Power-Efficient Wireless OFDMA Using Limited-Rate Feedback

Antonio G. Marques, *Member, IEEE*, Georgios B. Giannakis, *Fellow, IEEE*, Fadel F. Digham, *Member, IEEE*, and F. Javier Ramos, *Member, IEEE*

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. Along with the characterization of optimal bit, power and subcarrier allocation policies based on LRF, suboptimal yet simple schemes are developed for channel quantization. The novel algorithms proceed in two phases: (i) an off-line phase to construct the channel quantizer as well as the rate and power codebooks with moderate complexity; and (ii) an on-line phase to obtain, based on quantized channel state information, the optimum, rate, power and user-subcarrier allocation with linear complexity. Numerical examples corroborate the analytical claims and reveal that significant power savings result even with suboptimal schemes based on practically affordable LRF.**

Index Terms— **Feedback communication, frequency division multiaccess, minimum energy control, optimization methods, quantization, resource management.**

I. INTRODUCTION

ORTHOGONAL frequency-division multiplexing (OFDM) is the most common modulation for bandwidth limited wireline and wireless transmissions over frequency-selective multipath channels. Testament to its well-documented merits is provided by the fact that OFDM has been adopted by e.g., digital subscriber line (DSL) modems, digital audio and video broadcasting (DAB/DVB) standards and wireless local area networks [6], [14]. Spectral efficiency and error resilience in these applications are well known to improve with the knowledge of channel state information (CSI) at the transmitter (CSIT). For this reason, OFDM transmissions over wireline or slowly fading wireless links have traditionally relied on deterministic or *perfect* (P-) 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 [15].

While the assumptions of P-CSI at the transmitters and receiver render analysis and design tractable, they may not be as realistic due to wireless channel variations and estimation errors, feedback delay, bandwidth limitation, and jamming induced errors [8], [11]. 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., [12] and [17]. Q-CSIT entails a finite number of quantization regions describing different clusters of channel realizations [10], [16]. 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-CSIT is robust to channel uncertainties since transmitters adapt to a few regions rather than individual channel realizations. Prompted by these LRF features, we recently introduced a power-efficient OFDM scheme based on Q-CSIT for a point-to-point link [10]. In this paper, we considerably broaden the scope of [10] by casting our design in the challenging multi-user, i.e., orthogonal frequency-division multiple access (OFDMA), setup.

Resource allocation in OFDMA minimizing the transmit-power *per symbol* based on P-CSIT was first studied in [15]. Relying on fixed (as opposed to adaptive) Q-CSIT, recent works deal with optimization of power or rate performance *per* OFDMA *symbol* [2], [7]. 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. Specifically, we aim at on-line algorithms whose complexity is linear in the number of users and allocate optimally power, rate and user subcarriers per channel realization. Moreover, we rely on a general framework for modeling the Q-CSIT, which besides optimal resource allocation contributes jointly optimal designs for the channel quantizer.

The rest of the paper is organized as follows. After introducing preliminaries on the setup we deal with (Section II), we derive optimal and suboptimal subcarrier allocation as well as power and bit loading policies for OFDMA based on LRF for a given quantizer design (Section III). Once the optimum policies are characterized, we subsequently design an optimum quantizer and then derive a reduced-complexity alternative (Section IV). Numerical results and comparisons that corroborate our claims are presented (Section V), and concluding remarks wrap up this paper (Section VI).

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 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).

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 per subcarrier; i.e., time division multiple access (TDMA)[1]. This sharing process is described by the $K \times M$ 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 $N_{\max} := \max_{m \in [1,M]} N_m$. 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
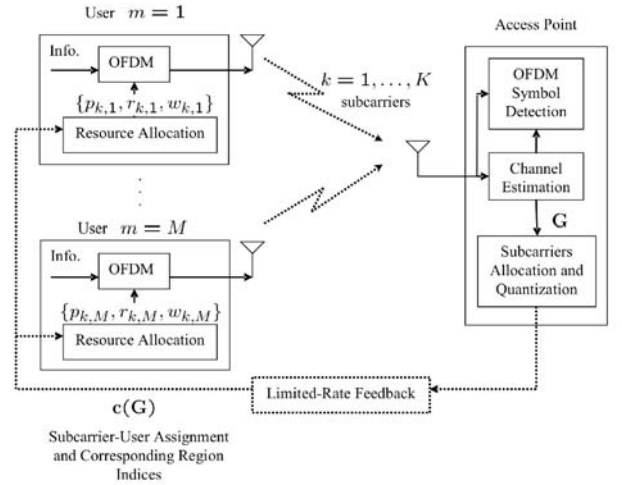

Fig. 1.   System block diagram.

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.

Channel estimation at the receiver relies on periodically inserted training (pilot) symbols. With the multi-user channel matrix $\tilde{\mathbf{H}}$ acquired, the receiver has available the channel-gain-to-noise-variance ratio 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.

The ultimate *goals* in this paper are to: (G1) design a channel quantizer to obtain $\mathbf{c}$, and (G2) given $\mathbf{c}$, find appropriate allocation matrices $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{W}$. Joint solution of this problem is analytically and computationally cumbersome, if not practically infeasible[2]. For this reason, our approach in this paper will proceed in two stages. First, we fix the quantizer design and rely on its instantaneous output $\mathbf{c}$ to design $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{W}$ so that the average power $\bar{P} := \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbb{E}_{\mathbf{G}}[p_{k,m}(\mathbf{G})w_{k,m}(\mathbf{G})]$ 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

---

[1]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.

[2]In contrast with designs based on P-CSIT, where the adaptation schemes can be found as the (sometimes closed-form) solution of convex optimization problems, designs based on Q-CSIT amount to solving optimally vector quantization problems that are non-convex and typically NP-hard.

across users. Subsequently, we design judiciously the channel quantizer to attain desirable tradeoffs among implementation complexity, feedback overhead and power efficiency. These tasks are carried out under 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 ($\kappa_1 = 0.2$, $\kappa_2 = 1.5$)*

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

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); and the accuracy of (as5) is widely accepted; see, e.g., [5].

III. TRANSMISSION MODE DESIGN

When designing the optimum transmission modes, we further suppose that:

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

Throughout this section the set $\{\mathcal{R}_{k,m|l}\}_{l=1}^{L_{k,m}}$ will be fixed. Selecting the set $\{\mathcal{R}_{k,m|l}\}_{l=1}^{L_{k,m}}$ is part of the quantizer design problem to be addressed in Section IV. Both the quantizer design as well as most of the transmission mode selection process in this section will entail off-line computations. In fact, we will see that only a simple algorithm needs to be executed on-line per fading realization in order to specify the matrices $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{W}$.

A. Problem Formulation

Given (as6), let $\mathcal{R}_{k,m|l} := \{\mathbf{G} : g_{k,m} \in \mathcal{R}_{k,m|l}\}$ 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 instantaneous power and rate loadings of user m on subcarrier k given that $\mathbf{G} \in \mathcal{R}_{k,m|l}$. Recall that $w_{k,m}(\mathbf{G}) \leq 1$, and thus the expected power and bit loadings for the $\mathbf{G}$ realization

---

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

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 no less than a prespecified minimum rate, i.e., $\sum_{k=1}^{K} \mathbb{E}_{\mathbf{G}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] \geq [\bar{\mathbf{r}}_0]_m$. As for the average BER requirement, one could upper bound with a prespecified maximum BER $[\bar{\epsilon}_0]_m$ 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 $\mathbb{E}_{\mathbf{G} \in \mathcal{R}_{k,m|l}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})\epsilon(p_{k,m|l(\mathbf{G})}, g_{k,m}, r_{k,m|l(\mathbf{G})})]/\mathbb{E}_{\mathbf{G} \in \mathcal{R}_{k,m|l}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})]\ \forall k, m, l$. Since the latter is a stricter constraint, the power consumption may be slightly higher. Setting the BER constraint per region, sub-carrier and user eliminates one degree of freedom in our design, but still leaves three degrees of freedom per region (namely thresholds, rate and power). As a consequence, the resultant performance does not degrade noticeably. Simulations will corroborate that in practice this degradation is indeed negligible.

Analytically, the constrained optimization problem we wish to solve is:

$$
\begin{cases}
\min_{\mathbf{P}(\mathbf{G}) \geq \mathbf{0}, \mathbf{R}(\mathbf{G}) \geq \mathbf{0}, \mathbf{W}(\mathbf{G}) \geq \mathbf{0}} \quad \bar{P}, \\
\quad \text{where } \bar{P} := \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbb{E}_{\mathbf{G}}[p_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] \\
\text{subject to :} \\
\sum_{k=1}^{K} \mathbb{E}_{\mathbf{G}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] \geq [\bar{\mathbf{r}}_0]_m, \quad \forall m; \\
\mathbb{E}_{\mathbf{G} \in \mathcal{R}_{k,m|l}}\left[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})\kappa_1 \exp\left(-\frac{p_{k,m|l(\mathbf{G})}\kappa_2 g_{k,m}}{2^{r_{k,m|l(\mathbf{G})}} - 1}\right)\right] \\
/\mathbb{E}_{\mathbf{G} \in \mathcal{R}_{k,m|l}}[r_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})] \leq [\bar{\epsilon}_0]_m, \quad \forall k, m, l; \\
\sum_{m=1}^{M} w_{k,m}(\mathbf{G}) \leq 1, \quad \forall k, \mathbf{G};
\end{cases}
\tag{2}
$$

where for the second constraint we substituted from (1), and through the third constraint we enforce the total utilization of any subcarrier by all users not to exceed one, per $\mathbf{G}$ realization.

