of users since the proposed resource allocation and scheduling algorithm is able to exploit MUD. In general, MUD introduces an extra power gain [1, Chapter 6.6] to the system which provides further energy savings. Indeed, since a large number of transmit antennas reduces the multipath propagation fluctuations in each channel and causes *channel hardening*, the potentially achievable MUD gain due to the multipath channel vanishes. Yet, the MUD gain obtained from path loss and shadowing is still beneficial for the system performance in terms of energy efficiency. For comparison, Figure 7 also contains the energy efficiency of the baseline scheme mentioned in Section V-C with $N_{T_{i,k}} = 60, \forall i,k$. Figure 7 shows that although the baseline scheme is able to exploit MUD, the energy efficiency of the proposed resource allocation algorithm is superior to the baseline scheme in all considered scenarios, due to the optimization of the number of antennas.

VI. Conclusion

In this paper, we formulated the resource allocation for energy-efficient OFDMA systems with a large number of antennas as a mixed non-convex and combinatorial optimization problem, in which circuit power consumption, minimum data rate requirements, and outage probability constraints were taken into consideration. An efficient iterative resource allocation algorithm with closed-form power adaption, antenna allocation, data rate adaption, and subcarrier allocation was derived for maximization of the number of received bit-per-Joule at the users. Simulation results did not only show that the proposed algorithm converges to the solution within a small number of iterations, but demonstrated also the trade-off between energy efficiency and the number of transmit antennas: The use of a large number of antennas is always beneficial for the system outage capacity, even if the CSIT is imperfect. However, an exceedingly large number of antennas may not be a cost effective solution for improving the system performance, at least not from an energy efficiency point of view.

Appendix

A. Proof of Algorithm Convergence

We follow a similar approach as in [19] for proving the convergence of Algorithm I. We first introduce two propositions. For the sake of notational simplicity, we define the equivalent objective function in (10) as $F(q') = \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \{U(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) - q' U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\}$.

Proposition 2: $F(q')$ is a strictly monotonic decreasing function in q', i.e., $F(q'') > F(q')$ if $q' > q''$.

Proof: Let $\{\mathcal{P}',\mathcal{A}',\mathcal{R}',\mathcal{S}'\} \in \mathcal{F}$ and $\{\mathcal{P}'',\mathcal{A}'',\mathcal{R}'',\mathcal{S}''\} \in \mathcal{F}$ be the two distinct optimal resource allocation policies for $F(q')$ and $F(q'')$, respectively. $F(q'')$

$$\begin{aligned}
&= \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \{U(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}) - q'' U_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})\} \\
&= U(\mathcal{P}'',\mathcal{A}'',\mathcal{R}'',\mathcal{S}'') - q'' U_{TP}(\mathcal{P}'',\mathcal{A}'',\mathcal{R}'',\mathcal{S}'') \\
&> U(\mathcal{P}',\mathcal{A}',\mathcal{R}',\mathcal{S}') - q'' U_{TP}(\mathcal{P}',\mathcal{A}',\mathcal{R}',\mathcal{S}') \\
&\geq U(\mathcal{P}',\mathcal{A}',\mathcal{R}',\mathcal{S}') - q' U_{TP}(\mathcal{P}',\mathcal{A}',\mathcal{R}',\mathcal{S}') \\
&= F(q'). \qquad (25)
\end{aligned}$$

Proposition 3: Let $\{\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}'\} \in \mathcal{F}$ be an arbitrary feasible solution and $q' = \frac{U(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}{U_{TP}(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}')}$, then $F(q') \geq 0$.

Proof:

$$\begin{aligned} & F(q') \\ = & \max_{\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S}} \{U(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S}) - q' U_{TP}(\mathcal{P}, \mathcal{A}, \mathcal{R}, \mathcal{S})\} \\ \geq & \; U(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}') - q' U_{TP}(\mathcal{P}', \mathcal{A}', \mathcal{R}', \mathcal{S}') = 0. \end{aligned} \quad (26)$$

We are now ready to prove the convergence of Algorithm 1.

