This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2007 proceedings.

Minimizing Power in Wireless OFDMA with Limited-Rate Feedback

Antonio G. Marques*, Georgios B. Giannakis†, Fadel F. Digham†, and F. Javier Ramos*
*Dept. of Signal Theory and Communications. Universidad Rey Juan Carlos. Fuenlabrada, 28943 Madrid (SPAIN)
Email: {antonio.garcia.marques, javier.ramos}@urjc.es
†Dept. of Electrical and Computer Engineering. University of Minnesota. Minneapolis, 55455 MN (USA)
Email: {georgios, fdigham}@ece.umn.edu

Abstract— **Emerging applications involving low-cost wireless sensor networks motivate well optimization of multi-user orthogonal frequency-division multiple access (OFDMA) in the power-limited regime. In this context, the present paper relies on limited-rate feedback (LRF) sent from the access point to terminals to minimize the total average transmit-power under individual average rate and error probability constraints. The characterization of optimal bit, power and subcarrier allocation policies based on LRF, as well as optimal channel quantization are provided. Numerical examples corroborate the analytical claims and reveal that significant power savings result even with few fed back bits.**

I. Introduction

Orthogonal frequency-division multiplexing (OFDM) is the most common modulation for bandwidth-limited wireline and wireless transmissions over frequency-selective multipath channels. OFDM transmissions over wireline or slowly fading wireless links have traditionally relied on deterministic or *perfect* (P-) channel state information at the transmitters (CSIT) to adaptively load power, bits and/or subcarriers so as to either maximize rate (capacity) for a prescribed transmit-power, or, minimize power subject to instantaneous rate constraints [9].

While the assumptions of P-CSI at the transmitters and receiver render analysis and design tractable, they may not be that realistic due to wireless channel variations and estimation errors, feedback delay, bandwidth limitation, and jamming induced errors [6]. These considerations motivate a *limited-rate feedback* (LRF) mode, where only *quantized* (Q-) CSIT is available through a (typically small) number of bits fed back from the receiver to the transmitters; see e.g., [10]. Q-CSIT entails a finite number of quantization regions describing different clusters of channel realizations [7]. Upon estimating the channel, the receiver feeds back the index of the region individual uplink channels belong to (channel codeword), based on which each terminal adapts its transmission parameters accordingly. This LRF-based mode of operation fulfills two requirements: (i) the feedback is pragmatically affordable in most practical wireless links, and (ii) the Q-CIST is robust to channel uncertainties since transmitters adapt to a few regions rather than individual channel realizations.

Work in this paper was supported by the US ARL under the CTA Program, Cooperative Agreement No. DAAD19-01-2-0011; by the USDoD ARO grant No. W911NF-05-1-0283; by the Spanish Gov't grant No. TEC2005-06766-C03-01/TCM; and by the C.A. Madrid grant No. P-TIC-000223-0505.

Resource allocation in orthogonal frequency-division multiple access (OFDMA) minimizing the transmit-power *per symbol* based on P-CSIT was first studied in [9]. Relying on fixed (as opposed to adaptive) Q-CSIT, recent works deal with optimization of power or rate performance per OFDMA symbol [2], [5]. Different from these works, here we jointly adapt power, rate, and subcarrier resources based on Q-CSIT to minimize the *average* transmit-power. Our focus is on allocation algorithms with negligible on-line computational complexity. Moreover, we rely on a general framework for modeling the Q-CSIT, which jointly optimizes resource allocation and channel quantization.

The rest of the paper is organized as follows. After introducing preliminaries on the setup we deal with (Section II), we derive optimal subcarrier, power, and bit allocation, as well as optimum quantizer design for OFDMA based on LRF (Section III). Once the optimum design is characterized (this is carried out off-line), we subsequently present an algorithm to choose the optimum LRF codeword that has to be executed on-line (Section IV). Numerical results and comparisons that corroborate our claims are presented (Section V), and concluding remarks wrap up this paper (Section VI)[1].

II. Preliminaries and Problem Statement

We consider a wireless OFDMA system (see Fig. 1) with M users, indexed by $m \in [1, M]$, sharing K subcarriers (subchannels), indexed by $k \in [1, K]$. The *instantaneous* (per symbol) power and rate user m loads on subcarrier k are denoted by $p_{k,m}$ and $r_{k,m}$, respectively. With these as entries we form $K \times M$ instantaneous power and rate matrices $\mathbf{P}$ and $\mathbf{R}$, that is $[\mathbf{P}]_{k,m} := p_{k,m}$ and $[\mathbf{R}]_{k,m} := r_{k,m}$. For a given feedback update, we consider a time sharing user access

[1]*Notation:* Lower and upper case boldface letters are used to denote (column) vectors and matrices, respectively; $(\cdot)^T$ denotes transpose; $[\cdot]_{k,l}$ the (k,l)th entry of a matrix, and $[\cdot]_k$ the kth entry of a vector; $\mathbf{X} \geq \mathbf{0}$ means all entries of $\mathbf{X}$ are nonnegative; $\mathbf{F}_N$ stands for the normalized FFT matrix with entries $[\mathbf{F}_N]_{n,k} = e^{-j\frac{2\pi}{N}kn}$, $n, k = 0, \ldots, N-1$; $f_{\mathbf{X}}(\mathbf{X})$ denotes the joint probability density function (PDF) of a matrix $\mathbf{X}$; likewise, $f_x(x)$ denotes the PDF of a scalar x; $\mathbb{E}_{\mathbf{X}}[\cdot]$ stands for the expectation operator over $\mathbf{X}$; $\lfloor \cdot \rfloor$ ($\lceil \cdot \rceil$) denotes the floor (ceiling) operation; $\mathbf{I}_{\{\cdot\}}$ is short for the indicator function; i.e., $\mathbf{I}_{\{x\}} = 1$ if x is true and zero otherwise; and $LHS(\mathrm{x})$ denotes the left hand side of equation (x).