Since the problem in (2) is non-convex for all $\mathbf{G}$, to render it more tractable we introduce the average weights $\bar{w}_{k,m|l} := \int_{\mathbf{G} \in \mathcal{R}_{k,m|l}} w_{k,m}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G})d\mathbf{G}$, and corresponding variables $\bar{p}_{k,m|l} := \int_{\mathbf{G} \in \mathcal{R}_{k,m|l}} p_{k,m|l(\mathbf{G})}w_{k,m}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G})d\mathbf{G} = p_{k,m|l}\bar{w}_{k,m|l}$ and $\bar{r}_{k,m|l} := r_{k,m|l}\bar{w}_{k,m|l}$. Recall that $p_{k,m|l}$ and $r_{k,m|l}$ are region specific and thus represent deterministic quantities over the region they index. With this change of

variables, (2) can be re-written as:

$$
\begin{cases}
\min_{\bar{p}_{k,m|l},\bar{r}_{k,m|l},w_{k,m}(\mathbf{G})}\ \bar{P}, \\
\quad \text{where } \bar{P} := \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{l=1}^{L_{k,m}}\bar{p}_{k,m|l} \\
\text{subject to :} \\
C1.\ -\sum_{k=1}^{K}\sum_{l=1}^{L_{k,m}}\bar{r}_{k,m|l}+[\bar{\mathbf{r}}_0]_m \le 0,\ \ \forall m; \\
C2.\ \mathbb{E}_{\mathbf{G}\in\mathcal{R}_{k,m|l}}\left[w_{k,m}(\mathbf{G})\kappa_1 \exp\left(-\dfrac{\frac{\bar{p}_{k,m|l(\mathbf{G})}}{\bar{w}_{k,m|l}}\kappa_2 g_{k,m}}{2^{\frac{\bar{r}_{k,m|l(\mathbf{G})}}{\bar{w}_{k,m|l}}}-1}\right)\right] \\
\qquad\qquad\qquad\qquad -[\bar{\boldsymbol{\epsilon}}_0]_m\bar{w}_{k,m|l}\le 0,\ \ \forall k,m,l; \\
C3.\ \sum_{m=1}^{M}w_{k,m}(\mathbf{G})-1 \le 0,\ \ \forall k,\mathbf{G}; \\
C4.\ -\bar{p}_{k,m|l}\le 0,\ \ \forall k,m,l; \\
C5.\ -\bar{r}_{k,m|l}\le 0,\ \ \forall k,m,l; \\
C6.\ -w_{k,m}(\mathbf{G})\le 0,\ \ \forall k,m,\mathbf{G};
\end{cases}
\tag{3}
$$

The optimization problem in (3) is in general not jointly convex in the variables $\bar{p}_{k,m|l}$, $\bar{r}_{k,m|l}$ and $w_{k,m}$ for some $\mathbf{G}$. For this reason, we will relax it in Section III-C to obtain a convex minimization problem for all $\mathbf{G}$, and thus ensure availability of solvers with with guaranteed convergence to the global minimum.

The objective in (3) 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 (3). 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., [13].

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 Lagrangian in (3) with respect to (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

---

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

KKT condition:

$$
\int_{\mathbf{G}\in\mathcal{R}_{k,m|l}} w_{k,m}^*(\mathbf{G})\kappa_2 g_{k,m}\left(\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}}{(1-\alpha_{k,m|l}^{p*})\frac{\bar{p}_{k,m|l}^*}{\bar{w}_{k,m|l}}\ln(2)}-1\right)
$$
$$
\times\kappa_1\exp\left(-\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}-(1-\alpha_{k,m|l}^{p*})\frac{\bar{p}_{k,m|l}^*}{\bar{w}_{k,m|l}}\ln(2)}{(1-\alpha_{k,m|l}^{p*})\ln(2)}\kappa_2 g_{k,m}\right)
$$
$$
\times f_{\mathbf{G}}(\mathbf{G})d\mathbf{G}=\frac{1-\alpha_{k,m|l}^{p*}}{\beta_{k,m|l}^{\epsilon*}}.
\tag{4}
$$

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

$$
\frac{\bar{r}_{k,m|l}^*}{\bar{w}_{k,m|l}}=\log_2\left(\frac{\beta_m^{r*}+\alpha_{k,m|l}^{r*}}{\beta_m^{r*}+\alpha_{k,m|l}^{r*}-(1-\alpha_{k,m|l}^{p*})\frac{\bar{p}_{k,m|l}^*}{\bar{w}_{k,m|l}}\ln(2)}\right).
\tag{5}
$$

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 $\alpha_{k,m|l}^{p*}\ne 0$ and/or $\alpha_{k,m|l}^{r*}\ne 0$, then $\bar{p}_{k,m|l}^*=\bar{r}_{k,m|l}^*=0$ and thus the region $\mathcal{R}_{k,m|l}$ 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 (4) and (5) yields $\bar{p}_{k,m|l}^*/\bar{w}_{k,m|l}<\beta_m^{r*}/\ln(2)$ which must hold for the region $\mathcal{R}_{k,m|l}$ to be active. Intuitively, if the channel in the region $\mathcal{R}_{k,m|l}$ 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.

Supposing that $\mathcal{R}_{k,m|l}$ is active, and differentiating the Lagrangian of (3) w.r.t. $w_{k,m}(\mathbf{G})$, we find at the optimum

$$
\begin{cases}
\dfrac{\bar{p}_{k,m|l}^*}{\bar{w}_{k,m|l}}-\beta_m^{r*}\dfrac{\bar{r}_{k,m|l}^*}{\bar{w}_{k,m|l}}+\beta_{k,m|l(\mathbf{G})}^{\epsilon*}\left[\kappa_1\exp\left(-\dfrac{\frac{\bar{p}_{k,m|l}^*}{\bar{w}_{k,m|l}}\kappa_2 g_{k,m}}{2^{\frac{\bar{r}_{k,m|l}^*}{\bar{w}_{k,m|l}}}-1}\right)\right. \\
\qquad\left.-[\bar{\boldsymbol{\epsilon}}_0]_m\right]\Big\}f_{\mathbf{G}}(\mathbf{G})+\beta_k^{w*}(\mathbf{G})-\alpha_{k,m}^{w*}(\mathbf{G})=0,
\end{cases}
\tag{6}
$$

where we made the dependence of β_k^w on $\mathbf{G}$ explicit. (Remember that $\beta_m^r\ne\beta_m^r(\mathbf{G})$ since β_m^r is associated with an average constraint.) Recall also that if $\bar{w}_{k,m|l}\ne 0$, then $\bar{p}_{k,m|l}^*/\bar{w}_{k,m|l}:=p_{k,m|l}^*$ and $\bar{r}_{k,m|l}^*/\bar{w}_{k,m|l}:=r_{k,m|l}^*$ in (4)-(6). To appreciate the implications of (6), let us define

$$
\begin{aligned}
\mathcal{P}_{k,m}(\mathbf{G}) :=\ & p_{k,m|l(\mathbf{G})}^*-\beta_m^{r*}r_{k,m|l(\mathbf{G})}^*+\beta_{k,m|l(\mathbf{G})}^{\epsilon*} \\
&\times\left[\kappa_1\exp\left(-\frac{p_{k,m|l(\mathbf{G})}^*\kappa_2 g_{k,m}}{2^{r_{k,m|l(\mathbf{G})}^*}-1}\right)-[\bar{\boldsymbol{\epsilon}}_0]_m\right],
\end{aligned}
\tag{7}
$$

to represent the *power cost* of user m utilizing subcarrier k. Then (6) can be rewritten as

$$
\mathcal{P}_{k,m}(\mathbf{G})f_{\mathbf{G}}(\mathbf{G})+\beta_k^{w*}(\mathbf{G})-\alpha_{k,m}^{w*}(\mathbf{G})=0,\ \forall\mathbf{G},\ \forall m\in[1,M].
\tag{8}
$$

It is useful to check three things: (i) $LHS(8)$ does not depend explicitly on $w_{k,m}^*(\mathbf{G})$ but only through the associated multipliers $\beta_k^{w*}(\mathbf{G})$ and $\alpha_{k,m}^{w*}(\mathbf{G})$; (ii) the multiplier $\beta_k^{w*}(\mathbf{G})$ 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

$$w_{k,m}^*(\mathbf{G})\alpha_{k,m}^{w*}(\mathbf{G}) = 0, \ \forall \mathbf{G}, \ \forall m \in [1, M]. \quad (9)$$