Proof of Convergence: We first prove that the energy efficiency q increases in each iteration. Then, we prove that if the number of iterations is large enough, the energy efficiency q converges to the optimal q^* such that it satisfies the optimality condition in Theorem 1, i.e., $F(q^*) = 0$. Let $\{\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n\}$ be the optimal resource allocation policies in the n-th iteration. Suppose $q_n \neq q^*$ and $q_{n+1} \neq q^*$ represent the energy efficiencies of the considered system in iterations n and $n+1$, respectively. By Theorem 1 and Proposition 3, $F(q_n) > 0$ and $F(q_{n+1}) > 0$ must be true. On the other hand, in the proposed algorithm, we calculate q_{n+1} as $q_{n+1} = \frac{U(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n)}{U_{TP}(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n)}$. Thus, we can express $F(q_n)$ as

$$\begin{aligned} & F(q_n) \\ = & \; U(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n) - q_n U_{TP}(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n) \\ = & \; U_{TP}(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n)(q_{n+1} - q_n) \\ > & \; 0 \Longrightarrow q_{n+1} > q_n, \; \because U_{TP}(\mathcal{P}_n, \mathcal{A}_n, \mathcal{R}_n, \mathcal{S}_n) > 0. \end{aligned} \quad (27)$$

By combining $q_{n+1} > q_n$, Proposition 2, and Proposition 3, we can show that as long as the number of iterations is large enough, $F(q_n)$ will eventually approach zero and satisfy the optimality condition as stated in Theorem 1.

B. Proof of Proposition 1

The outage probability requirement in C3 is a complicated non-convex function of data rates and powers, and a closed-form expression for the corresponding distribution function is not available. Therefore, we tackle this issue by the following approximations. We focus on an upper bound on the actual outage probability by bounding $\Pr(\Gamma_{i,k} < c) = \Pr(C_{i,k} < R_{i,k}), 1 \leq k \leq K$, with an outage probability requirement ε, where $\Gamma_{i,k}$ is defined in (3) and $c = 2^{\frac{R_{i,k}}{W}} - 1$. For notational simplicity, we define variables $\Phi_j = |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,j}|^2 P_{i,j} s_{i,j} l_k g_k \geq 0, \; \forall j \neq k$, $\Phi = \sum_{j \neq k} \Phi_j + N_0 W$, and $B = P_{i,k} l_k g_k |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,k}|^2$. Suppose now we restrict the resource allocator such that $\Pr\Big(\Phi \geq c_2\Big) \leq \frac{\varepsilon}{2}$ and $\Pr\Big(B \leq c_1\Big) = \frac{\varepsilon}{2}$, where $\frac{c_1}{c_2} = c = 2^{\frac{R_{i,k}}{W}} - 1$ is a function of the scheduled data rate, and c_1 and c_2 are positive constants that will be specified in the following. Hence, the actual outage probability can be expressed as

$$\begin{aligned} \Pr\Big(C_{i,k} < R_{i,k}\Big) = & \underbrace{\Pr\Big(\frac{B}{c_1} c_2 < \Phi \Big| B \leq c_1\Big)}_{a'} \Pr\Big(B \leq c_1\Big) \\ & + \underbrace{\Pr\Big(\frac{B}{c_1} c_2 < \Phi \Big| B > c_1\Big)}_{b'} \Pr\Big(B > c_1\Big). \end{aligned} \quad (28)$$

For calculating b', it can be observed that $b' \leq \frac{\varepsilon}{2}$ since $\frac{B}{c_1} > 1$ and $\Pr\Big(\Phi \geq c_2\Big) \leq \frac{\varepsilon}{2}$. On the other hand, $a' \leq 1$. Thus, the actual outage probability $\Pr\Big(C_{i,k} < R_{i,k}\Big)$ is bounded by

$$\begin{aligned} \Pr\Big(C_{i,k} < R_{i,k}\Big) & \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}(1 - \frac{\varepsilon}{2}) \\ & = \varepsilon - \frac{\varepsilon^2}{4} \approx \varepsilon \quad \text{for} \quad \varepsilon \ll 1. \end{aligned} \quad (29)$$

In other words, the outage probability requirement $\Pr\Big(C_{i,k} < R_{i,k}\Big) \leq \varepsilon$ is satisfied if we guarantee $\Pr\Big(\Phi \geq c_2\Big) \leq \frac{\varepsilon}{2}$ and $\Pr\Big(B \leq c_1\Big) = \frac{\varepsilon}{2}$.