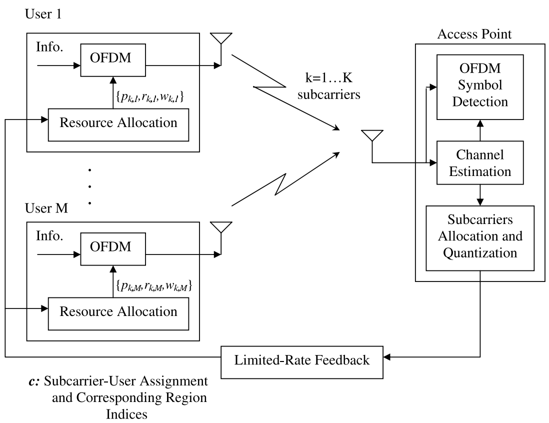

Fig. 1. System block diagram.

per subcarrier; i.e., time division multiple access (TDMA)[2]. This sharing process is described by a $K \times M$ weight matrix $\mathbf{W}$ whose (k, m)th entry $w_{k,m}$ represents the percentage of time the kth subcarrier is utilized by the mth user. Clearly, $\sum_{m=1}^{M} w_{k,m} \leq 1, \forall k$, and the average transmitted power and rate over the transmission period between successive feedback updates is $p_{k,m}w_{k,m}$ and $r_{k,m}w_{k,m}$ for the kth subcarrier of user m.

Each user's discrete-time baseband equivalent impulse response of the corresponding frequency-selective fading channel is $\mathbf{h}_m := [h_{m,0}, \ldots, h_{m,N_m}]^T$, where: $N_m := \lfloor D_{m,\max}/T_s \rfloor$ denotes the channel order, $D_{m,\max}$ the maximum delay spread, T_s the sampling period, and $N_{\max} := \max_{m \in [1,M]} N_{m,\max}$. As usual in OFDM, we suppose $K \gg N_{\max}$. For notational convenience, we collect the M impulse response vectors in a $K \times M$ matrix $\mathbf{H} := [\mathbf{h}_1, \ldots, \mathbf{h}_M]$, where the length of each column is increased to K by padding the appropriate number of zeros.

Each user applies a K-point inverse fast Fourier transform (I-FFT) to each snapshot of K-symbol streams, and subsequently inserts a cyclic prefix (CP) of size $N_{\max}$ to obtain a block of $K + N_{\max}$ symbols (i.e., one OFDM symbol), which are subsequently multiplexed and digital to analog converted for transmission. These operations along with the corresponding FFT and CP removal at the receiver convert each user's frequency-selective channel to a set of K parallel flat-fading subchannels, each with fading coefficient given by the frequency response of this user's channel evaluated on the corresponding subcarrier. Consider the $K \times M$ matrix $\tilde{\mathbf{H}} := (1/\sqrt{K})\mathbf{F}_K\mathbf{H}$, whose mth column comprises the frequency response of user m's channel.

With the multi-user channel matrix $\tilde{\mathbf{H}}$ acquired (via training symbols), the receiver has available a noise-normalized channel power gain matrix $\mathbf{G}$, where $[\mathbf{G}]_{k,m} := |[\tilde{\mathbf{H}}]_{k,m}|^2/\sigma_{k,m}^2$, with $\sigma_{k,m}^2$ denoting the known variance of the zero-mean additive white Gaussian noise (AWGN) at the receiver. We will use $g_{k,m} := [\mathbf{G}]_{k,m}$ to denote the instantaneous noise-normalized channel power gain for the kth subchannel of the mth user. Likewise, letting $\bar{\mathbf{G}} := \mathbb{E}_{\mathbf{G}}[\mathbf{G}]$, its generic entry $\bar{g}_{k,m} := [\bar{\mathbf{G}}]_{k,m}$ shall denote the average gain of the (k, m) subcarrier-user pair. Having (practically perfect) knowledge of each $\mathbf{G}$ realization, the access point (AP) allocates subcarriers to users after assigning entries of $\mathbf{G}$ to appropriate quantization regions they fall into. Using the indices of these regions, the receiver feeds back the codeword $\mathbf{c} = \mathbf{c}(\mathbf{G})$ for the users to adapt their transmission modes (power, rate and subcarriers) from a finite set of mode triplets.

Our work will rely on the following assumptions:

(as1) *Different user channels are uncorrelated; i.e., the columns of* $\mathbf{G}$ *are uncorrelated.*

(as2) *Each user's subchannels are allowed to be correlated, and complex Gaussian distributed; i.e.,* $g_{k,m}$ *obeys an exponential PDF* $f_{g_{k,m}}(g_{k,m}) = (1/\bar{g}_{k,m})\exp(-g_{k,m}/\bar{g}_{k,m})$.

(as3) *Subchannel states (regions) remain invariant over at least two consecutive OFDM symbols.*

(as4) *The feedback channel is error-free and incurs negligible delay.*

(as5) *Symbols are drawn from quadrature amplitude modulation (QAM) constellations so that the resulting instantaneous BER can be approximated as (*$\kappa_1 = 0.2$, $\kappa_2 = 1.5$*)*

$$\epsilon(p_{k,m}, g_{k,m}, r_{k,m}) \simeq \kappa_1 \exp\left(\frac{-p_{k,m}\kappa_2 g_{k,m}}{(2^{r_{k,m}} - 1)}\right). \quad (1)$$

(as6) *A realization of each* $g_{k,m}$ *gain falls into one of* $L_{k,m}$ *disjoint regions* $\{\mathcal{R}_{k,m|l}\}_{l=1}^{L_{k,m}}$.

Since users are sufficiently separated in space (as1) is generally true; (as2) corresponds to fading amplitudes adhering to the commonly encountered Rayleigh model but generalizations are possible; (as3) allows each subchannel to vary from one OFDM symbol to the next so long as the quantization region it falls into remains invariant; error-free feedback under (as4) is easily guaranteed with sufficiently strong error control codes (especially since data rates in the feedback link are typically low); the accuracy of (as5) is widely accepted; see, e.g., [3]; and (as6) represents an intuitive quantization that can be easily implemented.