Since $\mathcal{P}_{k,m}(\mathbf{G})$ is constant for a given $\mathbf{G}$ [cf. (7)], per subcarrier k (8) represents an undetermined system of M equations in $M + 1$ unknowns, namely $\beta_k^{w*}(\mathbf{G})$ and $\{\alpha_{k,m}^{w*}(\mathbf{G})\}_{m=1}^M$. But if we fix $\mathbf{G}$ and k, we must have $\alpha_{k,m}^{w*}(\mathbf{G}) = 0$ 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 $\alpha_{k,m}^{w*}(\mathbf{G}) \neq 0 \ \forall m$, then (9) 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 (8) requires at most one zero $\alpha_{k,m}^{w*}$ and at least one zero $\alpha_{k,m}^{w*}$ (to avoid the undesirable situation of having $\alpha_{k,m}^{w*}(\mathbf{G}) \neq 0 \ \forall m$), it follows that $\alpha_{k,m}^{w*}(\mathbf{G}) = 0$ 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 $\beta_k^{w*}(\mathbf{G})[\sum_{m=1}^M w_{k,m}^*(\mathbf{G}) - 1] = 0$ and $\beta_k^{w*}(\mathbf{G}) \neq 0$, 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 power cost function is minimum, i.e., $m_k = \arg\min_m\{\mathcal{P}_{k,m}(\mathbf{G})\}_{m=1}^M$.*
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 $\alpha_{k,m_k}^{w*}(\mathbf{G}) = 0$. For this user, (8) implies that $\beta_k^{w*}(\mathbf{G}) = -\mathcal{P}_{k,m_k}(\mathbf{G})f_\mathbf{G}(\mathbf{G})$. Now applying (8) to another user $m_k' \neq m_k$ yields $\mathcal{P}_{k,m_k'}(\mathbf{G})f_\mathbf{G}(\mathbf{G}) + \beta_k^{w*}(\mathbf{G}) - \alpha_{k,m_k'}^{w*}(\mathbf{G}) = [\mathcal{P}_{k,m_k'}(\mathbf{G}) - \mathcal{P}_{k,m_k}(\mathbf{G})]f_\mathbf{G}(\mathbf{G}) - \alpha_{k,m_k'}^{w*}(\mathbf{G}) = 0$. Satisfying the latter requires $\mathcal{P}_{k,m_k'}(\mathbf{G}) \geq \mathcal{P}_{k,m_k}(\mathbf{G})$, since $\alpha_{k,m_k'}^{w*}(\mathbf{G}) \geq 0$; that is, $\mathcal{P}_{k,m_k}(\mathbf{G}) = \min_m \ \mathcal{P}_{k,m}(\mathbf{G})$. ∎

There are two rare cases for which Proposition 1 must be carefully asserted because the optimal solution may not necessarily amount to a single user accessing each subcarrier. The first one corresponds to realizations $\mathbf{G}$ for which the minimum value of $\{\mathcal{P}_{k,m}(\mathbf{G})\}_{m=1}^M$ is attained by multiple users. In this case, any arbitrary time sharing of the subcarrier by multiple users can be, in principle, optimum. Under (as1) and (as2) however, this is an event with Lebesgue-measure zero that, if happens, can be solved by simply picking one of the users at random (the power cost will remain invariant and the claim of a single user accessing each subcarrier will still be optimum). The second case corresponds to realizations $\mathbf{G}$ for which $\mathcal{P}_{k,m}(\mathbf{G}) > 0$ $\forall m$. In this case, we have from (8) that $\alpha_{k,m}^{w*}(\mathbf{G}) \neq 0$, $\forall m$ since $\beta_k^{w*}(\mathbf{G}) \geq 0$; and the optimal solution will not allocate the corresponding subcarrier to any user. This can be easily incorporated to Proposition 1 by introducing a fictitious user $\mathcal{P}_{k,0}(\mathbf{G}) := 0 \ \forall k, \mathbf{G}$. Taking into account these clarifications, we are ready to express $w_{k,m}^*(\mathbf{G})$ in compact form using the indicator function as

$$w_{k,m}^*(\mathbf{G}) = \mathbf{I}_{\{m = \arg\min_{m'}\{\mathcal{P}_{k,m'}(\mathbf{G})\}_{m'=0}^M\}}. \quad (10)$$

Interestingly, finding the optimum user-subcarrier allocation dictated by (10) incurs complexity that is linear in the number of subcarriers and users. Note that linear complexity in assigning subcarriers is a direct consequence of the winner-takes-all access asserted by Proposition 1, which in turn follows from the fact that we optimize average (as opposed to instantaneous) performance over fading channels [cf. (3)].

Remark 2: From (4), (5) and (10), we can readily infer that: (i) the only coupling among subcarriers is through the multiplier β_m^r (i.e., given $\beta_m^r \ \forall m$, the allocation of rate and power on each subcarrier can be performed independently); (ii) given β_m^r, 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 (8) jointly $\forall m$.

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 (GRA)*

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

(S1.1) For each subcarrier k:

(S1.1.1) Set arbitrary non-negative $\beta_{k,m|l}^\epsilon \ \forall m, l$.

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

(S1.1.3) For $\alpha_{k,m|l}^p = 0$ and $\alpha_{k,m|l}^r = 0$, use (5) 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 (10).

(S1.1.5) Check (4) $\forall m, l$. If $|LHS(4) - 1/\beta_{k,m|l}^\epsilon| < \delta\beta_{k,m|l}^\epsilon \ \forall m, l$ go to (S1.1.6); otherwise increase $p_{k,m|l}$ if (m, l) is such that $LHS(4) > 1/\beta_{k,m|l}^\epsilon$; decrease $p_{k,m|l}$ if (m, l) is such that $LHS(4) < 1/\beta_{k,m|l}^\epsilon$, and go to (S1.1.3).

(S1.1.6) Check constraint $C2$ in (3) $\forall m, l$. If $|C2| < \delta[\bar{\boldsymbol{\epsilon}}_0]_m \ \forall m, l$, move to the next subcarrier $k+1$, and go to (S1.1); otherwise increase $\beta_{k,m|l}^\epsilon$ if (m, l) is such that $C2 > 0$; decrease $\beta_{k,m|l}^\epsilon$ if (m, l) is such that $C2 < 0$, and go to (S1.1.3).

(S1.2) Check constraint $C1$ in (3) $\forall m$. If $|C1| < \delta[\bar{\mathbf{r}}_0]_m \ \forall m$ then *Stop*; otherwise increase β_m^r if a user index m is such that $C1 > 0$; decrease β_m^r if m is such that $C1 < 0$, and go to (S1.1).

Performance and convergence of the above mentioned algorithm will be clearly affected by the schemes used to increase/decrease $p_{k,m|l}, \beta_{k,m|l}^\epsilon$, and β_m^r in steps (S1.1.5), (S1.1.6), and (S1.2). The reason we did not specify these schemes is twofold: (i) we want to illustrate that besides characterizing the solution, the KKT conditions can in principle solve the optimization problem at hand; and (ii) for practical implementation we advocate a simplified algorithm which we describe in the next section.

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 [5]. On the other hand, optimizing CR loadings is analytically more convenient and leads to more power savings compared with DR. In fact, even for systems that implement DR, the CR solution, offers intuition and useful guidelines for the DR design. 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 [6].

C. Reduced-Complexity Allocation Algorithm

Instead of minimizing $\bar{P}$, in this section we will minimize $\bar{P}^+$, an upper bound on $\bar{P}^*$. Minimizing an upper bound is a commonly used technique when direct optimization of the original objective function is difficult [1]. To this end, let ϵ_P^{-1} denote the inverse function involved when solving (1) w.r.t. $p_{k,m}$, and $g_{k,m|l}^{\min} := \min\{g_{k,m} \mid g_{k,m} \in \mathcal{R}_{k,m|l}\}$ represent the worst channel gain. Then we can upper bound $\bar{P}$ in (2) using $\bar{P}^+ := \sum_{m=1}^M \sum_{k=1}^K \mathbb{E}_\mathbf{G}[p_{k,m|l(\mathbf{G})}^+ w_{k,m}(\mathbf{G})]$, where

$$p_{k,m|l(\mathbf{G})}^+ := \epsilon_P^{-1}(r_{k,m|l(\mathbf{G})}, g_{k,m|l(\mathbf{G})}^{\min}, [\bar{\epsilon}_0]_m). \quad (11)$$

The design parameters using this upper bound approach will be identified with the superscript "+". The upper bound in (11) brings the following advantages: (i) constraint $C2$ is automatically satisfied and thus $\beta_{k,m|l}^{\epsilon+} = 0$, and (ii) a direct link is established between $p_{k,m|l}^+$ and $r_{k,m|l}^+$ through ϵ_P^{-1} and therefore we need only to optimize over either variable, thus reducing the dimensionality.

With power variables eliminated, the new problem can be expressed as:

$$\begin{cases} \min_{\mathbf{R}^+(\mathbf{G}) \geq \mathbf{0}, \mathbf{W}^+(\mathbf{G}) \geq \mathbf{0}} \bar{P}^+, \text{ where } \bar{P}^+ := \\ \sum_{k=1}^K \sum_{m=1}^M \mathbb{E}_\mathbf{G}[\epsilon_P^{-1}(r_{k,m|l(\mathbf{G})}^+, g_{k,m|l(\mathbf{G})}^{\min}, [\bar{\epsilon}_0]_m) w_{k,m}^+(\mathbf{G})] \\ \\ \text{subject to :} \\ \sum_{k=1}^K \mathbb{E}_\mathbf{G}[r_{k,m|l(\mathbf{G})}^+ w_{k,m}^+(\mathbf{G})] \geq [\bar{\mathbf{r}}_0]_m, \ \forall m; \\ \sum_{m=1}^M w_{k,m}^+(\mathbf{G}) \leq 1, \ \forall k. \end{cases}$$
$$(12)$$

Using $\bar{r}_{k,m|l}^+ := \mathbb{E}_\mathbf{G}[r_{k,m|l(\mathbf{G})}^+ w_{k,m}^+(\mathbf{G})]$ and following steps similar to those in [15, eq. (11)], the optimization in (12) can be easily cast as a convex problem.