Next, we calculate $\Pr\Big(\Phi \geq c_2\Big)$ which represents the probability that the sum power of the $K-1$ inter-user interferers exceeds c_2. Let $c_2 = \sum_{j \neq k} \eta_j + N_0 W$, where η_j are dummy variables. We obtain

$$\begin{aligned} & \Pr\Big(\Phi \geq c_2\Big) = \Pr\Big(\sum_{j \neq k} \Phi_j + N_0 W \geq \sum_{j \neq k} \eta_j + N_0 W\Big) \\ \overset{(a)}{\leq} & \frac{\mathcal{E}\{\sum_{j \neq k} \Phi_j\}}{\sum_{j \neq k} \eta_j} = \frac{\sum_{j \neq k} P_{i,j} s_{i,j} l_k g_k}{\sum_{j \neq k} \eta_j}, \end{aligned} \quad (30)$$

where (a) is due to Markov's inequality [29], [30]. Note that although Markov's inequality may not be the tightest upper bound for the corresponding outage probability, it has been widely adopted in the literature [29], [30] for calculating the outage probability in interference channels, since it only requires the first moment of the random variable. As a result, if we set $\eta_j = P_{i,j} s_{i,j} l_k g_k (\frac{2}{\varepsilon})$, then we have[12]

$$\begin{aligned} & \Pr\Big(\Phi \geq c_2\Big) \\ = & \; \Pr\Big(\sum_{j \neq k} |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,j}|^2 P_{i,j} s_{i,j} l_k g_k \geq \sum_{j \neq k} \frac{2}{\varepsilon} P_{i,j} s_{i,j} l_k g_k\Big) \\ \leq & \; \frac{\varepsilon}{2}. \end{aligned} \quad (31)$$

For calculating $\Pr\Big(B \leq c_1\Big)$, we consider

$$\begin{aligned} & |\mathbf{h}_{i,k}^T \hat{\mathbf{f}}_{i,k}|^2 \\ = & \; \Big[(\hat{\mathbf{h}}_{i,k}^T + \Delta \mathbf{h}_{i,k}^T) \frac{\hat{\mathbf{h}}_{i,k}}{\|\hat{\mathbf{h}}_{i,k}\|}\Big]^2 \\ = & \; \|\hat{\mathbf{h}}_{i,k}\|^2 + 2\Re(\Delta \mathbf{h}_{i,k}^T \hat{\mathbf{h}}_{i,k}) + \frac{\|\Delta \mathbf{h}_{i,k}^T \hat{\mathbf{h}}_{i,k}\|^2}{\|\hat{\mathbf{h}}_{i,k}\|^2} \\ \overset{(c)}{\approx} & \; \|\hat{\mathbf{h}}_{i,k}\|^2 = \Theta\Big(N_{T_{i,k}}(1 - \sigma_e^2)\Big) \qquad \text{for } N_{T_{i,k}} \to \infty, \end{aligned} \quad (32)$$

where (c) is due to the fact that $\|\hat{\mathbf{h}}_{i,k}\|^2$ scales with $N_{T_{i,k}}$ in the order of $\Theta(N_{T_{i,k}}(1 - \sigma_e^2))$ for $N_{T_{i,k}} \to \infty$, thanks to the law of large numbers. Note that $\|\hat{\mathbf{h}}_{i,k}\|^2$ is a random variable if $N_{T_{i,k}}$ is an unknown before solving the optimization problem. On the other hand, the term $2\Re(\Delta \mathbf{h}_{i,k}^T \hat{\mathbf{h}}_{i,k}) + \frac{\|\Delta \mathbf{h}_{i,k}^T \hat{\mathbf{h}}_{i,k}\|^2}{\|\hat{\mathbf{h}}_{i,k}\|^2}$ scales only in the order of $\mathcal{O}(1)$ which can be neglected for

[12]Note that in [4], [31], the denominator of the SINR is approximated by only its mean value. However, this approximation cannot guarantee a small channel outage probability requirement ε.

large $N_{T_{i,k}}$. By choosing $c_1 = P_{i,k} l_k g_k N_{T_{i,k}} (1-\sigma_e^2)(1-\delta)$, $\Pr\Big(B \le c_1\Big)$ can be upper bounded by its Chernoff bound as

$$\begin{aligned}\Pr\Big(B \le c_1\Big) &\approx \Pr\Big(\|\hat{\mathbf{h}}_{i,k}\|^2 \le N_{T_{i,k}}(1-\sigma_e^2)(1-\delta)\Big) \\ &\le \phi^{N_{T_{i,k}}} \exp\Big((1-\phi)N_{T_{i,k}}\Big) = \frac{\varepsilon}{2}, \qquad (33)\end{aligned}$$