The ultimate *goal* in this paper is twofold: (G1) design a channel quantizer to obtain $\mathbf{c}$, and (G2) given $\mathbf{c}$, find appropriate allocation matrices $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{W}$. We want to design $\mathbf{P}$, $\mathbf{R}$, $\mathbf{W}$, and $\{\mathcal{R}_{k,m|l}\}_{l=1}^{L_{k,m}}$ $\forall k, m$, so that the average power $\bar{P}$ is minimized under prescribed average rate, $\bar{\mathbf{r}}_0 := [\bar{r}_{0,1}, \ldots, \bar{r}_{0,M}]^T$, and average bit error rate (BER), $\bar{\boldsymbol{\epsilon}}_0 := [\bar{\epsilon}_{0,1}, \ldots, \bar{\epsilon}_{0,M}]^T$, constraints across users.

III. Quantizer and Transmission Mode Design

A. Problem Formulation

Given (as6), let $\boldsymbol{\mathcal{R}}_{k,m|l} := \{\mathbf{G} : g_{k,m} \in \mathcal{R}_{k,m|l}\}$ denote the set of matrices $\mathbf{G}$ for which $g_{k,m}$ belongs to the region $\mathcal{R}_{k,m|l}$. Furthermore, let $p_{k,m|l}$ and $r_{k,m|l}$ denote[3] respectively, the

[2]Orthogonal access schemes other than TDMA are also possible. But as we will see later, the one chosen is not particulary important because the optimal choice will typically correspond to no sharing; i.e., each subcarrier will be owned by a single user.

[3]The subscript l here will be also written explicitly as $l(\mathbf{G})$ in places that this dependence must be emphasized.

instantaneous power and rate loadings of user m on subcarrier k given that $\mathbf{G} \in \boldsymbol{\mathcal{R}}_{k,m|l}$. Recall that $w_{k,m}(\mathbf{G}) \leq 1$, and thus the expected power and bit loadings for the $\mathbf{G}$ realization over the time between successive feedback updates will be $p_{k,m|l}w_{k,m}(\mathbf{G})$ and $r_{k,m|l}w_{k,m}(\mathbf{G})$, respectively.

Our goal is to minimize the average transmit-power $\mathbb{E}_{\mathbf{G}}[p_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})]$ over all subcarriers and users while satisfying average rate and BER requirements. Specifically, we want the average rate of any user (say the mth) across all subcarriers to be $\sum_{k=1}^{K}\mathbb{E}_{\mathbf{G}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] \geq [\bar{\mathbf{r}}_0]_m$ (this inequality will be represented as constraint $C1$ in our optimization problem). As for the average BER requirement, one could upper bound with a prespecified maximum BER $[\bar{\boldsymbol{\epsilon}}_0]_m$ the expected number of erroneous bits $\sum_{k=1}^{K}\mathbb{E}_{\mathbf{G}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})\epsilon(p_{k,m|l(\mathbf{G})}, g_{k,m}, r_{k,m|l(\mathbf{G})})]$ over the expected total number of bits $\sum_{k=1}^{K}\mathbb{E}_{\mathbf{G}}[r_{k,m|l(\mathbf{G})}$ $w_{k,m}(\mathbf{G})]$ transmitted by each user $m \in [1, M]$. But since in this average constraint $p_{k,m|l(\mathbf{G})}$ and $r_{k,m|l(\mathbf{G})}$ variables of all $\sum_{k=1}^{K}L_{k,m}$ regions are coupled, it is more convenient to impose a BER constraint where averaging is performed separately over individual regions. In particular, we will upper bound the ratio $\mathbb{E}_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})\epsilon(p_{k,m|l(\mathbf{G})}, g_{k,m}, r_{k,m|l(\mathbf{G})})]/$ $\mathbb{E}_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] < [\bar{\boldsymbol{\epsilon}}_0]_m \ \ \forall k, m, l$ (this inequality will give rise to the constraint $C2$ below).

Analytically, the constrained optimization problem we wish to solve can be written as:

$$
\begin{cases}
\min_{\boldsymbol{\mathcal{R}}_{k,m|l},\mathbf{P}(\mathbf{G}),\mathbf{R}(\mathbf{G}),\mathbf{W}(\mathbf{G})} \ \bar{P}, \quad \text{where } \ \bar{P} := \\
\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{l=1}^{L_{k,m}}\int_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}[p_{k,m|l}w_{k,m}(\mathbf{G})]f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} \\
\text{subject to :} \\
C1. \ \sum_{k=1}^{K}\sum_{l=1}^{L_{k,m}}\int_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}[r_{k,m|l}w_{k,m}(\mathbf{G})]f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} \\
\qquad \geq [\bar{\mathbf{r}}_0]_m, \forall m, \\
C2. \ \int_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}\left[r_{k,m|l}w_{k,m}(\mathbf{G})\kappa_1\exp\left(-\frac{p_{k,m|l}\kappa_2 g_{k,m}}{2^{r_{k,m|l}}-1}\right)\right] \\
\qquad f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} - [\bar{\boldsymbol{\epsilon}}_0]_m\int_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}}[r_{k,m|l}w_{k,m}(\mathbf{G})] \\
\qquad f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} \leq 0, \ \forall k, m, l, \\
C3. \ \sum_{m=1}^{M}w_{k,m}(\mathbf{G}) - 1 \leq 0, \ \forall k, \mathbf{G} \\
C4. \ -\bar{p}_{k,m|l} \leq 0, \ \forall k, m, l, \quad C5. \ -\bar{r}_{k,m|l} \leq 0, \ \forall k, m, l, \\
C6. \ -w_{k,m}(\mathbf{G}) \leq 0, \ \forall k, m, \mathbf{G},
\end{cases}
\tag{2}
$$

where for the second constraint we substituted ϵ from (1) and we separated the numerator and denominator with respect to (w.r.t.) the original $C2$; and through the third constraint $C3$ we enforce the total utilization of any subcarrier by all users not to exceed one, per $\mathbf{G}$ realization.

The optimization in (2) is not jointly convex over all the variables; thus, only local optimality can be guaranteed. However, numerical results will show that the solution of (2) is very close to the P-CSIT benchmark.

The objective in (2) is to minimize the average power over all possible channel realizations. However, the constraints involve different forms of CSI: $C1$ is an average requirement; $C2$ pertains to an average per region; $C3$ needs to be satisfied per channel realization; and $C4-C6$ entail different CSI types pertinent to the constrained variables. In the following subsection, we will derive the Karush-Kuhn-Tucker (KKT) conditions associated with (2). These will lead us not only to the expressions determining the optimal loading variables but will also provide valuable insights about the structure of the power-efficient resource allocation policies. But before presenting the KKT conditions, a remark is due on another aspect related to power efficiency of OFDM-based systems.

Remark 1: Although the peak-to-average-power-ratio (PAPR) plays an important role in power (battery) consumption of OFDM systems, in (2) we did not impose PAPR constraints. The underlying reason is that available digital predistortion schemes can be applied to the users' OFDM symbols to meet such constraints, see e.g., [8].

B. Optimal Policies

Let β_m^r, $\beta_{k,m|l}^{\epsilon}$, β_k^w, $\alpha_{k,m|l}^p$, $\alpha_{k,m|l}^r$, $\alpha_{k,m}^w$ denote the positive Lagrange multipliers associated with $C1$-$C6$, respectively. Setting the derivative of the dual Lagrangian function in (2) w.r.t. $\bar{p}_{k,m|l}$ equal to zero at the optimum[4], $\bar{p}_{k,m|l}^*$, yields after tedious but straightforward manipulations the following KKT condition:

$$
\begin{aligned}
&\int_{\mathbf{G}\in\boldsymbol{\mathcal{R}}_{k,m|l}} w_{k,m}^*(\mathbf{G})\kappa_2 g_{k,m}\left(\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}}{(1-\alpha_{k,m|l}^{p*})p_{k,m|l}^*\ln(2)}-1\right) \\
&\times\kappa_1\exp\left(-\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}-(1-\alpha_{k,m|l}^{p*})p_{k,m|l}^*\ln(2)}{(1-\alpha_{k,m|l}^{p*})\ln(2)}\kappa_2 g_{k,m}\right) \\
&\times f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} = \frac{1-\alpha_{k,m|l}^{p*}}{\beta_{k,m|l}^{\epsilon*}}.
\end{aligned}
\tag{3}
$$

Likewise, differentiating (2) w.r.t. $\bar{r}_{k,m|l}$ and setting the result equal to zero yields at the optimum

$$
r_{k,m|l}^* = \log_2\left(\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}}{\beta_m^{r*}+\alpha_{k,m|l}^{r*}-(1-\alpha_{k,m|l}^{p*})p_{k,m|l}^*\ln(2)}\right).
\tag{4}
$$

KKT conditions for $C4$ and $C5$ also dictate $\bar{p}_{k,m|l}^*\alpha_{k,m|l}^{p*} = 0$ and $\bar{r}_{k,m|l}^*\alpha_{k,m|l}^{r*} = 0$ [1]. These equations imply that $\bar{p}_{k,m|l}^* > 0$ if and only if (iff) $\alpha_{k,m|l}^{p*} = 0$, and $\bar{r}_{k,m|l}^* > 0$ iff $\alpha_{k,m|l}^{r*} = 0$. When $\alpha_{k,m|l}^{p*} \neq 0$ and/or $\alpha_{k,m|l}^{r*} \neq 0$, then $\bar{p}_{k,m|l}^* = \bar{r}_{k,m|l}^* = 0$ and thus the region $\boldsymbol{\mathcal{R}}_{k,m|l}$ is inactive in the sense that it does not affect resource allocation. On the other hand, setting $\alpha_{k,m|l}^{p*} = \alpha_{k,m|l}^{r*} = 0$ in (3) and (4) yields $\bar{p}_{k,m|l}^*/\bar{w}_{k,m|l} < \beta_m^{r*}/\ln(2)$ which must hold for the region $\boldsymbol{\mathcal{R}}_{k,m|l}$ to be active. Intuitively, if the channel in the region $\boldsymbol{\mathcal{R}}_{k,m|l}$ is so poor that for satisfying the BER the power required exceeds the price level represented by $\beta_m^{r*}/\ln(2)$, then the optimum power and rate loadings for this region are zero.

Before analyzing the optimality condition for $w_{k,m}(\mathbf{G})$, let

[4]Henceforth, x^* will denote the optimal value of x.

us first define the power cost of user m utilizing subcarrier k:

$$\mathcal{P}_{k,m}(\mathbf{G}) := p^*_{k,m|l(\mathbf{G})} - \beta^{r*}_m r^*_{k,m|l(\mathbf{G})} + \beta^{\epsilon*}_{k,m|l(\mathbf{G})} \times \left[\kappa_1 \exp\left(-\frac{p^*_{k,m|l(\mathbf{G})}\kappa_2 g_{k,m}}{2^{r^*_{k,m|l(\mathbf{G})}}-1}\right) - [\bar{\boldsymbol{\epsilon}}_0]_m\right], \quad (5)$$

where we made the dependence of β^w_k on $\mathbf{G}$ explicit. (Remember that $\beta^r_m \neq \beta^r_m(\mathbf{G})$ since β^r_m is associated with an average constraint.)

Now supposing that $\boldsymbol{\mathcal{R}}_{k,m|l}$ is active, using (5), and differentiating the Lagrangian of (2) w.r.t. $w_{k,m}(\mathbf{G})$, we find at the optimum

$$\mathcal{P}_{k,m}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G})+\beta^{w*}_k(\mathbf{G})-\alpha^{w*}_{k,m}(\mathbf{G}) = 0,\ \forall\mathbf{G},\ \forall m \in [1,M]. \quad (6)$$

It is useful to check three things: (i) $LHS(6)$ does not depend explicitly on $w^*_{k,m}(\mathbf{G})$ but only through the associated multipliers $\beta^{w*}_k(\mathbf{G})$ and $\alpha^{w*}_{k,m}(\mathbf{G})$; (ii) the multiplier $\beta^{w*}_k(\mathbf{G})$ is common $\forall m$; and (iii) for the same subcarrier k and a given realization $\mathbf{G}$, the power cost $\mathcal{P}_{k,m}(\mathbf{G})$ is fixed and in general different for each user m. Furthermore, for each k, the KKT condition corresponding to $C6$ also dictates

$$w^*_{k,m}(\mathbf{G})\alpha^{w*}_{k,m}(\mathbf{G}) = 0,\ \forall\mathbf{G},\ \forall m \in [1,M]. \quad (7)$$

Since $\mathcal{P}_{k,m}(\mathbf{G})$ is constant for a given $\mathbf{G}$ [c.f. (5)], per subcarrier k, (6) represents an undetermined system of M equations in $M+1$ unknowns, namely $\beta^{w*}_k(\mathbf{G})$ and $\{\alpha^{w*}_{k,m}(\mathbf{G})\}^M_{m=1}$. But if we fix $\mathbf{G}$ and k, we must have $\alpha^{w*}_{k,m}(\mathbf{G}) = 0$ for no more than one m, since otherwise the system of M equations becomes overdetermined and can not be *exactly* satisfied $\forall m \in [1,M]$. On the other hand if $\alpha^{w*}_{k,m}(\mathbf{G}) \neq 0\ \forall m$, then (7) implies $w^*_{k,m}(\mathbf{G}) = 0\ \forall m$ and subcarrier k is wasted $\forall\mathbf{G}$, since no user loads power and rate on it. Because exact solution of (6) requires at most one zero $\alpha^{w*}_{k,m}$ and at least one zero $\alpha^{w*}_{k,m}$ (to avoid the undesirable situation of having $\alpha^{w*}_{k,m}(\mathbf{G}) \neq 0\ \forall m$), it follows that $\alpha^{w*}_{k,m}(\mathbf{G}) = 0$ for exactly one user m per subcarrier k and channel realization $\mathbf{G}$. In other words, the optimal subcarrier allocation allows only one user m_k to transmit on the kth subcarrier. As $\beta^{w*}_k(\mathbf{G})[\sum^M_{m=1} w^*_{k,m}(\mathbf{G}) - 1] = 0$ and $\beta^{w*}_k(\mathbf{G}) \neq 0$, this implies $w^*_{k,m_k}(\mathbf{G}) = 1$ and $w^*_{k,m}(\mathbf{G}) = 0$ for $m \neq m_k$. The next proposition specifies the user m_k who "owns" subcarrier k.

Proposition 1: *The optimal user m_k assigned to utilize the kth subchannel is the one whose subcarrier cost function is minimum, i.e., $m_k = \arg\min_m\{\mathcal{P}_{k,m}(\mathbf{G})\}^M_{m=1}$.*

Proof: Assume that m_k is the candidate user to utilize the subchannel k, i.e., the one for which $w^*_{k,m_k}(\mathbf{G}) = 1$ and $\alpha^{w*}_{k,m_k}(\mathbf{G}) = 0$. For this user, (6) implies that $\beta^{w*}_k(\mathbf{G}) = -\mathcal{P}_{k,m_k}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G})$. Now applying (6) to another user $m'_k \neq m_k$ yields $\mathcal{P}_{k,m'_k}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G}) + \beta^{w*}_k(\mathbf{G}) - \alpha^{w*}_{k,m'_k}(\mathbf{G}) = [\mathcal{P}_{k,m'_k}(\mathbf{G}) - \mathcal{P}_{k,m_k}(\mathbf{G})]f_{\mathbf{G}}(\mathbf{G}) - \alpha^{w*}_{k,m'_k}(\mathbf{G}) = 0$. Satisfying the latter requires $\mathcal{P}_{k,m'_k}(\mathbf{G}) \geq \mathcal{P}_{k,m_k}(\mathbf{G})$, since $\alpha^{w*}_{k,m'_k}(\mathbf{G}) \geq 0$; that is, $\mathcal{P}_{k,m_k}(\mathbf{G}) = \min_m\ \mathcal{P}_{k,m}(\mathbf{G})$. ■

Notice that if the minimum value of $\{\mathcal{P}_{k,m}\}^M_{m=1}$ is attained by more than one user, any arbitrary time sharing of the subcarrier k among them is optimum; or we can simply pick one of them at random (the power cost is invariant). Finally, if there exists a realization $\mathbf{G}$ such that $\mathcal{P}_{k,m}(\mathbf{G}) > 0\ \forall m$, then since $\beta^{w*}_k(\mathbf{G}) \geq 0$ it follows from (6) that $\alpha^{w*}_{k,m}(\mathbf{G}) \neq 0,\ \forall m$; and the optimal solution will not allocate this subcarrier to any user. Therefore, introducing a fictitious $\mathcal{P}_{k,0}(\mathbf{G}) = 0\ \forall k, \mathbf{G}$, we can express $w^*_{k,m}(\mathbf{G})$ in compact form using the indicator function as

$$w^*_{k,m}(\mathbf{G}) = \mathbf{I}_{\{m=\arg\min_{m'}\{\mathcal{P}_{k,m'}(\mathbf{G})\}^M_{m'=0}\}}. \quad (8)$$

Remark 2: From (3), (4) and (8), we can readily infer that: (i) the only coupling among subcarriers is through the multiplier β^r_m (i.e., given $\beta^r_m\ \forall m$, the allocation of rate and power on each subcarrier can be performed independently); (ii) given β^r_m, the optimal rate and power allocation for user m does not depend on the loadings in other regions; and (iii) for a subcarrier k, optimal assignment of users amounts to satisfying (6) jointly $\forall m$.

Once $\mathbf{P}^*, \mathbf{R}^*$, and $\boldsymbol{W}^*$ are characterized, we must find the optimum regions $\{\boldsymbol{\mathcal{R}}^*_{k,m|l}\}^{L_{k,m}}_{l=1}\quad \forall(k,m)$ for optimizing $\bar{P}$ in (2). Before proceeding, with this joint optimization, let us recall that: (i) the optimum resource allocation in Section III can be decomposed for each user m and subcarrier k; and (ii) $\boldsymbol{\mathcal{R}}_{k,m|l}$ represents a quantization region of a single variable $g_{k,m}$. In fact, $\{\boldsymbol{\mathcal{R}}^*_{k,m|l}\}^{L_{k,m}}_{l=1}$ can be equivalently represented by a set of thresholds $\{\tau^*_{k,m|l}\}^{L_{k,m}+1}_{l=1}$, with $\tau^*_{k,m|1} = 0$ and $\tau^*_{k,m|L_{k,m}+1} = \infty\ \forall(k,m)$. In other words, thanks to (i) and (ii) our vector quantization problem reduces to KM scalar quantization problems.

The necessary KKT condition, to find $\{\tau^*_{k,m|l}\}^{L_{k,m}}_{l=2}\ \forall k,m$ sets the derivative of the Lagrangian function $\mathcal{L}$ of (2) to zero, yielding

$$p_{k,m|l-1} - \beta^r_m r_{k,m|l-1} + \beta^\epsilon_{k,m|l-1}\kappa_1 e^{-\frac{p_{k,m|l-1}\kappa_2\tau^*_{k,m|l}}{2^{r_{k,m|l-1}}-1}} = -p_{k,m|l} + \beta^r_m r_{k,m|l} - \beta^\epsilon_{k,m|l}\kappa_1 e^{-\frac{p_{k,m|l}\kappa_2\tau^*_{k,m|l}}{2^{r_{k,m|l}}-1}} \quad (9)$$

which can be solved for $\tau^*_{k,m|l}$ using line search to find the wanted thresholds $\forall(k,m)$ pair.

So far, we obtained the conditions that the optimal allocation policies must satisfy. We next outline the steps of an algorithm that can be implemented to fulfill these conditions.

Algorithm 1: *Generic Resource Allocation*

(S1.0) Let δ be a small positive number and $\boldsymbol{\beta^r}$ the vector formed by $\{\beta^r_m\}^M_{m=1}$. Start with arbitrary non-negative $\boldsymbol{\beta^r}$.

(S1.1) For each subcarrier k:

(S1.1.1) Set arbitrary non-negative $\beta^\epsilon_{k,m|l}\ \forall m,l$ and $\{\tau_{k,m|l}\}^{L_{k,m}}_{l=2}\ \forall m$ such that $\tau_{k,m|l} < \tau_{k,m|l+1}$.

(S1.1.2) Set initial $p_{k,m|l} = \bar{p}_{k,m|l}/\bar{w}_{k,m}$, such that $0 < p_{k,m|l} < \beta^r_m/\ln(2),\ \forall m,l$.

(S1.1.3) For $\alpha^p_{k,m|l} = 0$ and $\alpha^r_{k,m|l} = 0$, use (4) to obtain $r_{k,m|l} = \bar{r}_{k,m|l}/\bar{w}_{k,m}\ \forall m,l$.

(S1.1.4) Find $w_{k,m}(\mathbf{G})$ $\forall m$ as in (8).

(S1.1.5) Check (3) $\forall m,l$. If $|LHS(3) - 1/\beta^{\epsilon}_{k,m|l}| < \delta\beta^{\epsilon}_{k,m|l}$ $\forall m,l$ go to (S1.1.6); otherwise increase $p_{k,m|l}$ if (m,l) is such that $LHS(3) > 1/\beta^{\epsilon}_{k,m|l}$; decrease $p_{k,m|l}$ if (m,l) is such that $LHS(3) < 1/\beta^{\epsilon}_{k,m|l}$, and go to (S1.1.3).

(S1.1.6) Check constraint $C2$ in (2) $\forall m,l$. If $|C2| < \delta[\bar{\boldsymbol{\epsilon}}_0]_m$ $\forall m,l$, move to the next subcarrier $k+1$, and go to (S1.1.7); otherwise increase $\beta^{\epsilon}_{k,m|l}$ if (m,l) is such that $C2 > 0$; decrease $\beta^{\epsilon}_{k,m|l}$ if (m,l) is such that $C2 < 0$, and go to (S1.1.3).

(S1.1.7) Given $\{p_{k,m|l}, r_{k,m|l}, \beta^{\epsilon}_{k,m|l}\}_{l=1}^{L_{k,m}}, w_{k,m}(\mathbf{G})$ and β^{r}_{m} $\forall m$ from (S1.1.6), update the values of $\{\tau_{k,m|l}\}_{l=2}^{L_{k,m}}$ $\forall m$ as in (9). If the change w.r.t. the previous values of $\{\tau_{k,m|l}\}_{l=2}^{L_{k,m}}$ is smaller than δ, then go to (S1.1); otherwise go back to (S1.1.3).

(S1.2) Check constraint $C1$ in (2) $\forall m$. If $|LHS(C1) - [\bar{\mathbf{r}}_0]_m| < \delta[\bar{\mathbf{r}}_0]_m$ $\forall m$ then *Stop*; otherwise increase β^{r}_{m} if a user index m is such that $[\bar{\mathbf{r}}_0]_m > C1$; decrease β^{r}_{m} if m is such that $[\bar{\mathbf{r}}_0]_m < C1$, and go to (S1.1).

Performance and convergence of Algorithm 1 will be clearly affected by the schemes used to increase/decrease $p_{k,m|l}, \beta^{\epsilon}_{k,m|l}$, and β^{r}_{m} in steps (S1.1.5), (S1.1.6), and (S1.2). Different recursive updates for these steps can be found in e.g., [1]. We remark that although (9) returns a quantizer that attains a local optimum, global optimality is not guaranteed due to lack of convexity.

Remark 3: Instead of discrete-rate (DR) loadings, optimization throughout this paper is carried out for continuous-rate (CR). The hardware complexity for implementing CR modulations (through non-square constellations) is higher than the required for DR [3], but leads to more power savings compared with DR. Furthermore, it turns out that CR performs very close to the DR solution and can be optimally transformed to it by extending the results in [4].

IV. Codeword Structure

Given the quantizer design, we developed so far resource allocation policies to assign rate, power and subcarriers across users. Once the quantizer and resource allocation strategy are designed, the AP quantizes each fading state and feeds back a codeword that identifies the user-subcarrier assignment and the region index each subchannel falls into per fading realization $\mathbf{G}$. Based on this form of Q-CSIT, each user is informed about its own subset of subcarriers (if any) and relies on the region indices to retrieve the corresponding power and rate levels from a lookup table. The following proposition describes the construction of this codeword.

Proposition 2: *Given the quantizer design and the optimal allocation parameters* $(\mathbf{P}^*, \mathbf{R}^*, \boldsymbol{W}^*(\boldsymbol{\beta}^{r*}), \{\beta^{\epsilon*}\}_{l=1}^{L_{k,m}})$ *returned by Algorithm 1, the AP broadcasts to the users the codeword* $\mathbf{c}^*(\mathbf{G}) = [\mathbf{c}_1^*(\mathbf{G}), \ldots, \mathbf{c}_K^*(\mathbf{G})]$ *specifying the optimal resource allocation for the current fading state, where* $\mathbf{c}_k^*(\mathbf{G}) = [m_k^*(\mathbf{G}), l_k^*(\mathbf{G})]^T$ *is determined* $\forall k$ *as:*

1) $m_k^*(\mathbf{G}) = \arg\min_m \{\mathcal{P}_{k,m}(\mathbf{G}, \mathbf{P}^*, \mathbf{R}^*, \boldsymbol{\beta}^{r*}, \{\beta^{\epsilon*}\}_{l=1}^{L_{k,m}})\}_{m=1}^{M}$ *(pick randomly any user* m_k^* *when multiple minima occur); and*

2) $l_k^*(\mathbf{G}) = \{\ l \mid \mathbf{G} \in \boldsymbol{\mathcal{R}}_{k,m_k^*(\mathbf{G}),l},\ \ l = 1, \ldots, L_k\}$.

The structure of $\mathbf{c}^*(\mathbf{G})$ in Proposition 2 encodes information pertinent to each subcarrier (namely, its region and assigned user) which is more efficient in terms of the number of feedback bits relative to encoding each user's individual information (i.e., set of subcarriers and corresponding regions). Since in each subcarrier we have $L_k - 1$ active regions and one inactive or outage region, we can save additional feedback bits by encoding only the active regions. Only when all users' channel gains belong to inactive regions, we will need to index an outage for the corresponding subcarrier. This can be readily done by indexing a virtual user (e.g., $m = 0$) with a unique region. Including all these indices, the codeword length will be $\left\lceil \sum_{k=1}^{K} \log_2 \left(\sum_{m=1}^{M} (L_{k,m} - 1) + K \right) \right\rceil$ bits.

We conclude this section by emphasizing that $\mathbf{P}$, $\mathbf{R}$, and $\{\mathcal{R}_{k,m|l}\}_{l=1}^{L_{k,m}} \forall k,m$, in (2) are involved only in average quantities. Hence, $\mathbf{P}^*$, $\mathbf{R}^*$, and $\{\mathcal{R}^*_{k,m|l}\}_{l=1}^{L_{k,m}} \forall k,m$, are computed off-line and only the subcarrier-user assignment (involved in instantaneous constraints) and the indexing of the corresponding entries of these matrices need to be fed back on-line. Thus, almost all the complexity is carried out off-line (Algorithm 1), while only a light computation (Proposition 2) has to be carried out on-line.

V. Numerical Examples

To numerically test our power-efficient designs, we consider an adaptive OFDMA system with $M = 3$ users, $K = 64$ subcarriers, noise power per user and subcarrier at 0 dB_W, $L_{k,m} = 5$ regions (i.e., 4 active regions) per subcarrier, and $[\bar{\boldsymbol{\epsilon}}_0]_m = \bar{\epsilon}_0 = 10^{-3}$ $\forall m$. The average signal-to-noise-ratio ($\overline{\text{SNR}}$) considered was 0 dB; and three uncorrelated Rayleigh taps were simulated per user.

Test Case 1 *(Comparison of allocation schemes):* For different $\overline{\text{SNR}}$ values and $\bar{\mathbf{r}}_0 = [20,\ 40,\ 60]^T$, Fig. 2 depicts the power gap of two different Q-CSIT allocation schemes based on Q-CSIT: Q-CSIT-A1 and Q-CSIT-US, w.r.t. the benchmark allocation based on P-CSIT. Q-CSIT-A1 represents the allocation scheme when Algorithm 1 is implemented, while Q-CSIT-US is a heuristic scheme with uniform subcarrier allocation and optimum rate and power loading. The striking observation here is the almost equivalent performance of Q-CSIT-A1 and P-CSIT schemes. These results certify the usefulness of optimum Q-CSIT-A1 and validate the optimality of Algorithm 1. Moreover, the power loss of at least 6 dB w.r.t. Q-CSIT-US shows the important role subcarrier allocation plays in terms of minimizing the transmit-power.

Test Case 2 *(Different parameter values):* Numerical results assessing the performance of P-CSIT, Q-CSIT-A1 and Q-CSIT-US schemes over a wide range of parameter values are summarized in Table I, where the reference case is $K = 64$, $M = 3$, $\bar{\mathbf{r}}_0 = [60,\ 60,\ 60]^T$, $\overline{\text{SNR}} = 0$ dB; and the column "case" entails only a single variation w.r.t. the reference case.

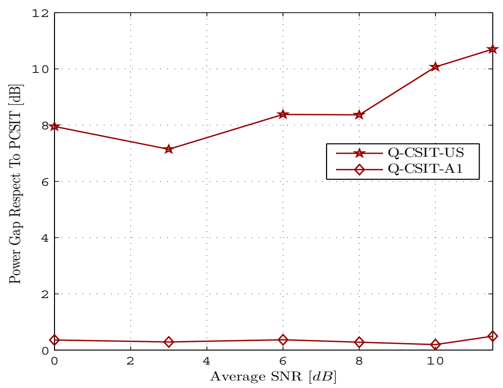

Fig. 2. Comparison of the transmit-power gap w.r.t. the PCSIT solution for different allocation schemes based on Q-CSIT ($K = 64$, $M = 3$, $\bar{\mathbf{r}}_0 = [20,\ 40,\ 60]^T$).

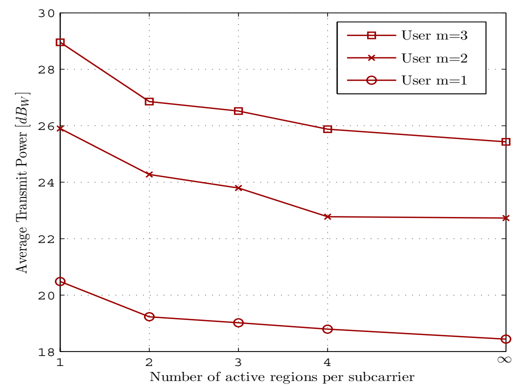

Fig. 3. Effect of the number of quantization regions (feedback bits) per subcarrier using Algorithm 1 ($K = 64$, $M = 3$, $\bar{\mathbf{r}}_0 = [20,\ 40,\ 60]^T$).

TABLE I

TOTAL AVERAGE TRANSMIT POWER (IN dB_W) FOR P-CSIT, Q-CSIT-A1 AND Q-CSIT-US SCHEMES. (CASE ENTAILS ONLY ONE VARIATION W.R.T. THE REFERENCE CASE.)

CASE	Q-CSIT-US	Q-CSIT-A1	P-CSIT
Reference Case	38.9	31.5	31.2
$\bar{\epsilon}_0 = 10^{-4}$	40.6	33.3	32.8
$\bar{\mathbf{r}}_0 = [30,\ 30,\ 30]^T$	31.3	26.6	26.3
$K = 128$	34.7	30.1	29.2
$M = 6$	46.2	39.8	39.2

We observe that these results confirm our previous conclusions, namely: (i) the near optimality of Q-CSIT-A1, and (ii) the performance loss exhibited by the heuristic schemes exemplified by Q-CSIT-US.

Test Case 3 *(Number of quantization regions):* Finally, Fig. 3 plots the average transmit-power versus the number of active regions per subcarrier for $\bar{\mathbf{r}}_0 = [20,\ 40,\ 60]^T$. Recall that the number of active regions is equal to $L_{k,m} - 1$; e.g., $L_{k,m} = 2$ implies one active region and one outage region. Simulation results in this figure demonstrate that joint optimization of resource allocation and quantizing thresholds leads to a power loss no greater than 3-5 dB w.r.t. the P-CSIT case ($L_{k,m} = \infty$). Moreover, the resulting power gap shrinks as the number of regions increases reaching a power loss of approximately only 0.5 dB in the case of four active regions.

VI. CONCLUDING SUMMARY AND FUTURE RESEARCH

Based on Q-CSIT, we devised a power-efficient OFDMA scheme under prescribed individual average rate and BER constraints. In this setup, an access point quantizes the subcarrier gains and feeds back to the users a codeword conveying the optimum power, rate, and subcarrier allocation. The resulting near-optimal transceivers are attractive because they only incur a power loss as small as 1 dB relative to the benchmark design based on P-CSIT which requires often unrealistic feedback information.

We proposed an optimal design that jointly optimizes over power, rate and subcarriers across users as well as quantization regions. We ended up with a lightweight resource allocation protocol where both rate and power are available at the transmitter through a lookup table and only the subcarrier assignment needs be determined on-line.

To build on the presented framework, interesting future directions include reduction of the complexity of the optimum algorithm as well as further reduction of the feedback overhead by exploiting the possible correlation across subcarriers to group subcarriers and then index each group; or, by applying differential quantization techniques along the lines of [5].[5]

REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[2] M. Cho, W. Seo, Y. Kim, D. Hong, "A Joint Feedback Reduction Scheme Using Delta Modulation for Dynamic Channel Allocation in OFDMA Systems", *Proc. of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Commun.*, Berlin, Germany, vol. 4, pp. 2747–2750, Sep. 2005.

[3] A. J. Goldsmith and S. -G. Chua, "Variable-rate variable-power M-QAM for fading channels," *IEEE Trans. on Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.

[4] B. S. Krongold, K. Ramchandran, and D. L. Jones, "Computationally efficient optimal power allocation algorithms for multicarrier communication systems," *IEEE Commun. Lett.*, vol. 48, pp. 23–27, Jan. 2000.

[5] J. H. Kwon, D. Rhee, I. M. Byun, Y. Whang, K. S. Kim, "Adaptive Modulation Technique with Partial CQI for Multiuser OFDMA Systems," *Proc. of Int. Conf. on Advanced Commun. Technology*, Phoenix Park, Korea, vol. 2, pp. 1283–1286, Feb. 2006.

[6] A. Lapidoth and S. Shamai, "Fading channels: how perfect need 'perfect side information' be?," *IEEE Trans. on Info. Theory*, vol. 48, no. 5, pp. 1118-1134, May 2002.

[7] A. G. Marques, F. F. Digham, and G. B. Giannakis, "Optimizing power efficiency of OFDM using quantized channel state information," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1581–1592, Aug. 2006.

[8] H. Qian, C. Xiao, N. Chen, and G. T. Zhou, "Dynamic selected mapping for OFDM," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, pp. 325–328, Mar. 2005.

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

[10] S. Zhou and B. Li, "BER criterion and codebook construction for finite-rate precoded spatial multiplexing with linear receivers," *IEEE Trans. on Signal Proc.*, vol. 54, no. 5, pp. 1653–1665, May 2006.

[5]The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U. S. Government.