This worst-case design based on the upper bound, will yield $w_{k,m}^{+*}$ and $r_{k,m|l}^{+*}$ as a solution of (12). Specifically, upon defining $\kappa_{3,m} := \kappa_2^{-1} \ln(\kappa_1/[\bar{\epsilon}_0]_m)$ we can solve (12) to find

$$r_{k,m|l(\mathbf{G})}^{+*} = \log_2\left(\frac{\beta_m^{r+*} g_{k,m|l(\mathbf{G})}^{\min}}{\ln(2)\kappa_{3,m}}\right), \quad (13)$$

$$\mathcal{P}_{k,0}^+(\mathbf{G}) = 0$$

$$\mathcal{P}_{k,m}^+(\mathbf{G}) = \frac{(2^{r_{k,m|l(\mathbf{G})}^{+*}} - 1)\kappa_{3,m}}{g_{k,m|l(\mathbf{G})}^{\min}} - \beta_m^{r+*} r_{k,m|l(\mathbf{G})}^{+*}, (14)$$

$$w_{k,m}^{+*}(\mathbf{G}) = \mathbf{I}_{\{m=\arg\min_{m'}\{\mathcal{P}_{k,m'}^+(\mathbf{G})\}_{m'=0}^M\}}, \quad (15)$$

where the subcarrier assignments given by (14) and (15) follows the same principle (described by Proposition 1) that the optimum assignment does for the original problem in (7) and (10). Furthermore, it is worth noting that due to the

logarithmic expression of ϵ_P^{-1} [cf. (1)], the rate loading is reminiscent of the classical capacity water-filling solution.

However, to obtain the final power loading, instead of substituting $\mathbf{R}^{+*}$ and $\mathbf{W}^{+*}$ into (11) to find the upper bound $\mathbf{P}^{+*}$, we will use $\mathbf{R}^{+*}$ and $\mathbf{W}^{+*}$ to tightly satisfy the BER constraint $C2$ in (3). Let $\mathbf{P}^{++*}$ denote the exact power loading found after substituting $\mathbf{R}^{+*}$ and $\mathbf{W}^{+*}$ of (13) and (15) into $C2$ in (3). The generic entry $p_{k,m|l}^{++*}$ is then found as the solution of

$$\int_{\mathbf{G} \in \mathcal{R}_{k,m|l}} w_{k,m}^{+*}(\mathbf{G})\kappa_1 \exp\left(-\frac{p_{k,m|l(\mathbf{G})}^{++*}\kappa_2 g_{k,m}}{2^{r_{k,m|l(\mathbf{G})}^{+*}} - 1}\right) f_\mathbf{G}(\mathbf{G})d\mathbf{G}$$
$$= [\bar{\epsilon}_0]_m \bar{w}_{k,m|l}^+. \quad (16)$$

Although $p_{k,m|l}^{++*}$ in (16) can not be found in closed form, the monotonicity of the exponential function allows it to be obtained through line search (using e.g., the bisection method). The accuracy of this section's low-complexity allocation will be evaluated through simulations in Section V. The following algorithm summarizes the main steps for solving the set of equations (13), (15), and (16).

Algorithm 2: *Practical Resource Allocation (PRA)*

(S2.0) Select a small positive number δ, and initialize $\boldsymbol{\beta}^{r+}$ using an arbitrary non-negative vector.

(S2.1) For each (k, m) pair per iteration:

(S2.1.1) Use β_m^{r+} to determine $r_{k,m|l}^+ \ \forall l$ via (13).

(S2.1.2) Use β_m^{r+} to determine $w_{k,m}^+(\mathbf{G}) \ \forall l$ via (15).

(S2.2) Check the rate constraint in (12) $\forall m$; if $\left|\sum_{k=1}^K \mathbb{E}_\mathbf{G}[r_{k,m|l(\mathbf{G})}^+ w_{k,m}^+(\mathbf{G})] - [\bar{\mathbf{r}}_0]_m\right| < \delta[\bar{\mathbf{r}}_0]_m$ $\forall m$, then go to (S2.3); otherwise, increase β_m^{r+} for the users m whose average rate is smaller than $[\bar{\mathbf{r}}_0]_m$; decrease β_m^{r+} for the users m whose average rate is greater than $[\bar{\mathbf{r}}_0]_m$; and go to (S2.1).

(S2.3) Once $\mathbf{R}^{+*}, \boldsymbol{\beta}^{r+*}, \mathbf{W}^{+*}$ are obtained, use (16) to calculate the finally allocated power.

The convex relaxation of this section not only reduces the complexity but also enables efficient methods to update $\boldsymbol{\beta}^{r+}$ [1]. For example, we can set the initial value of $\boldsymbol{\beta}^{r+}$ $\beta_{m'}^{r+}$ equal to any small number and update each component β_m^{r+} independently $\forall m$ by fixing $\beta_{m'}^{r+}, \ \forall m' \neq m$ from the previous iteration. The adaptation of each β_m^{r+} is then performed using line search until the rate constraint for the mth user is tightly satisfied. This simple algorithm has guaranteed convergence and facilitates computation *distributed* across users.

D. 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* $(\mathbf{P}^{+*}, \mathbf{R}^{+*}, \mathbf{W}^{+*}(\boldsymbol{\beta}^{r+*}))$ *returned by Algorithm 2, the AP broadcasts to the users the codeword* $\mathbf{c}^*(\mathbf{G}) = [\mathbf{c}_1^*(\mathbf{G}), \dots, \mathbf{c}_K^*(\mathbf{G})]$ *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) $m_k^*(\mathbf{G}) = \arg\min_m \{\mathcal{P}_{k,m}^+(\mathbf{G}, \mathbf{P}^{+*}, \mathbf{R}^{+*}, \boldsymbol{\beta}^{r+*})\}_{m=1}^M$ *(pick randomly any user* m_k^* *when multiple minima occur); and*

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

Note that Proposition 2 applies also to Algorithm 1 after dropping the superscript "+" from the corresponding design parameters.

Remark 4: 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 fictitious user (e.g., $m = 0$) with a unique region. Including all these indices, the codeword length will be $\left\lceil \sum_{k=1}^K \log_2 \left(\sum_{m=1}^M (L_{k,m} - 1) + 1 \right) \right\rceil$ bits.

We conclude this section by emphasizing that $\mathbf{P}$ and $\mathbf{R}$ in (2) and (12) are involved only in average quantities. Hence, all possible values for the entries of $\mathbf{P}^{+*}$ and $\mathbf{R}^{+*}$ ($L_{k,m}$ for the (k, m)th entry) are computed off-line and only the subcarrier-user assignment (involved in instantaneous constraints) and the indexing of the corresponding regions need to be fed back on-line.

IV. QUANTIZER DESIGN

In the previous section we addressed our objective (G2) to derive optimum and suboptimum subcarrier, rate, and power allocation policies assuming the quantization regions, $\mathcal{R}_{k,m|l}$, are given. In this section, we will address (G1) by deriving an optimum iterative scheme for the selection of $\mathcal{R}_{k,m|l}^*$ as well as a suboptimum non-iterative alternative which greatly simplifies the quantizer design.

A. Optimum Quantization

Besides $\mathbf{P}^*$, $\mathbf{R}^*$, and $\mathbf{W}^*$, in this section $\{\mathcal{R}_{k,m|l}^*\}_{l=1}^{L_{k,m}}$ will be considered as variables $\forall(k, m)$ in 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) $\mathcal{R}_{k,m|l}$ represents a quantization region of a single variable $g_{k,m}$. In fact, $\{\mathcal{R}_{k,m|l}^*\}_{l=1}^{L_{k,m}}$ can be equivalently represented by a set of thresholds $\{\tau_{k,m|l}^*\}_{l=1}^{L_{k,m}+1}$, 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.

To determine $\{\tau_{k,m|l}^*\}_{l=2}^{L_{k,m}} \ \forall k, m$ we start with the Lagrangian $\mathcal{L}$ of (2). Expressing the latter in terms of the thresholds and defining $\mathcal{G}_{k,m}\{x, y\} := \{\mathbf{G} : g_{k,m} \in [x, y)\}$, we can write

$$\mathcal{L} := \sum_{k=1}^K \sum_{m=1}^M \sum_{l=1}^{L_{k,m}} \int_{\mathbf{G} \in \mathcal{G}_{k,m}\{\tau_{k,m|l}, \tau_{k,m|l+1}\}} \Bigg[p_{k,m|l}(\mathbf{G})$$
$$-\beta_m^r r_{k,m|l}(\mathbf{G}) + \beta_{k,m|l}^{\epsilon*} \kappa_1 \exp\left(-\frac{p_{k,m|l}(\mathbf{G})^{\kappa_2} g_{k,m}}{2^{r_{k,m|l}(\mathbf{G})} - 1} \right)$$
$$-\beta_{k,m|l}^\epsilon [\bar{\epsilon}_0]_m \Bigg] w_{k,m}(\mathbf{G}) f_{\mathbf{G}}(\mathbf{G}) d\mathbf{G}, \quad (17)$$

where for clarity we omitted the contribution of constraints that do not depend on $\{\tau_{k,m|l}\}_{l=1}^{L_{k,m}+1}$.

The necessary KKT condition found after setting $\partial \mathcal{L} / \partial \tau_{k,m|l}^* = 0$, yields

$$p_{k,m|l-1} - \beta_m^r r_{k,m|l-1} + \beta_{k,m|l-1}^\epsilon \kappa_1 \exp\left(-\frac{p_{k,m|l-1}^{\kappa_2} \tau_{k,m|l}^*}{2^{r_{k,m|l-1}} - 1} \right)$$
$$= -p_{k,m|l} + \beta_m^r r_{k,m|l} - \beta_{k,m|l}^\epsilon \kappa_1 \exp\left(-\frac{p_{k,m|l}^{\kappa_2} \tau_{k,m|l}^*}{2^{r_{k,m|l}} - 1} \right). \quad (18)$$

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

As (18) links $\{\tau_{k,m|l}^*\}_{l=2}^{L_{k,m}}$, $\{p_{k,m|l}^*\}_{l=1}^{L_{k,m}}$ and $\{r_{k,m|l}^*\}_{l=1}^{L_{k,m}}$, per inner iteration inside (S.1.1) of Algorithm 1, we have to solve not only for $\{p_{k,m|l}^*\}_{l=1}^{L_{k,m}}$, $\{r_{k,m|l}^*\}_{l=1}^{L_{k,m}}$, and $w_{k,m}^*(\mathbf{G}) \ \forall m$ (as before), but also for $\{\tau_{k,m|l}^*\}_{l=2}^{L_{k,m}} \ \forall m$ via (18). This requires modifying the steps (S1.1.1) and (S1.1.6) of Algorithm 1, and incorporating an extra step (S1.1.7) as follows:

Algorithm 1': *GRA with Iterative Channel Quantization*

(S1.0) Same as in Algorithm 1.

(S1.1) Execute this step as in Algorithm 1:

(S1.1.1) Besides $\beta_{k,m|l}^\epsilon \ \forall m, l$, initialize also $\{\tau_{k,m|l}\}_{l=2}^{L_{k,m}} \ \forall m$.

(S1.1.2)-(S1.1.5) Same as in Algorithm 1.

(S1.1.6) After checking $C2$, if $|C2| < \delta [\bar{\epsilon}_0]_m \ \forall m, l$, instead of updating k and going to step (S1.1), keep the current k and go to the new step (S1.1.7).

(S1.1.7) Given $\{p_{k,m|l}, r_{k,m|l}, \beta_{k,m|l}^\epsilon\}_{l=1}^{L_{k,m}}$, $w_{k,m}(\mathbf{G})$ and $\beta_m^r \ \forall m$ from (S1.1.6), update the values of $\{\tau_{k,m|l}\}_{l=2}^{L_{k,m}} \ \forall m$ as in (18). 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) Same as in Algorithm 1.

With this augmentation of Algorithm 1, it is possible to iteratively solve (per $\boldsymbol{\beta}^r$ value and subcarrier k) for the variables $\{p_{k,m|l}^*\}_{l=1}^{L_{k,m}}$, $\{r_{k,m|l}^*\}_{l=1}^{L_{k,m}}$ and $w_{k,m}^*(\mathbf{G})$, and $\{\tau_{k,m|l}^*\}_{l=2}^{L_{k,m}} \ \forall m$.

The optimality condition in (18) pertains to the optimization problem in (2). Analogously, for the practical resource allocation of Section III-C, the optimization in (12) can be

carried out jointly over the thresholds. The corresponding KKT condition for optimality is

$$\frac{2^{r^+_{k,m|l}} - 1}{\tau^*_{k,m|l-1}} \kappa_{3,m} - \beta^{r+}_m r^+_{k,m|l} = \frac{2^{r^+_{k,m|l+1}} - 1}{\tau^*_{k,m|l}} \kappa_{3,m} \quad (19)$$

$$-\beta^{r+}_m r^+_{k,m|l+1} + \frac{2^{r^+_{k,m|l+1}} - 1}{(\tau^*_{k,m|l})^2} \kappa_{3,m} A_{k,m|l},$$

where $A_{k,m|l} := \dfrac{\int_{\mathbf{G} \in \mathcal{G}_{k,m}\{\tau^*_{k,m|l}, \tau^*_{k,m|l+1}\}} w_{k,m}(\mathbf{G}) f_{\mathbf{G}}(\mathbf{G}) d\mathbf{G}}{\int_{\mathbf{G} \in \mathcal{G}_{k,m}\{\tau^*_{k,m|l}\}} w_{k,m}(\mathbf{G}) f_{\mathbf{G}}(\mathbf{G}) d\mathbf{G}}$, and $\mathcal{G}_{k,m}\{x\} := \{\mathbf{G} : g_{k,m} = x\}$.

Notice that the main difference between (18) and (19) is that in (12) we have $g^{\min}_{k,m|l} = \tau_{k,m|l}$. This implies that $p^+_{k,m|l+1}$ is an explicit function of $\tau_{k,m|l}$ and consequently contributes to the optimality condition via $A_{k,m|l}$.

In Algorithm 2, (S.2.0) has to be modified and the extra step (S2.1.3) has to be included:

Algorithm 2': *PRA with Iterative Channel Quantization*

(S2.0) Besides δ and $\boldsymbol{\beta^r}$, initialize also $\{\tau_{k,m|l}\}^{L_{k,m}}_{l=2} \; \forall k, m$.
(S2.1) Execute this step as in Algorithm 2:

(S2.1.1)-(S2.1.2) Same as in Algorithm 2.
(S2.1.3) Use $\{r^+_{k,m|l}\}^{L_{k,m}}_{l=1}$, $w^+_{k,m}(\mathbf{G})$ and $\beta^{r+}_m \; \forall m$, to update the values of $\{\tau_{k,m|l}\}^{L_{k,m}}_{l=2} \; \forall m$ as in (19). If the change w.r.t. the previous values of $\{\tau_{k,m|l}\}^{L_{k,m}}_{l=2}$ is smaller than $\delta \; \forall m, l$, then go to (S2.1); otherwise go back to (S2.1.1).

(S2.2) Same as in Algorithm 2.

Although both (18) and (19) return quantizers that attain a local optimum, global optimality is not guaranteed due to lack of convexity. On the other hand, it is clear that the modified versions of Algorithms 1 and 2 are more complex. For this reason, we will introduce in the next subsection a non-iterative quantizer design that yields a simple yet effective channel quantizer. Before proceeding, we close this section with a remark that links the proposed solution with classical quantization theory.

Remark 5: Since we are dealing with a constrained minimization problem, the Lagrangian in (17) can be interpreted as the overall average distortion metric associated with the quantizer design [4]. In this context, we can find the quantization thresholds and codewords minimizing the average cost $\mathcal{L}$ through the Lloyd algorithm (LA) [9]. LA minimizes an average distortion metric iterating between two generic steps: (L1) given the regions, find the optimal codewords associated with them *(centroid condition)*; and (L2) given the updated codewords, find the corresponding optimal regions *(nearest neighbor rule)*. Notice that from an optimization point of view, (L1) is equivalent to setting $\partial \mathcal{L}/\partial p_{k,m|l}$, $\partial \mathcal{L}/\partial r_{k,m|l}$, and $\partial \mathcal{L}/\partial w_{k,m}$ equal to zero ($p_{k,m|l}, r_{k,m|l}, w_{k,m}$ represent the centroids), while (L2) is equivalent to $\partial \mathcal{L}/\partial \tau_{k,m|l} = 0$ ($\tau_{k,m|l}$ characterize the regions). To run the LA we only need to define the distortion metric, $d^{\mathcal{L}}$, as a function of $\mathbf{G}$. If $\mathbf{c}$ is a codeword adhering to Proposition 2 (i.e., $\mathbf{c}$ indexes

$p_{k,m|l}, r_{k,m|l}, w_{k,m}, \forall k, m$), we can eventually express $d^{\mathcal{L}}$ as

$$d^{\mathcal{L}}(\mathbf{G}, \mathbf{c}(\mathbf{G})) := \sum_{k=1}^{K} \sum_{m=1}^{M} \left[p_{k,m|l(\mathbf{G})} - \beta^{r*}_m r_{k,m|l(\mathbf{G})} \right.$$
$$\left. + \beta^{\epsilon*}_{k,m|l} \kappa_1 \exp\left(-\frac{p_{k,m|l(\mathbf{G})} \kappa_2 g_{k,m}}{2^{r_{k,m|l(\mathbf{G})}} - 1} \right) \right] w_{k,m}(\mathbf{G}). \quad (20)$$

If we compare (20) with (18), we can interpret $\tau^*_{k,m|l}$ in (18) either as the solution of the KKT condition applied to (2), or, as the point whose distance $d^{\mathcal{L}}$ from the centroid $(p_{k,m|l}, r_{k,m|l}, w_{k,m})$ is equal to the distance from its neighboring centroid $(p_{k,m|l-1}, r_{k,m|l-1}, w_{k,m})$.

B. Equally Probable Region Quantizer

The metric (cost function) for calculating $\tau_{k,m|l}$ in the previous section relied on minimizing $\bar{P}$. In this section, we first solve the optimal resource allocation problem supposing CSI is available without quantization (i.e., $L_{k,m} \to \infty$), and subsequently calculate $\tau_{k,m|l}$ to satisfy

$$\int_{\tau_{k,m|l}}^{\tau_{k,m|l+1}} \Pr(w_{k,m} = 1, g_{k,m}) dg_{k,m}$$
$$= \int_{\tau_{k,m|1}}^{\tau_{k,m|L_{k,m}}} \Pr(w_{k,m} = 1, g_{k,m}) dg_{k,m}/L_{k,m}, \quad (21)$$

with $\tau_{k,m|1} = 0$ and $\tau_{k,m|L_{k,m}+1} = \infty$. If the joint probabilities can be computed, solving (21) yields thresholds $\{\tau_{k,m|l}\}^{L_{k,m}}_{l=1}$ per subcarrier k and user m that divide the joint probability $\Pr(w_{k,m} = 1, g_{k,m})$ into regions of equal area; hence the term *equally probable region quantizer*. Intuitively speaking, this quantizer design tries to maximize the entropy in the feedback link.

To evaluate $\Pr(w_{k,m} = 1, g_{k,m})$ needed in (21), we apply Bayes' rule to re-write it as $\Pr(w_{k,m} = 1|g_{k,m}) f_{g_{k,m}}(g_{k,m})$ and recall that $f_{g_{k,i}}(g_{k,m})$ is known per (as2). To calculate $\Pr(w_{k,m} = 1|g_{k,m})$, we will need to first solve the optimal resource allocation problem assuming no quantization. Clearly, as $L_{k,m} \to \infty$, we have $g_{\min} \to g$ in (14). Letting $\mathcal{P}^{\infty}_{k,m}$ and $\beta^{r\infty*}_m$ denote, respectively, the cost indicator and the Lagrange multiplier when $L_{k,m} \to \infty$, we can write [cf. (14)]

$$\mathcal{P}^{\infty}_{k,m}(g_{k,m}) = \beta^{r\infty*}_m / \ln(2) - \kappa_{3,m}/g_{k,m}$$
$$- \beta^{r\infty*}_m \log_2 \left(g_{k,m} \beta^{r\infty*}_m / (\kappa_{3,m} \ln(2)) \right). (22)$$

Because condition in (15) establishes that $w_{k,m}(\mathbf{G}) = \mathbf{I}_{\{m = \arg\min_{m'}\{\mathcal{P}^{\infty}_{k,m'}(\mathbf{G})\}^{M}_{m'=0}\}}$, if $\mathcal{P}^{\infty}_{k,m}(g_{k,m}) < \mathcal{P}^{\infty}_{k,0} = 0$, then $g_{k,m} > \ln(2)\kappa_{3,m}/\beta^{r\infty*}_m$ is a necessary condition for the user m to be active. Taking also into account that channels of different users are uncorrelated [cf. (as1)], we can write

$$\Pr(w_{k,m} = 1|g_{k,m}) =$$
$$\mathbf{I}_{\{g_{k,m} > \frac{\ln(2)\kappa_{3,m}}{\lambda^{r*}_m}\}} \prod_{\mu=1, \mu \neq m}^{M} \Pr(\mathcal{P}^{\infty}_{k,m} < \mathcal{P}^{\infty}_{k,\mu} | g_{k,m}). \quad (23)$$

Interestingly, for the active users we have $\mathcal{P}^{\infty}_{k,m}(g_{k,m}) = \mathbf{I}_{\{g_{k,m} > \frac{\ln(2)\kappa_{3,m}}{\beta^{r\infty*}_m}\}} \left(\frac{\beta^{r\infty*}_m}{\ln(2)} - \frac{\kappa_{3,m}}{g_{k,m}} - \beta^{r\infty*}_m \log_2 \left(\frac{g_{k,m}\beta^{r\infty*}_m}{(\ln(2)\kappa_{3,m})} \right) \right)$, and therefore $\frac{\partial \mathcal{P}^{\infty}_{k,m}(g_{k,m})}{\partial g_{k,m}} = \mathbf{I}_{\{g_{k,m} > \frac{\ln(2)\kappa_{3,m}}{\beta^{r\infty*}_m}\}} \left(\frac{\kappa_{3,m}}{g_{k,m}} - \frac{\beta^{r\infty*}_m}{\ln(2)} \right) \frac{\kappa_{3,m}}{g_{k,m}} \leq 0$, which implies

that $\mathcal{P}_{k,m}^{\infty}(g_{k,m})$ is monotonically decreasing. Therefore, we can find unique channel gains $\gamma_{k,\mu}$, $\forall \mu \neq m$, such that

$$\mathcal{P}_{k,m}^{\infty}(g_{k,m}) = \mathcal{P}_{k,\mu}^{\infty}(\gamma_{k,\mu}). \tag{24}$$

It is then clear that $\mathcal{P}_{k,m}^{\infty}(g_{k,m}) \leq \mathcal{P}_{k,\mu}^{\infty}(g_{k,\mu})$ if $g_{k,\mu} \in [0, \gamma_{k,\mu}]$, and $\mathcal{P}_{k,m}^{\infty}(g_{k,m}) > \mathcal{P}_{k,\mu}^{\infty}(g_{k,\mu})$ if $g_{k,\mu} \in (\gamma_{k,\mu}, \infty)$. And consequently, $\mathrm{Pr}(\mathcal{P}_{k,m}^{\infty} < \mathcal{P}_{k,\mu}^{\infty} \mid g_{k,m}) = \mathrm{Pr}(g_{k,\mu} < \gamma_{k,\mu} \mid g_{k,m})$.

Excluding the case $\mathcal{P}_{k,m}^{\infty}(g_{k,m}) > 0$, which amounts to $w_{k,m} = 0$ as discussed before and solving (24) w.r.t. $\gamma_{k,\mu}$ yields

$$\gamma_{k,\mu}(g_{k,m}) = -\frac{\kappa_{3,\mu} \ln(2)/\beta_{\mu}^{r\infty*}}{f^W \left[-2^{-\frac{\kappa_{3,m}}{g_{k,m}\beta_{\mu}^{r\infty*}}} e^{-1} \left(\frac{e \ln(2)\kappa_{3,m}}{g_{k,m}\beta_{m}^{r\infty*}} \right)^{\frac{\beta_{m}^{r\infty*}}{\beta_{\mu}^{r\infty*}}} \right]}, \tag{25}$$

where $f^W[x] = y$ is the real-valued Lambert's f^W function which solves the equation $y \exp(y) = x$ for $-1 \leq y \leq 0$ and $-1/e \leq x \leq 0$ [3].

Using (23)-(25), we can express $\mathrm{Pr}(w_{k,m} = 1, g_{k,m})$ in (21) as

$$\mathrm{Pr}(w_{k,m} = 1, g_{k,m}) = \mathbf{I}_{\{g_{k,m} > \frac{\ln(2)\kappa_{3,m}}{\beta_m^{r\infty*}}\}} \frac{\exp(-g_{k,m}/\bar{g}_{k,m})}{\bar{g}_{k,m}}$$
$$\times \prod_{\mu=1, \mu \neq m}^{M} \left(1 - \exp(-\gamma_{k,\mu}(g_{k,m})/\bar{g}_{k,\mu}) \right). \tag{26}$$

Since (26) depends on $\{\beta_{\mu}^{r\infty*}\}_{\mu=1}^{M}$, we need to solve the optimal allocation problem as $L_{k,m} \to \infty$. The thresholds $\{\tau_{k,m|l}\}_{l=2}^{L_{k,m}}$ are obtained by solving (21) using a line search. Notice that we can also take advantage of the condition $\mathcal{P}_{k,m}^{\infty}(g_{k,m}) < 0$, by setting $\tau_{k,m|2} \geq \kappa_{3,m} \ln(2)/\beta_m^{r\infty*}$.

An example illustrating the quantizer of this section is given in Fig. 2 which depicts $\mathrm{Pr}(w_{k,m} = 1|g_{k,m})$, $f_{g_{k,m}}(g_{k,m})$ and $\mathrm{Pr}(w_{k,m} = 1, g_{k,m})$ versus $g_{k,m}/\bar{g}_{k,m}$ for $M = 6$, $L_{k,m} = 6$, equal average subcarrier gains, and equal rate constraints. The first subplot in this figure, $\mathrm{Pr}(w_{k,m} = 1|g_{k,m})$, reveals that the better the channel the more likely the corresponding user is to be selected. Coupling this observation with the exponential behavior of $f_{g_{k,m}}(g_{k,m})$ in the second subplot, the bell-ring characteristic of the joint PDF, $\mathrm{Pr}(w_{k,m} = 1, g_{k,m})$, results naturally in the third subplot (after pair-wise multiplication of the functions in the first two subplots), where the quantization thresholds (and regions) resulting from (21) are also identified.

To conclude this section, it is instructive to summarize complexity aspects when carrying out the resource allocation, (G2), as well as the quantizer design, (G1). Algorithm 1 in Section IV-A solves iteratively the joint (G2)/(G1) problem. With the method of Section IV-B, calculation of $\tau_{k,m|l}$ based on (21) has to be executed only once (to solve (G1)). With the thresholds available, the resource allocation can be easily obtained through the reduced-complexity Algorithm 2 in Section III-C (to solve (G2)). Numerical results will show that this reduced complexity non-iterative design exhibits power consumption similar to the one exhibited by the optimal alternative which solves iteratively the joint resource allocation and quantization problem.
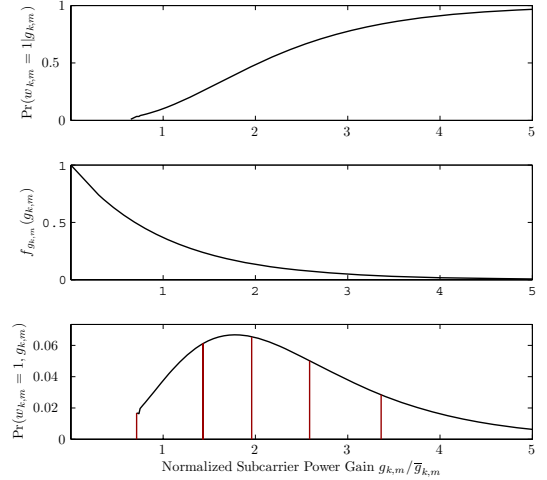

Fig. 2. Quantization based on equally probable regions ($K = 64$, $M = 6$, $L = 6$).

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 (per symbol), $L_{k,m} = 5$ regions (i.e., 4 active regions) per subcarrier, $[\bar{\epsilon}_0]_m = \bar{\epsilon}_0 = 10^{-3}$ $\forall m$, and two different rate requirements: $\bar{\mathbf{r}}_0 = [60, \ 60, \ 60]^T$ and $\bar{\mathbf{r}}_0 = [20, \ 40, \ 60]^T$ bits (per symbol). The average signal-to-noise ratio for the different users is set to 0 dB, while the power profile considered for the multi-path channel corresponds to the three main taps[5] as the test channel *Vehicular A* recommended by the ITU in [18, Table 5].

Test Case 1 *(Comparison of allocation schemes):* For different $\overline{\mathrm{SNR}}$ values and equal average rate constraints across users, Fig. 3 compares the *total* average transmit power for four different allocation schemes based on: (i) the benchmark P-CSIT, (ii) Algorithm 1' (we will denote this alternative as Q-CSIT1), (iii) Algorithm 2 using $\bar{P}^{++}$ and quantization regions given by (21) (we will denote this alternative as Q-CSIT2), and (iv) Algorithm 2 using $\bar{P}^{+}$ and quantization regions given by (21) (we will denote this alternative as Q-CSIT3). The striking observation here is the almost equivalent performance of Q-CSIT1, Q-CSIT2 and P-CSIT schemes and their small (less than 2 dB) gain compared to the Q-CSIT3 one. The negligible difference between Q-CSIT1 and Q-CSIT2 is justified since the power difference between Q-CSIT2 and Q-CSIT3 is less than 2 dB, which means that the upper bound in (11) is tight in terms of power consumption and turns out to yield performance similar to the optimum in terms of resources allocated. Furthermore, these results validate the usefulness of the simple approach to quantization based on equiprobable regions in the multi-user scenario ([10] already hinted that this simple quantizer is near-optimal for single user OFDM).

---

[5]The reason for including only three taps (we picked those with relative loss less than 10dB) is twofold: (i) the simulated performance does not degrade noticeably with this reduced set of taps; and (ii) by fixing the number of taps to three the reference test case provides the same order of multi-user and multi-path (a.k.a. frequency) diversity.
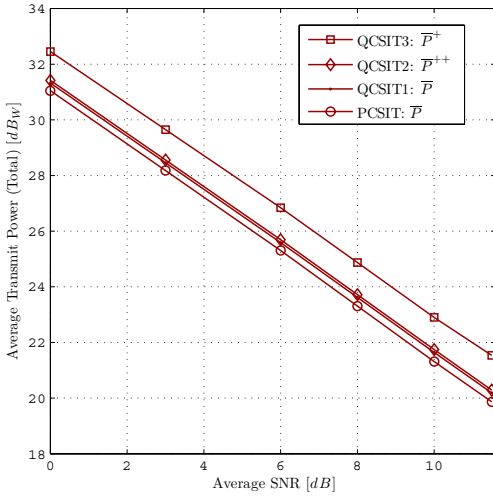
Fig. 3.   Comparison of average (total) transmit power for different proposed loading schemes ($K = 64$, $M = 3$, $\overline{\mathbf{r}}_0 = [60, \ 60, \ 60]^T$ bits).
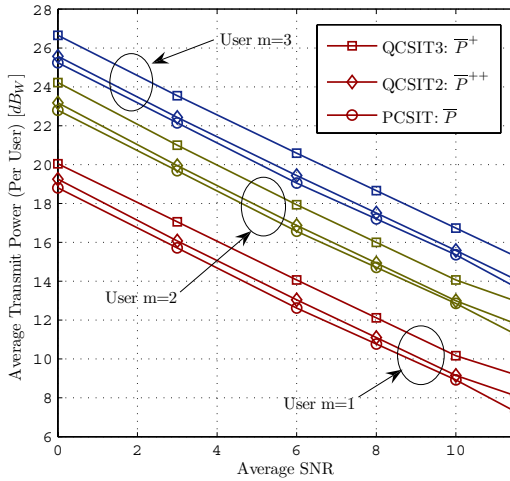

Fig. 4.   Comparison of average (per user) transmit power for different proposed loading schemes ($K = 64$, $M = 3$, $\overline{\mathbf{r}}_0 = [20, \ 40, \ 60]^T$ bits).

With regard to complexity, Q-CSIT2 and Q-CSIT3 schemes attain the best complexity-performance tradeoffs. Specifically, Q-CSIT2 yields performance similar to that of the P-CSIT scheme, even though it only involves a one dimensional search as per (16). On the other hand, Q-CSIT3 yields closed-form expressions but only suffers less than 2 dB loss compared to the P-CSIT scheme. Due to the simplicity and near-optimum power consumption of Q-CSIT2, we will use it in the ensuing examples as a possible candidate scheme. Given different average rate constraints across users, Fig. 4 depicts the average transmit-power per user and corroborates the conclusions illustrated also by Fig. 3.

Test Case 2 *(Comparing Q-CSIT2 to alternative Q-CSIT based schemes):* We here compare the performance of Q-CSIT2 to that of other Q-CSIT based alternatives including: (i) optimum resource allocation with arbitrarily chosen quantization thresholds (call it Q-CSIT4), (ii) optimum rate and
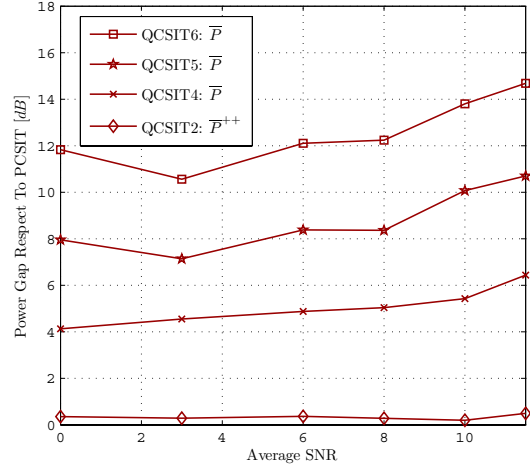

Fig. 5.   Average (total) transmit power gap of different proposed schemes relative to P-CSIT based loading ($K = 64$, $M = 3$, $\overline{\mathbf{r}}_0 = [20, \ 40, \ 60]^T$ bits).

TABLE I
TOTAL AVERAGE TRANSMIT POWER (IN dB_W) FOR P-CSIT, Q-CSIT2 AND Q-CSIT5 SCHEMES (REFERENCE CASE: $K = 64$, $M = 3$, $\overline{\mathbf{r}}_0 = [60, \ 60, \ 60]^T$ BITS, $\overline{\text{SNR}} = 0 \ dB$; OTHER CASES IMPLY ONLY ONE VARIATION W.R.T. THE REFERENCE CASE.)

Case	Q-CSIT5: $\overline{P}$	Q-CSIT2:$\overline{P}^{++}$	P-CSIT:$\overline{P}$
Reference Case	38.9	31.6	31.2
$\overline{\epsilon}_0 = 10^{-4}$	40.8	33.3	32.8
$\overline{\mathbf{r}}_0 = [30, \ 30, \ 30]^T$	31.4	26.7	26.3
$K = 128$	34.8	30.1	29.2
$M = 6$	46.8	39.9	39.2

power allocation with uniform and non-adaptive subcarrier allocation (call it Q-CSIT5), and (iii) Q-CSIT5 with only one bit of feedback (call it Q-CSIT6). Note that since the last scheme considers just two regions per sub-carrier (i.e., just two transmit-configurations per sub-carrier), the optimally designed allocation for this case comprises one inactive region for poor channel realizations plus a unique pair of optimum rate and power loadings for all other channel realizations. Fig. 5 depicts the relative power losses of these schemes compared to the benchmark P-CSIT based one. As shown, while Q-CSIT2 attains performance similar to P-CSIT as asserted before, Q-CSIT4-6 suffer from a power loss of at least 4, 7 and 11 dB, respectively.

Test Case 3 *(Different parameter values):* Numerical results assessing the performance of P-CSIT, Q-CSIT2 and Q-CSIT5 schemes over a wide range of parameter values are summarized in Table I. We observe that these results confirm our previous conclusions, namely: (i) the near optimality of Q-CSIT2, and (ii) the performance loss exhibited by the heuristic schemes exemplified by Q-CSIT5.

Test Case 4 *(Number of users):* To gauge the performance of our allocation schemes when the number of users varies, Fig. 6 depicts the average transmit-power for different values of M with $[\overline{\mathbf{r}}_0]_m = 60 \ \forall m$ remaining fixed. From Fig. 6 we confirm that the gap between P-CSIT and Q-CSIT2 solutions remains small for all configurations tested. More interestingly, we see
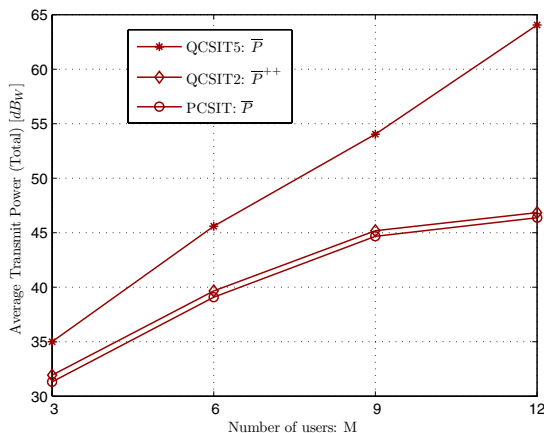
Fig. 6.   Comparison of average (total) transmit power for different proposed loading schemes ($K = 64$, $[\bar{\mathbf{r}}_0]_m = 60$ bits $\forall m$).
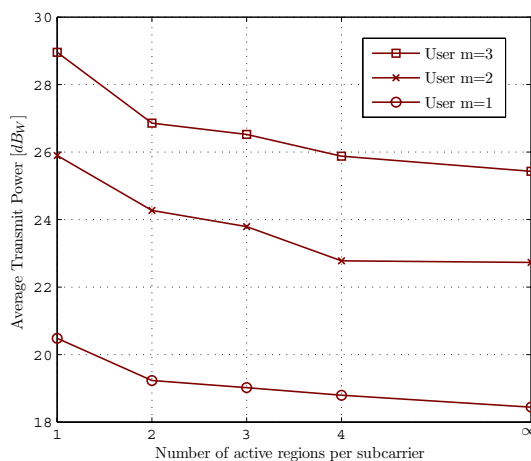

Fig. 7.   Effect of the number of quantization regions (feedback bits) per subcarrier using Algorithm 2 ($K = 64$, $M = 3$, $\bar{\mathbf{r}}_0 = [20, \ 40, \ 60]^T$ bits).

how the gap among P-CSIT and Q-CSIT2 (that dynamically allocate sub-carriers to users) and Q-CSIT5 (that implements fixed subcarrier allocation) increases with the number of users. (Note that Q-CSIT5 is unable to exploit the multi-user diversity provided by the channel.)

Test Case 5 *(Number of quantization regions):* Finally, Fig. 7 plots the average transmit-power versus the number of active regions per subcarrier for three users with different average rate requirements. 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 the joint optimization of resource allocation and quantizing thresholds lead 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 con-

straints. 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 policy. 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.

Besides optimal solutions, to increase the potential for practical deployment, we focused on suboptimal but reduced-complexity designs. The most complex version requires joint optimization of power, rate and subcarriers across users as well as quantization regions. Our approach to lowering complexity was to decouple resource allocation from the quantizer design. Given a quantizer design, the resource allocation problem was further simplified by considering an upper bound to the allocated power. While this upper bound facilitated the optimization task, exact power values were used in the final stage. 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. In regard to the off-line quantizer design, we devised a simple scheme in which quantizing thresholds per user and subcarrier are selected to ensure equiprobable regions under the joint probability of the subchannel gain and the user-specific utilization of the corresponding subcarrier.

To build on the presented framework, interesting future directions include 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 [7]. We also plan to explore the possibility of employing multi-antenna links in the OFDMA setup which can be co-located or distributed.[6]

References

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

[2] M. Cho, W. Seo, Y. Kim, and D. Hong, "A joint feedback reduction scheme using delta modulation for dynamic channel allocation in OFDMA systems," *Proc. IEEE International Symposium Pers., Indoor, Mobile Radio Commun.*, Sep. 2005, vol. 4, pp. 2747–2750.

[3] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances Computational Math.*, vol. 5, pp. 329–359, 1996.

[4] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

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

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

[7] J. H. Kwon, D. Rhee, I. M. Byun, Y. Whang, and K. S. Kim, "Adaptive modulation technique with partial CQI for multiuser OFDMA systems," in *Proc. International Conf. Advanced Commun. Technol.*, Feb. 2006, vol. 2, pp. 1283–1286.

[8] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inform. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[9] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. 28, pp. 129–137, Mar. 1982.

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

[11] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

[12] K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite-rate feedback in multiple-antenna systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

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

[14] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE Signal Processing Mag.*, vol. 17, no. 3, pp. 29–48, May 2000.

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

[16] P. Xia, S. Zhou, and G. B. Giannakis, "Multi-antenna adaptive modulation and transmit beamforming with bandwidth-constrained feedback," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 526–536, Mar. 2005.

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

[18] "Guidelines for evaluation of radio transmission technologies for IMT-2000," *Recomendation ITU-R*, M.1225, 1997.

Antonio G. Marques (M'07) received the degree in Telecommunication Engineering and the doctorate degree (B.Sc., M.Sc., and Ph.D. degrees in electrical engineering) with highest honors from the Universidad Carlos III de Madrid, Madrid, Spain, in 2002 and 2007 respectively. In 2003 he joined the Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain, where he currently develops his research and teaching activities as an Assistant Professor. Since 2005 he has also been a Visiting Researcher at the Department of Electrical Engineering, University of Minnesota, MN, USA.

His research interests lie in the areas of communication theory, signal processing, and networking. His current research focuses on channel state information designs, energy-efficient resource allocation, and wireless ad hoc and sensor networks.

Dr. Marques' work brought him several awards in distinctive international conferences including ICASSP 2007.

Georgios B. Giannakis (F'97) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. Since 1999 he has been a professor with the ECE Department at the Univ. of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications.

His general interests span the areas of communications, networking, and statistical signal processing - subjects on which he has published more than 250 journal papers, 450 conference papers, two edited books and two research monographs. His current research focuses on complex-field and space-time coding, multicarrier cooperative wireless communications, cognitive radios, cross-layer designs, mobile ad hoc networks, and wireless sensor networks.

Dr. Giannakis is the (co-) recipient of six paper awards from the IEEE Signal Processing (SP) and Communications Societies including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He has served the IEEE in a number of posts, and is currently a Distinguished Lecturer for the IEEE-SP Society.

Fadel F. Digham (M'06) was born in Egypt. He received the B.Sc. (with honor) and the M.Sc. in Electronics and Communications both from Cairo University, Cairo, Egypt, in 1995 and 1999. He received both the M.E.E and PhD in Electrical Engineering from the University of Minnesota, Minneapolis, MN, USA, in 2002 and 2005, respectively. He worked for the Alcatel-Telecom, Egypt branch from 1996 until 2000. He was a Postdoctoral researcher with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA from 2005 until 2006. He is currently the director of the R&D department at the National Telecom Regulatory Authority (NTRA) of Egypt and a part-time lecturer at the American University in Cairo (AUC).

Dr. Digham's current research interests are mainly tailored to wireless communications including cognitive radios, WiMAX standards, next generation networks, wireless sensor networks, cross-layer design, distributed signal processing, and detection and estimation. Other interests encompass the management of technology, market analysis, and regulatory-related aspects.

F. Javier Ramos (M'95) received the degree in Telecommunication Engineering and the doctorate degree (B.Sc., M.Sc., and Ph.D. degrees in electrical engineering) from the Universidad Politecnica de Madrid, Spain, in 1990 and 1995 respectively. Between 1992 and 1995 he cooperated in several research projects at Purdue University, Indiana, USA, working in the field of Signal Processing for Communications. During 1996 he was a Post-Doctoral Research Associate at Purdue University. From 1997 to 2003 he was an associate professor at Universidad Carlos III de Madrid. Since 2003 he has been the Dean of the Telecommunications Engineering School at Universidad Rey Juan Carlos in Madrid, Spain.

His research interests lie in the areas of signal processing, mobile communications, and networking.

Dr. Ramos received the COIT-Ericsson Award for the best Ph.D. dissertation on Mobile Communications in 1996.