where $\phi = (1-\sigma_e^2)(1-\delta)$ and $0 < \delta < 1$ is a constant backoff factor. Mathematically, δ represents the deviation of $(1-\sigma_e^2)$ from $\frac{\|\hat{\mathbf{h}}_{i,k}\|^2}{N_{T_{i,k}}}$, for a finite value of $N_{T_{i,k}}$. For a given outage probability requirement ε and backoff factor δ, solving (33) for $N_{T_{i,k}}$ results in the minimum required N_{th} for satisfying the outage requirement. Note that for a target outage probability requirement ε, the actual outage probability for the case of $N_{T_{i,k}} \ge N_{th}$ will be less than ε since $\phi^{N_{T_{i,k}}} \exp\Big((1-\phi)N_{T_{i,k}}\Big)$ is a decreasing function of $N_{T_{i,k}}$. Therefore, by combining (31) and (33), a scheduled data rate of $R_{i,k} = (1-\varepsilon)W\log_2\Big(1+\frac{c_1}{c_2}\Big) = (1-\varepsilon)W\log_2\Big(1+\frac{P_{i,k}l_k g_k N_{T_{i,k}}(1-\sigma_e^2)(1-\delta)}{WN_0+\sum_{j\neq k}(\frac{2}{\varepsilon})P_{i,j}s_{i,j}l_k g_k}\Big)$ can satisfy the outage probability requirement $\Pr\Big(C_{i,k} < R_{i,k}\Big) \le \varepsilon$ which proves in Proposition 1. We note that the use of the strong law of large numbers in (32) makes the optimization of $N_{T_{i,k}}$ possible since $N_{T_{i,k}}$ becomes a part of the equivalent channel gain.

C. Proof of the Concavity of the Transformed Problem in (14)

For notational simplicity, we drop the *subindices* and *scaling constants* of all optimization variables in this section such that the transformed objective function in (14) can be expressed as the summation of two functions with variables P, s, and N_T, i.e., $y = f + t$, where $f = s\log_2(P/s) + s\log_2(N_T/s)$ and $t = -q\tilde{U}_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})$. Let $\mathbf{H}(f)$ and $\lambda_1, \lambda_2, \lambda_3$ be the Hessian matrix of function f and the eigenvalues of $\mathbf{H}(f)$, respectively. The Hessian matrix of function f and the trace of the Hessian matrix are given by

$$\mathbf{H}(f) = \begin{bmatrix} \frac{-2}{s\ln(2)} & \frac{1}{P\ln(2)} & \frac{1}{N_T\ln(2)} \\ \frac{1}{P\ln(2)} & \frac{-s}{P^2\ln(2)} & 0 \\ \frac{1}{N_T\ln(2)} & 0 & \frac{-s}{N_T^2\ln(2)} \end{bmatrix} \text{ and}$$

$$\text{tr}\Big(\mathbf{H}(f)\Big) = \sum_{t=1}^{3}\lambda_t = -\frac{s^2P^2 + s^2N_T^2 + 2P^2N_T^2}{sP^2N_T^2\ln(2)}, \qquad (34)$$

respectively. Besides, it can be shown that the eigenvalues of the Hessian matrix are given by

$$\lambda_1 \times \lambda_2 = \frac{s^2+P^2+N_T^2}{P^2N_T^2\ln^2(2)} \ge 0, \quad \lambda_3 = 0. \qquad (35)$$

From (35), λ_1 and λ_2 must be either both positive or both negative. Therefore, by combining the above with $\text{tr}\Big(\mathbf{H}(f)\Big) \le 0$, we conclude that $\lambda_1, \lambda_2 \le 0$. Since $\lambda_t \le 0$, $\forall t$, so $\mathbf{H}(f)$ is a negative semi-definite matrix and f is jointly concave w.r.t. P, s, and N_T. On the other hand, function t is a jointly concave function[13] of P, s, and N_T so the concavity of function f is not destroyed by adding function f and function t. Therefore, the transformed objective function is jointly concave w.r.t. all the optimization variables.

[13]Note that $\max_{i,k}\{\tilde{N}_{T_{i,k}}\}$ is a convex function with respect to $\tilde{N}_{T_{i,k}}$. Therefore, $-q\tilde{U}_{TP}(\mathcal{P},\mathcal{A},\mathcal{R},\mathcal{S})$ is a concave function with respect to $\tilde{N}_{T_{i,k}}$ since $-q\max_{i,k}\{\tilde{N}_{T_{i,k}}\}$ is a concave function with respect to $\tilde{N}_{T_{i,k}}$.

REFERENCES

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, 1st edition. Cambridge University Press, 2005.

[2] V. K. N. Lau and Y. K. Kwok, *Channel Adaptation Technologies and Cross Layer Design for Wireless Systems with Multiple Antennas - Theory and Applications*, 1st edition. Wiley John Proakis Telecom Series, 2005.

[3] D. Aktas, M. Bacha, J. Evans, and S. Hanly, "Scaling results on the sum capacity of cellular networks with MIMO links," *IEEE Trans. Inf. Theory*, vol. 52, pp. 3264–3274, 2006.

[4] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3590–3600, Nov. 2010.

[5] J. Zeng and H. Minn, "A novel OFDMA ranging method exploiting multiuser diversity," *IEEE Trans. Commun.*, vol. 58, pp. 945–955, Mar. 2010.

[6] P. Chan and R. Cheng, "Capacity maximization for zero-forcing MIMO-OFDMA downlink systems with multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1880–1889, May 2007.

[7] T. Maciel and A. Klein, "On the performance, complexity, and fairness of suboptimal resource allocation for multiuser MIMO-OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 59, pp. 406–419, Jan. 2010.

[8] V. Papoutsis, I. Fraimis, and S. Kotsopoulos, "User selection and resource allocation algorithm with fairness in MISO-OFDMA," *IEEE Commun. Lett.*, vol. 14, pp. 411–413, May 2010.

[9] E. Lo, P. Chan, V. Lau, R. Cheng, K. Letaief, R. Murch, and W. Mow, "Adaptive resource allocation and capacity comparison of downlink multiuser MIMO-MC-CDMA and MIMO-OFDMA," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1083–1093, Mar. 2007.

[10] R. Prabhu and B. Daneshrad, "Energy-efficient power loading for a MIMO-SVD system and its performance in flat fading," in *Proc. 2010 IEEE Global Telecommun. Conf.*, pp. 1–5.

[11] A. Akbari, R. Hoshyar, and R. Tafazolli, "Energy-efficient resource allocation in wireless OFDMA systems," in *Proc. 2010 IEEE Personal, Indoor and Mobile Radio Commun. Sympos.*, pp. 1731–1735.

[12] X. Xiao, X. Tao, Y. Jia, and J. Lu, "An energy-efficient hybrid structure with resource allocation in OFDMA networks," in *Proc. 2011 IEEE Wireless Commun. and Networking Conf.*, pp. 1466–1470.

[13] G. Miao, N. Himayat, and G. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, pp. 545–554, Feb. 2010.

[14] C. Isheden and G. P. Fettweis, "Energy-efficient multi-carrier link adaptation with sum rate-dependent circuit power," in *Proc. 2010 IEEE Global Telecommun. Conf.*, pp. 1–6.

[15] ——, "Energy-efficient link adaptation with transmitter CSI," in *Proc. 2011 IEEE Wireless Commun. and Networking Conf.*, pp. 1–6.

[16] Z. Hasan, G. Bansal, E. Hossain, and V. Bhargava, "Energy-efficient power allocation in OFDM-based cognitive radio systems: a risk-return model," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 6078–6088, Dec. 2009.

[17] M. Tao, Y.-C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2190–2201, June 2008.

[18] W. W. L. Ho and Y.-C. Liang, "Optimal resource allocation for multiuser MIMO-OFDM systems with user rate constraints," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 1190–1203, Mar. 2009.

[19] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, pp. 492–498, Mar. 1967. Available: http://www.jstor.org/stable/2627691

[20] "Report of the Spectrum Ffficiency Working," FCC Spectrum Policy Task Force, Tech. Rep., Nov. 2002, http://www.fcc.gov/sptf/reports.html.

[21] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[23] W. Yu and J. M. Cioffi, "FDMA capacity of Gaussian multiple-access channels with ISI," *IEEE Trans. Commun.*, vol. 50, pp. 102–111, Jan. 2002.

[24] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," notes for EE392o Stanford University Autumn, 2003-2004.

[25] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, pp. 1310–1321, July 2006.
[26] D. P. Bertsekas, *Nonlinear Programming*, 2nd edition. Athena Scientific, 1999.
[27] R. Kumar and J. Gurugubelli, "How green the LTE technology can be?" in *2011 Intern. Conf. on Wireless Commun., Veh. Techn., Inform. Theory and Aerosp. Electron. Syst. Techn.*
[28] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. 2010 Future Network and Mobile Summit*, pp. 1–8.
[29] R. Ganti and M. Haenggi, "Interference and outage in clustered wireless ad hoc networks," *IEEE Trans. Inf. Theory*, no. 9, pp. 4067–4086, 2009.
[30] N. Jindal, J. Andrews, and S. Weber, "Rethinking MIMO for wireless networks: linear throughput increases with multiple receive antennas," in *Proc. 2009 IEEE Intern. Commun. Conf.*, pp. 1–6.
[31] C. Fischione, M. D'Angelo, and M. Butussi, "Utility maximization via power and rate allocation with outage constraints in Nakagami-lognormal channels," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 1108–1120, Apr. 2011.

Derrick Wing Kwan Ng (S'06) received the bachelor degree with first class honors and master of philosophy (M.Phil.) degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively. He is currently working toward the Ph.D. degree in the University of British Columbia (UBC). In the summer of 2011 and spring of 2012, he was a visiting scholar at the Centre Tecnològic de Telecomunicacions de Catalunya - Hong Kong (CTTC-HK). His research interests include cross-layer optimization for wireless communication systems, resource allocation in OFDMA wireless systems, and communication theory.

He received the Best Paper Awards at the IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, and the IEEE Third International Conference on Communications and Networking in China 2008. He was awarded the IEEE Student Travel Grants for attending the IEEE WCNC 2010, the IEEE International Conference on Communications (ICC) 2011, and the IEEE Globecom 2011. He was also the recipient of the 2009 Four Year Doctoral Fellowship from the UBC, Sumida & Ichiro Yawata Foundation Scholarship in 2008, and R&D Excellence Scholarship from the Center for Wireless Information Technology in HKUST in 2006. He has served as an editorial assistant to the Editor-in-Chief of the TRANSACTIONS ON COMMUNICATIONS since Jan. 2012. He has been a TPC member of various conferences, including the ICC 2012 workshop on Green Communications and Networking, the IEEE Globecom'12 Workshop on Heterogeneous, Multi-hop, Wireless and Mobile Networks, the ISIEA 2012, and the PEMOS 2012, etc.

Ernest S. Lo (S'02-M'08) is the Founding Director and Chief Representative of the Centre Tecnolgic de Tecnològic de Catalunya - Hong Kong (CTTC-HK). He was a Croucher Postdoc Fellow at Stanford University, and received his Ph.D., M.Phil., and B.Eng. (1st Hons.) from the Hong Kong University of Science and Technology. He has a broad spectrum of research interests, including channel coding, resource allocation, and wireless system and architectural design, all with a goal of finding new resources and inventing new technologies for realizing a flexible, spectrally-efficient, and energy-efficient wireless multiuser network. He contributed to the standardization of the IEEE 802.22 cognitive radio WRAN system and holds a few pending and granted US and China patents. Some of them were transferred to other companies.

Dr. Lo has received three Best Paper Awards, one at the IEEE International Conference on Communications (ICC) 2007, Glasgow, and another two at the IEEE Global Communications Conference (GLOBECOM), 2011, Houston, and the IEEE Wireless Communications and Networking Conference (WCNC) 2012, Paris, respectively. He served as an Editorial Assistant of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS when it was founded and has been a TPC member of various conferences, including the IEEE ICC'10,11,12, IEEE GLOBECOM'10,11,12, and IEEE ICCC'12, etc. He was honoured as an Exemplary Reviewer of the IEEE COMMUNICATIONS LETTERS.

Robert Schober (M'01, SM'08, F'10) was born in Neuendettelsau, Germany, in 1971. He received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from the University of Erlangen-Nuermberg in 1997 and 2000, respectively. From May 2001 to April 2002 he was a Postdoctoral Fellow at the University of Toronto, Canada, sponsored by the German Academic Exchange Service (DAAD). Since May 2002 he has been with the University of British Columbia (UBC), Vancouver, Canada, where he is now a Full Professor and Canada Research Chair (Tier II) in Wireless Communications. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at the Friedrich Alexander University (FAU), Erlangen, Germany. His research interests fall into the broad areas of Communication Theory, Wireless Communications, and Statistical Signal Processing.

Dr. Schober received the 2002 Heinz MaierLeibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, and a 2012 NSERC E.W.R. Steacie Fellowship. In addition, he received best paper awards from the German Information Technology Society (ITG), the European Association for Signal, Speech and Image Processing (EURASIP), IEEE WCNC 2012, IEEE Globecom 2011, IEEE ICUWB 2006, the International Zurich Seminar on Broadband Communications, and European Wireless 2000. Dr. Schober is a Fellow of the Canadian Academy of Engineering and a Fellow of the Engineering Institute of Canada. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS.