Energy-Efficient Quantization and Resource Allocation for TDMA with Finite Rate Feedback

Xin Wang, *Member, IEEE,* Antonio G. Marques, *Student Member, IEEE* and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*— We deal with energy efficient time-division multiple access (TDMA) over fading channels with finite-rate feedback for use in the power-limited regime. Through finite-rate feedback from the access point, users acquire quantized channel state information. The goal is to map channel quantization states to adaptive modulation and coding modes and allocate optimally time slots to users so that the total average transmit-power is minimized. To this end, we develop a joint quantization and resource allocation approach, which decouples the complicated problem at hand into three minimization sub-problems and relies on a coordinate descent approach to iteratively effect energy efficiency. A sub-optimal yet simplified alternative algorithm which decouples the original problem into two sub-problems is also presented. Numerical results are presented to evaluate the energy savings and compare the novel approaches.**

***Index Terms*— Convex optimization, quantization, power control and energy efficiency, resource management, multiple access channel processing.**

I. Introduction

With energy efficiency emerging as a critical issue in both commercial and tactical radios designed to extend battery lifetime, energy-efficient resource allocation has attracted growing attention for additive white Gaussian noise (AWGN) channels [1], [2], [3], and time division multi-access (TDMA) fading channels [4], [5]. Resource allocation for fading channels has been studied in [6], [7] and energy-efficiency policies for TDMA have been investigated from an information theoretic perspective in [8]. Assuming that both transmitters and receivers have available perfect (P-) channel state information (CSI), the approaches in [8] not only provide fundamental power limits when each user can support an infinite number of capacity-achieving codebooks, but also yield guidelines for practical designs where users can only support a finite number of adaptive modulation and coding (AMC) modes with prescribed symbol error probabilities.

While the assumption of P-CSI renders analysis and design tractable, it may not be always realistic due to possible channel estimation errors at the receiver, feedback delay and jamming [9], [10]. These considerations motivate a *finite-rate* feedback model, where only *quantized* (Q-) CSI is available at the transmitter through a finite number of bits of feedback from the receiver; see e.g., [11]. Based on the finite-rate feedback, [12] minimized transmit-power of orthogonal frequency-division multiplexing (OFDM) systems. In this paper, we consider energy efficiency issues for TDMA over fading channels with finite-rate feedback. Availability of Q-CSI at the transmitters entails a finite number of quantization states. These states are indexed by the bits that the receiver feeds back to transmitters and for each of them the resource allocation is fixed. In this scenario, identifying each quantization state with a time allocation and AMC mode selection per user, emerges as a natural framework to address the Q-CSI problem. Compared with the energy-efficient resource allocation based on finite AMC modes developed for the P-CSI case in [8], the finite-rate feedback case presents three noticeable differences:

- *d1)* Q-CSI calls for designing the quantizer to determine per user the optimal fading region and optimal fixed transmit-power associated with each AMC mode.
- *d2)* With finite-rate feedback, the user can not vary its transmit-power per AMC mode (quantization state) and consequently the BER constraints have to be explicitly imposed.
- *d3)* Quantization and resource allocation in the multi-user setting are interwined.

To tackle these challenges, we need to optimize three subsets of variables: transmit-power, quantization regions and time allocation policies. Instead of optimizing them jointly, we decouple the complicated problem at hand into three sub-problems. In each sub-problem, we optimize over one subset of variables with the other variables remaining fixed. Starting with an initial point provided by the P-CSI solution, we iteratively solve the decoupled sub-problems to descend the global objective. The key finding is an iterative algorithm converging to an energy-efficient quantization and resource allocation solution (Section IV). A sub-optimal yet simplified alternative algorithm which decouples the original problem into two sub-problems is also introduced in Section V. Section VI provides numerical results, before concluding.

II. Modeling Preliminaries

Consider K users linked wirelessly to an access point. The input-output relationship in discrete time is

$$y(n) = \sum_{k=1}^{K} \sqrt{h_k(n)} x_k(n) + z(n), \tag{1}$$

Work in this paper was supported by the ARO Grant No. W911NF-05-1-0283 and was prepared under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

Xin Wang and Georgios B. Giannakis are with the Dept. of Electrical and Computer Engr., Univ. of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, Tel/fax: (612)626-7781/625-4583; Antonio G. Marques is with Dept. of Signal Theory and Communications, Rey Juan Carlos University, Cmno. del molino s/n, Fuenlabrada, Madrid 28943, Spain. Email: {xinwest,antonio,georgios}@ece.umn.edu.

where $x_k(n)$ and $h_k(n)$ are the transmitted signal and fading process of the kth user, respectively, and $z(n)$ denotes AWGN with variance $\sigma^2 = 1$. As in [8], we confine ourselves to TDMA; i.e., when $x_k(n) \neq 0$ in (1), we have $x_i(n) = 0$ for $\forall i \neq k$. We also assume that $\{h_k(n)\}_{k=1}^K$ are jointly stationary and ergodic with continuous stationary distribution. Each channel is slowly time-varying relative to the codeword's length and adheres to a block flat fading model which remains constant for a time block T, but is allowed to change in an independent identically distributed (i.i.d.) fashion from block to block [13, Chapter 2]. Because a frequency-selective channel can be decomposed into a set of parallel time-invariant Gaussian channels, our results apply readily to frequency-selective channels as well. User transmissions to the access point are naturally frame-based, where the frame length is chosen equal to the block length. Given an AMC pool containing a finite number of modes, each user can vary its transmission rate via AMC per block [14]. Having perfect knowledge of $\{h_k\}_{k=1}^K$, the access point assigns time fractions to users and indicates the AMC mode indices (a.k.a. Q-CSI) through a message (uplink map) before an uplink frame, as in e.g., IEEE 802.16 systems [15]. Users then transmit with the indicated AMC modes at the assigned time fractions. Finite-rate feedback from the access point to users consists of a few bits indexing predetermined AMC modes and time slots.

*N*otation: Using boldface lower-case letters to denote column vectors, we let $\mathbf{h} := [h_1, \ldots, h_K]^T$ denote the joint fading state over a block, $F(\mathbf{h})$ the cumulative distribution function (cdf) of joint fading states and $E_{\mathbf{h}}[\cdot]$ the expectation operator over fading states. Furthermore, ^{T} denotes transposition, $\lceil x \rceil$ the minimum integer $\geq x$, $\mathbf{I}_{\{\cdot\}}$ the indicator function ($\mathbf{I}_{\{x\}} = 1$ if x is true and zero otherwise), and $[x]^+ := \max(x, 0)$.

III. Resource Allocation with Finite AMC Modes and Perfect CSI

In this section, we review briefly the energy efficient resource allocation scheme in [8] with finite AMC modes and P-CSI. Besides introducing notation, this solution will be used later to initialize our quantization and resource allocation policies with finite-rate feedback.

We wish to minimize total power under individual average rate constraints in a TDMA system. Given a rate allocation policy $\mathbf{r}(\cdot)$ and a time allocation policy $\boldsymbol{\tau}(\cdot)$, let $\tau_k(\mathbf{h})$ and $r_k(\mathbf{h})$ denote the time *fraction* allocated to user k and the corresponding transmission rate during $\tau_k(\mathbf{h})$. Taking into account that user k does not transmit over the remaining $1-\tau_k(\mathbf{h})$ fraction of time, the kth user's *overall* transmission rate *per block* is $\tau_k(\mathbf{h})r_k(\mathbf{h})$. Also notice that with transmit-power $p_k(\mathbf{h})$ during $\tau_k(\mathbf{h})$ fraction of time in any given block, the kth user's *overall* transmit-power per block is $P_k(\mathbf{h}) = \tau_k(\mathbf{h})p_k(\mathbf{h})$. Suppose that each user can support a finite number of AMC modes. For user $k \in [1, K]$, an AMC mode corresponds to a rate-power pair $(\rho_{k,l}, p_{k,l})$, $l = 1, \ldots, M_k$, where M_k denotes the number of AMC modes. A pair $(\rho_{k,l}, p_{k,l})$ indicates that for transmission rate $\rho_{k,l}$ provided by the lth AMC mode, $p_{k,l}$ is the minimum *receive*-power required to maintain a prescribed BER. Although the kth user only supports M_k AMC modes, this user can still support through time-sharing continuous rates up to a maximum value determined by the highest-rate AMC mode ρ_{k,M_k}. By setting $\rho_{k,0} = 0$ and $p_{k,0} = 0$ and defining $\gamma_{k,l} := (p_{k,l} - p_{k,l-1})/(\rho_{k,l} - \rho_{k,l-1})$, we consider the following *piece-wise linear* function relating transmit-power with rate as (see also [8, Fig. 2])

$$\Upsilon_k(r_k(\mathbf{h})) = \begin{cases} p_{k,l-1}/h_k + \gamma_{k,l}(r_k(\mathbf{h}) - \rho_{k,l-1})/h_k, \\ \qquad \rho_{k,l-1} \leq r_k(\mathbf{h}) \leq \rho_{k,l}, \;\; l \in [1, M_k]; \\ \infty, \quad r_k(\mathbf{h}) > \rho_{k,M_k}. \end{cases} \tag{2}$$

Notice that in order to support rate $\rho_{k,l}$ over a channel h_k, the required transmit-power is scaled as $p_{k,l}/h_k$. For practical modulation-coding schemes with M-QAM constellations and error-control codes, $\Upsilon_k(r_k(\mathbf{h}))$ is guaranteed to be convex [1].

Let $\mathcal{F}$ denote the set of all possible rate and time allocation policies satisfying the individual rate constraints $\left\{E_{\mathbf{h}}[\tau_k(\mathbf{h})r_k(\mathbf{h})] \geq \bar{R}_k\right\}_{k=1}^K$ and $\sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1, \forall \mathbf{h}$. Upon defining $\bar{\mathbf{r}} := [\bar{R}_1, \ldots, \bar{R}_K]^T$, the *power region* under the individual rate constraints is defined as [8]

$$\mathcal{P}(\bar{\mathbf{r}}) = \bigcup_{(\mathbf{r}(\cdot), \boldsymbol{\tau}(\cdot)) \in \mathcal{F}} \mathcal{P}_{TD}(\mathbf{r}(\cdot), \boldsymbol{\tau}(\cdot)), \tag{3}$$

where

$$\mathcal{P}_{TD}(\mathbf{r}(\cdot), \boldsymbol{\tau}(\cdot)) = \left\{\bar{\mathbf{p}} : \forall k, \; \bar{P}_k \geq E_{\mathbf{h}}\left[\tau_k(\mathbf{h})\Upsilon_k(r_k(\mathbf{h}))\right]\right\}. \tag{4}$$

It is easy to show that the K-dimensional $\mathcal{P}(\bar{\mathbf{r}})$ is feasible and convex. With power cost weights $\boldsymbol{\mu} := [\mu_1, \ldots, \mu_K]^T$, the energy-efficient resource allocation policies solve the optimization problem

$$\min_{\bar{\mathbf{p}}} \boldsymbol{\mu}^T \bar{\mathbf{p}}, \qquad \text{subject to (s.t.)} \quad \bar{\mathbf{p}} \in \mathcal{P}(\bar{\mathbf{r}}). \tag{5}$$

The solution $\bar{\mathbf{p}}$ yielding the optimal rate and time allocation is on the boundary of $\mathcal{P}(\bar{\mathbf{r}})$ due to its convexity. By solving (5) for all $\boldsymbol{\mu} \geq \mathbf{0}$, we determine all the boundary points, and thus the whole power region $\mathcal{P}(\bar{\mathbf{r}})$.

Using $\Upsilon_k(x)$, the problem in (5) is equivalent to

$$\begin{cases} \min_{\mathbf{r}(\cdot), \boldsymbol{\tau}(\cdot)} E_{\mathbf{h}}\left[\sum_{k=1}^K \mu_k \tau_k(\mathbf{h}) \Upsilon_k(r_k(\mathbf{h}))\right] \\ \text{s.t.} \;\; \forall \mathbf{h}, \; \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \qquad \forall k, \; E_{\mathbf{h}}\left[\tau_k(\mathbf{h}) r_k(\mathbf{h})\right] \geq \bar{R}_k. \end{cases} \tag{6}$$

Since $\Upsilon_k(r_k(\mathbf{h}))$ is a piece-wise linear function, every $(r_k(\mathbf{h}), \Upsilon_k(r_k(\mathbf{h})))$ pair can be achieved by time-sharing between points $(\rho_{k,l}, p_{k,l}/h_k)$; i.e., we can find a pair of $\check{l} \in [0, M_k]$ and $\check{\tau} \in [0, 1]$ such that $r_k(\mathbf{h}) = \check{\tau}\rho_{k,\check{l}} + (1-\check{\tau})\rho_{k,\check{l}+1}$, and $\Upsilon_k(r_k(\mathbf{h})) = \check{\tau} p_{k,\check{l}}/h_k + (1-\check{\tau}) p_{k,\check{l}+1}/h_k$. Therefore, if we let $\tilde{\boldsymbol{\tau}}(\mathbf{h}) := \{\{\tilde{\tau}_{k,l}\}_{l=0}^{M_k}\}_{k=1}^K$ collect the time fractions per AMC mode, then finding the optimal solution for (6) is equivalent to solving

$$\begin{cases} \min_{\tilde{\boldsymbol{\tau}}(\mathbf{h})} \;\; \sum_{k=1}^K \;\; E_{\mathbf{h}}\left[\sum_{l=0}^{M_k} \mu_k \frac{\tilde{\tau}_{k,l}(\mathbf{h})}{h_k} p_{k,l}\right] \\ \text{s.t.} \;\; \forall \mathbf{h}, \; \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \qquad \forall k, \; E_{\mathbf{h}}\left[\sum_{l=0}^{M_k} \tilde{\tau}_{k,l}(\mathbf{h}) \rho_{k,l}\right] \geq \bar{R}_k. \end{cases} \tag{7}$$

It turns out that the optimal resource allocation policies are obtained via greedy water-filling as summarized next (c.f. [8, Theorem 6]).

Proposition 1: *If $\bar{\mathbf{r}}$ is feasible, $\forall \mathbf{h}$, we have the optimal solution $\tilde{\tau}^*_{k,l}(\mathbf{h})$ ($k \in [1,K]$, $l \in [0, M_k]$) to (7), and subsequently the optimal allocation $r^*_k(\mathbf{h})$ and $\tau^*_k(\mathbf{h})$ for (6) as follows. Given a positive $\boldsymbol{\lambda}^{P*} := [\lambda_1^{P*}, \ldots, \lambda_K^{P*}]^T$, for each fading state $\mathbf{h}$, let $l^*_k := \max\{l : \mu_k \gamma_{k,l}/h_k \le \lambda_k^{P*}\}$ ($l^*_k = 0$ if no such l), and define $\varphi_k(\mathbf{h}) := \mu_k p_{k,l^*_k}/h_k - \lambda_k^{P*}\rho_{k,l^*_k}$.*

1) *If the functions $\{\varphi_k(\mathbf{h})\}_{k=1}^K$ have a single minimum $\varphi_i(\mathbf{h})$, i.e., if $i = \arg\min_k \varphi_k(\mathbf{h})$, then $\tilde{\tau}_{i,l^*_i} = 1$ and all other $\tilde{\tau}_{k,l} = 0$. Consequently,*

$$r^*_i(\mathbf{h}) = \rho_{i,l^*_i}, \qquad \tau^*_i(\mathbf{h}) = 1; \tag{8}$$

and $\forall k \neq i$, $k \in [1,K]$, $r^*_k(\mathbf{h}) = 0$ and $\tau^*_k(\mathbf{h}) = 0$.

2) *If $\{\varphi_k(\mathbf{h})\}_{k=1}^K$ have multiple minima $\{\varphi_{i_j}(\mathbf{h})\}_{j=1}^J$, then $\tilde{\tau}_{i_j,l^*_{i_j}} = \tau^*_j$ with arbitrary $\sum_{j=1}^J \tau^*_j = 1$, and all other $\tilde{\tau}_{k,l} = 0$. Consequently,*

$$r^*_{i_j}(\mathbf{h}) = \rho_{i_j,l^*_{i_j}}, \qquad \tau^*_{i_j}(\mathbf{h}) = \tau^*_j, \tag{9}$$

and $\forall k \neq i_j$, $k \in [1,K]$, $r^*_k(\mathbf{h}) = 0$ and $\tau^*_k(\mathbf{h}) = 0$.

In (8) and (9), $\boldsymbol{\lambda}^{P*}$ and $\{\tau^*_j\}_{j=1}^J$ should satisfy the individual rate constraints

$$E_{\mathbf{h}}\left[\tau^*_k(\mathbf{h}) r^*_k(\mathbf{h})\right] = \bar{R}_k, \quad k = 1, \ldots, K. \tag{10}$$

Moreover, $\boldsymbol{\lambda}^{\boldsymbol{P}^*}$ is almost surely unique and can be iteratively computed by [8, Algorithm 4].

What Proposition 1 asserts is that with P-CSI the optimal access policy per $\mathbf{h}$ consists of the user with smallest cost $\varphi_i(\mathbf{h})$ accessing the channel while the others remaining silent.

IV. Quantization and Resource Allocation with Finite-Rate Feedback

With finite-rate feedback from the access point, particularly in frequency division duplex (FDD) systems, users can only adopt a finite number of resource allocation vectors determined by the Q-CSI of each realization $\mathbf{h}$. For all $k \in [1,K]$ and $l \in [1, M_k]$, let $Q_{k,l}$ denote the quantization region such that when $\mathbf{h} \in Q_{k,l}$, the kth user's lth AMC mode is adopted if user k is selected for transmission. Corresponding to $Q_{k,l}$, an AMC mode can be represented by a rate-power pair $(\rho_{k,l}, \pi_{k,l})$, where $\pi_{k,l}$ is the transmit-power for user k to support rate $\rho_{k,l}$ when $\mathbf{h} \in Q_{k,l}$. Notice that for P-CSI, we represent an AMC mode with a $(\rho_{k,l}, p_{k,l})$ pair where user k varies its *transmit* power for its lth AMC mode to achieve a fixed *receive* power $p_{k,l}$ satisfying the instantaneous BER. However, with Q-CSI, user k is only allowed to use a fixed *transmit* power $\pi_{k,l}$ for its lth mode. While $p_{k,l}$ can be determined by the prescribed BER requirement, we need to optimize $\pi_{k,l}$ in our finite-rate feedback setup.

In this setup, the optimization variables consist of quantization regions $\mathbf{Q} := \{\{Q_{k,l}\}_{l=1}^{M_k}\}_{k=1}^K$, transmit powers $\boldsymbol{\pi} := \{\{\pi_{k,l}\}_{l=1}^{M_k}\}_{k=1}^K$ and the time allocation policy $\boldsymbol{\tau}(\cdot)$. Note that by the definition of $Q_{k,l}$, the rate allocation is absorbed in the quantization design. Let $\epsilon_{k,l}(\gamma)$ denote the BER for a given SNR γ for the kth user's lth AMC mode. For practical modulation-coding schemes with e.g., M-QAM constellations and error-control codes, $\epsilon_{k,l}(\gamma)$ is decreasing and convex [1], [14]. With $\bar{\boldsymbol{\epsilon}} := [\bar{\epsilon}_1, \ldots, \bar{\epsilon}_K]^T$ collecting the prescribed BER requirements, we let $\tilde{\mathcal{F}}$ denote the set of all possible quantization and resource allocation vectors satisfying the rate and BER constraints; i.e., with $\sum_{k=1}^K \tau_k(\mathbf{h}) \le 1$, $\forall \mathbf{h}$, we have

$$\forall k \in [1,K], \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \ge \bar{R}_k; \tag{11}$$

$$\frac{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h})}{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})} \le \bar{\epsilon}_k. \tag{12}$$

where the left-hand side of (12) is the BER averaged over all different regions and rates [12]. Similar to (3), we define the power region as

$$\mathcal{P}(\bar{\mathbf{r}}, \bar{\boldsymbol{\epsilon}}) = \bigcup_{(\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)) \in \tilde{\mathcal{F}}} \mathcal{P}_{TD}(\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)), \tag{13}$$

where

$$\mathcal{P}_{TD}(\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)) = \left\{ \bar{\mathbf{p}} : \forall k, \ \bar{P}_k \ge \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \right\}. \tag{14}$$

Note that for a fixed quantizer, the power region in (13) may not be convex.

With weights $\boldsymbol{\mu}$, the energy-efficient quantization and resource allocation problem is

$$\min_{\bar{\mathbf{p}}} \boldsymbol{\mu}^T \bar{\mathbf{p}}, \qquad \text{s.t.} \quad \bar{\mathbf{p}} \in \mathcal{P}(\bar{\mathbf{r}}, \bar{\boldsymbol{\epsilon}}). \tag{15}$$

Using the previous definitions, this problem is equivalent to

$$\begin{cases} \min_{\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t.} \ \ \forall \mathbf{h}, \ \sum_{k=1}^K \tau_k(\mathbf{h}) \le 1; \\ \quad \forall k, \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \ge \bar{R}_k; \\ \quad \dfrac{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h})}{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})} \le \bar{\epsilon}_k. \end{cases} \tag{16}$$

As the term $\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})$ appears in both rate and BER constraints, we can enhance the BER constraint using the rate constraint as a lower bound. This simplifies the problem to

$$\begin{cases} \min_{\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t.} \ \ \forall \mathbf{h}, \ \sum_{k=1}^K \tau_k(\mathbf{h}) \le 1; \\ \quad \forall k, \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \ge \bar{R}_k; \\ \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \le \bar{\epsilon}_k. \end{cases} \tag{17}$$

If all rate requirements are met with equality in the optimal solution to (16), solving (17) yields the same optimal solution. However, if some rate requirements are over-satisfied in the optimum of (16), the solution of (17) will upperbound that of (16) since we impose stricter BER constraints. Moreover, when $\mathcal{P}(\bar{\mathbf{r}}, \bar{\boldsymbol{\epsilon}})$ is not convex, we may not be able to determine all the boundary points (and thus the whole power region) by solving (15) for all $\boldsymbol{\mu} \ge \mathbf{0}$. However, we will still presumingly use the solutions of (17) for all $\boldsymbol{\mu} \ge \mathbf{0}$ as the

"boundary points" to approximately characterize the power region $\mathcal{P}(\bar{\mathbf{r}}, \bar{\boldsymbol{\epsilon}})$. Clearly, the resultant power region will be a conservative approximation contained in $\mathcal{P}(\bar{\mathbf{r}}, \bar{\boldsymbol{\epsilon}})$.

The problem (17) is still complicated and not convex. To solve it, we divide it into three separate sub-problems and then solve each of them in an optimal way; i.e., we resort to a coordinate descent [16] approach to come up with an iterative algorithm which assembles the different sub-solutions to solve the main problem. Notice that this is a well appreciated strategy in the field of quantization theory, and a good example is the Lloyd algorithm [18]. To optimally quantize a real-vector with a fixed number of bits, Lloyd's algorithm iterates between the following two steps: i) given the codewords, find the optimal regions; and ii) given the regions, update the optimal codewords.

A. Initialization

We first use the resource allocation policies of Section III to initialize our coordinate descent method. Given AMC modes and P-CSI, Proposition 1 yields the energy-efficient rate and time allocation policies $\mathbf{r}^*(\cdot)$ and $\boldsymbol{\tau}^*(\cdot)$ via greedy water-filling. With the associated Lagrange multiplier vector $\boldsymbol{\lambda}^{P*}$, we can derive the quantization regions $\mathbf{Q}^*$ corresponding to the rate allocation $\mathbf{r}^*(\cdot)$:

Lemma 1: *With rate allocation $\mathbf{r}^*(\cdot)$, the optimal region $Q_{k,l}^*$ for user $k \in [1, K]$ is given by*

$$Q_{k,l}^* = \left\{ \mathbf{h} : \ h_k \in [q_{k,l}^*, q_{k,l+1}^*) \right\}, \tag{18}$$

where $q_{k,l}^* = \mu_k \gamma_{k,l} / \lambda_k^{P*}$ for $l \in [1, M_k]$ and $q_{k,M_k+1}^* = \infty$.

Proof: Since user selection is determined by the time allocation, region $Q_{k,l}^*$ must be specified only when the lth AMC mode is employed by user k. From $\mathbf{r}^*(\cdot)$, user k selects mode index $l_k^* := \max \{l : \ \mu_k \gamma_{k,l}/h_k \leq \lambda_k^{P*}\}$, $\forall \mathbf{h}$. By the convexity of $\Upsilon_k(r_k(\mathbf{h}))$, this implies that when $\mu_k \gamma_{k,l}/\lambda_k^{P*} \leq h_k < \mu_k \gamma_{k,l+1}/\lambda_k^{P*}$, the lth mode is picked, and thus (18) follows. □

With the allocation of time slots specified by $\boldsymbol{\tau}^*(\cdot)$ and quantization regions by $\mathbf{Q}^*$ provided by Proposition 1 and Lemma 1, we are ready to optimize over the transmit powers $\boldsymbol{\pi}$.

B. Optimal Transmit-Powers

It is clear from (17) that the rate constraints affect to $\boldsymbol{\tau}(\cdot)$ and $\mathbf{Q}$. Since $\mathbf{r}^*(\cdot)$ and $\boldsymbol{\tau}^*(\cdot)$ in Proposition 1 satisfy the rate constraints, so do the equivalent $\mathbf{Q}^*$ and $\boldsymbol{\tau}^*(\cdot)$. Moreover, in each step of our coordinate descent algorithm, we will descend the global objective within the feasible set. This guarantees that we can always start with a pair of $\mathbf{Q}$ and $\boldsymbol{\tau}(\cdot)$ already satisfying rate constraints to find the optimal $\boldsymbol{\pi}$. Now given these $\mathbf{Q}$ and $\boldsymbol{\tau}(\cdot)$, finding the optimal $\boldsymbol{\pi}$ is to solve

$$\begin{cases} \min_{\boldsymbol{\pi}} \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall k, \ \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \leq \bar{\epsilon}_k. \end{cases} \tag{19}$$

Let us define $A_{k,l} := \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})$. To prevent the trivial solution, we assume that all $A_{k,l} \neq 0$. If some $A_{k,l}$ are zero, we can just remove the corresponding AMC modes from consideration and reformulate (19) using a compact $\mathbf{Q}$ containing AMC modes with non-zero measures. Since the functions $\epsilon_{k,l}(x)$ are convex, (19) is a convex optimization problem. Its solution can be analytically obtained as follows.

Proposition 2: *Given a positive $\boldsymbol{\nu}^{\pi*} := [\nu_1^{\pi*}, \ldots, \nu_K^{\pi*}]^T$, and with $\epsilon'_{k,l}(\gamma)$ denoting the first derivative of $\epsilon_{k,l}(\gamma)$, the optimal $\pi_{k,l}^*$ is the unique value such that $\pi_{k,l}^* = 0$ or*

$$\int_{Q_{k,l}} \tau_k(\mathbf{h}) h_k \epsilon'_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) = -\frac{\mu_k \bar{R}_k A_{k,l}}{\rho_{k,l} \nu_k^{\pi*}}. \tag{20}$$

And $\forall k \in [1, K]$, each Lagrange multiplier $\nu_k^{\pi*}$ is determined by satisfying the BER constraint

$$\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) / \bar{R}_k = \bar{\epsilon}_k. \tag{21}$$

Proof: Since (19) is a convex optimization problem, strong duality holds and the Karush-Kuhn-Tucker (KKT) conditions are sufficient and necessary for the optimality [17]. With implicit constraints $\pi_{k,l} \geq 0$, $\forall k \in [1, K]$, $l \in [1, M_k]$, the KKT conditions are

$$\mu_k A_{k,l} + \nu_k^{\pi*} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) h_k \epsilon'_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) - \alpha_{k,l}^{\pi*} = 0; \tag{22}$$

where $\alpha_{k,l}^{\pi*} \geq 0$ is the Lagrange multiplier corresponding to the constraint $\pi_{k,l} \geq 0$, and $\nu_k^{\pi*} \geq 0$ is the Lagrange multiplier for the BER constraint. We next show that $\nu_k^{\pi*} \neq 0$, $\forall k$. Supposing that a certain $\nu_k^{\pi*} = 0$, we find [c.f. (22)]

$$\mu_k A_{k,l} = \alpha_{k,l}^{\pi*}, \quad l \in [1, M_k]. \tag{23}$$

Since $\forall l$, $A_{k,l} > 0$, we have that $\forall l$, $\alpha_{k,l}^{\pi*} > 0$. By complementary slackness [17] between $\alpha_{k,l}^{\pi*}$ and constraint $\pi_{k,l}^* \geq 0$, this implies that $\pi_{k,l}^* = 0$, $\forall l$. But then the BER for user k becomes 0.5 (we exclude the trivial case where $\bar{\epsilon}_k \geq 0.5$). This contradiction implies that $\forall k$, we have $\nu_k^{\pi*} > 0$. Also by the complementary slackness, we have from (22) that $\pi_{k,l}^* = 0$, or when $\pi_{k,l}^* > 0$ (thus $\alpha_{k,l}^{\pi*} = 0$),

$$\mu_k A_{k,l} + \nu_k^{\pi*} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) h_k \epsilon'_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) = 0; \tag{24}$$

which readily leads to (20). Since $\nu_k^{\pi*} > 0$, $\forall k$, the complementary slackness conditions imply that the BER constraints are achieved with equality as in (21). □

Notice that given $\tau_k(\mathbf{h})$, users are decoupled. Solving (19) is equivalent to solving K small problems; i.e., $\forall k$, $\min \ \ \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})$ subject to $\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) / \bar{R}_k \leq \bar{\epsilon}_k$. Given $\nu_k^{\pi*}$ and monotonically decreasing $\epsilon_{k,l}(\gamma)$, the solution to (20) is unique for $\pi_{k,l}^* > 0$ and we can use a one-dimensional, e.g., bi-sectional, search to obtain this $\pi_{k,l}^*$. Then we can use another one-dimensional search to solve for $\nu_k^{\pi*}$ from (21). And the optimal transmit-powers $\boldsymbol{\pi}^*$ are in turn obtained.

C. Optimal Quantization Regions

Given $\boldsymbol{\pi}$ and $\boldsymbol{\tau}(\cdot)$, users are decoupled as in Proposition 2. To find the optimal $\mathbf{Q}$ (fading regions), we need to solve $\forall k$,

$$\begin{cases} \min_{\{Q_{k,l}\}_{k=1}^{M_k}} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t.} \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k; \\ \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \leq \bar{\epsilon}_k. \end{cases} \tag{25}$$

Proposition 3: *Given non-negative λ_k^{q*} and ν_k^{q*}, we define $\psi_{k,l}(h_k) := \mu_k \pi_{k,l} - \lambda_k^{q*} \rho_{k,l} + \nu_k^{q*} \rho_{k,l} \epsilon_{k,l}(\pi_{k,l} q)/\bar{R}_k$ for $l \in [1, M_k]$ and $\psi_{k,0}(h_k) = 0$. Then we can obtain the optimal $Q_{k,l}^*$ as: $\forall l \in [1, M_k]$,*

$$Q_{k,l}^* = \{\mathbf{h} : \ \psi_{k,l}(h_k) \leq \psi_{k,j}(h_k); \ \forall j \neq l, j \in [0, M_k]\}. \tag{26}$$

Moreover, λ_k^{q*} and ν_k^{q*} are determined by satisfying the conditions

$$\lambda_k^{q*} = 0 \quad \text{or} \quad \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \bar{R}_k; \tag{27}$$

$$\nu_k^{q*} = 0 \quad \text{or} \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) = \bar{\epsilon}_k. \tag{28}$$

Proof: Similar to the constrained vector quantization in [19] and [20], by using Lagrange multipliers λ_k^{q*} and ν_k^{q*}, we can define a distortion measure

$$\begin{aligned} d(\mathbf{h}, Q_{k,l}) \ := \ & \mu_k \pi_{k,l} \tau_k(\mathbf{h}) - \lambda_k^{q*} \rho_{k,l} \tau_k(\mathbf{h}) \\ & + \nu_k^{q*} \rho_{k,l} \epsilon_{k,l}(\pi_{k,l} q) \tau_k(\mathbf{h})/\bar{R}_k. \end{aligned} \tag{29}$$

Then the optimal $Q_{k,l}^*$ solving (25) should minimize the overall average distortion measure

$$\begin{aligned} E_{\mathbf{h}}\left[d(\mathbf{h}, Q_{k,l})\right] = & \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ & - \lambda_k^{q*} \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ & + \nu_k^{q*} \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}). \end{aligned} \tag{30}$$

This in turn implies that $Q_{k,l}^*$ can be determined by the *nearest-neighbor rule* [19], [20]; i.e.,

$$\mathbf{h} \in Q_{k,l}, \quad \text{iff} \ \ d(\mathbf{h}, Q_{k,l}) \leq d(\mathbf{h}, Q_{k,j}), \ \forall j \neq l, j \in [0, M_k]. \tag{31}$$

Since $\tau_k(\mathbf{h}) \geq 0$, we can readily derive (26). And (27) and (28) directly come from the complementary slackness of the KKT conditions [17]. □

Note that we can also define a region $Q_{k,0}^*$ as the set complement of $\bigcup_{l \in [1, M_k]} Q_{k,l}^*$. When $\mathbf{h} \in Q_{k,0}^*$, user k will surely defer. To obtain the optimal $Q_{k,l}^*$, we need to find λ_k^{q*} and ν_k^{q*}. Since (25) is not a convex problem, we resort to a two-dimensional search. We can start the search in an exhaustive manner. However, once we have a pair of λ_k^{q*} and ν_k^{q*} satisfying (27) and (28), we stop the search and return these values. After obtaining $\boldsymbol{\lambda}^{q*} := [\lambda_1^{q*}, \ldots, \lambda_K^{q*}]^T$ and $\boldsymbol{\nu}^{q*} := [\nu_1^{q*}, \ldots, \nu_K^{q*}]^T$ with K two-dimensional searches, we in turn determine $\mathbf{Q}^*$. Interestingly, for all the simulations in Sec. VI the optimal quantization regions $\{Q_{k,l}^*\}_{l=1}^{M_k}$, $\forall k$, turn out to be a set of non-overlapping consecutive intervals of h_k; i.e., $\{Q_{k,l}^*\}_{l=1}^{M_k}$ can be determined by a set of thresholds $\{q_{k,l}^*\}_{l=1}^{M_k}$ as in (18).

D. Optimal Time Allocation

With $\mathbf{Q}$ and $\boldsymbol{\pi}$ given, finding the optimal time allocation policy is to solve

$$\begin{cases} \min_{\boldsymbol{\tau}(\cdot)} \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t.} \ \forall \mathbf{h}, \ \sum_{k=1}^{K} \tau_k(\mathbf{h}) \leq 1; \\ \quad \forall k, \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k; \\ \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \leq \bar{\epsilon}_k. \end{cases} \tag{32}$$

Similar to Proposition 1, we can also obtain the optimal $\boldsymbol{\tau}^*(\cdot)$ via a greedy approach.

Proposition 4: *Given $\boldsymbol{\lambda}^{\tau*} := [\lambda_1^{\tau*}, \ldots, \lambda_K^{\tau*}]^T \geq \mathbf{0}$ and $\boldsymbol{\nu}^{\tau*} := [\nu_1^{\tau*}, \ldots, \nu_K^{\tau*}]^T \geq \mathbf{0}$, for each fading state $\mathbf{h}$, let $l_k(\mathbf{h})$ denote the mode index for user k such that $\mathbf{h} \in Q_{k,l_k(\mathbf{h})}$, and define $\tilde{\varphi}_k(\mathbf{h}) := \mu_k \pi_{k,l_k(\mathbf{h})} - \lambda_k^{\tau*} \rho_{k,l_k(\mathbf{h})} + \nu_k^{\tau*} \rho_{k,l_k(\mathbf{h})} \epsilon_{k,l_k(\mathbf{h})} \left(h_k \pi_{k,l_k(\mathbf{h})}\right)/\bar{R}_k$. Then the optimal solution $\boldsymbol{\tau}^*(\cdot)$ to (32) can be obtained as follows:*

1) *If $\forall k \in [1, K]$, $\tilde{\varphi}_k(\mathbf{h}) \geq 0$, then $\forall k$, $\tau_k^*(\mathbf{h}) = 0$.*
2) *If $\{\tilde{\varphi}_k(\mathbf{h})\}_{k=1}^K$ have a single minimum $\tilde{\varphi}_i(\mathbf{h}) < 0$, then $\tau_i^*(\mathbf{h}) = 1$ and $\forall k \neq i$, $k \in [1, K]$, $\tau_k^*(\mathbf{h}) = 0$.*
3) *If $\{\tilde{\varphi}_k(\mathbf{h})\}_{k=1}^K$ have multiple minima $\left\{\tilde{\varphi}_{i_j}(\mathbf{h})\right\}_{j=1}^J < 0$, then $\tau_{i_j}^*(\mathbf{h}) = \tau_j^*$ with arbitrary $\sum_{j=1}^J \tau_j^* = 1$, and $\forall k \neq i_j$, $k \in [1, K]$, $\tau_k^*(\mathbf{h}) = 0$.*

Moreover, $\boldsymbol{\lambda}^{\tau*}$ and $\boldsymbol{\nu}^{\tau*}$ should satisfy the complementary slackness conditions $\forall k \in [1, K]$,

$$\lambda_k^{\tau*} = 0 \quad \text{or} \quad \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k^*(\mathbf{h}) dF(\mathbf{h}) = \bar{R}_k;$$

$$\nu_k^{\tau*} = 0 \quad \text{or} \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k^*(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) = \bar{\epsilon}_k.$$

Proof: See Appendix. □

As with P-CSI, Proposition 4 asserts that our optimal time allocation strategies are "greedy". Function $\tilde{\varphi}_k(\mathbf{h})$ can be viewed as a channel quality indicator (the smaller the better) for user k. Then for each time block, we should only allow the user with the "best" channel to transmit. When there are multiple users with "best" channels, arbitrary time division among them suffices. Since $\tilde{\varphi}_k(\mathbf{h})$ contains $\lambda_k^{\tau*}$ and $\nu_k^{\tau*}$, this implies that the user having smallest $\tilde{\varphi}_k(\mathbf{h})$ actually has the rate and BER constraints controlled "best" channel. For cases where $\tilde{\varphi}_k(\mathbf{h}) \geq 0 \ \forall k \in [1, K]$, we should let all users to defer. Imagine that there is a fictitious user which has no rate and BER requirements and always keeps silent. Then $\forall \mathbf{h}$, its channel quality indicator is zero. If $\tilde{\varphi}_k(\mathbf{h}) \geq 0 \ \forall k \in [1, K]$, picking this fictitious user is clearly most efficient. This implies that in these cases no user should transmit. Notice that in Proposition 1, the case $\varphi_k(\mathbf{h}) > 0$ never occurs, since it is easy to show that $\varphi_k(\mathbf{h}) = 0$ when $h_k = 0$ and $\varphi_k(\mathbf{h})$ is a decreasing function of h_k.

$$
\begin{aligned}
g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau) \;=\; & \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau, \mathbf{h}) dF(\mathbf{h}) \\
& - \sum_{k=1}^{K} \lambda_k^\tau \left(\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau, \mathbf{h}) dF(\mathbf{h}) - \bar{R}_k \right) \\
& + \sum_{k=1}^{K} \nu_k^\tau \left(\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau, \mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) / \bar{R}_k - \bar{\epsilon}_k \right); \qquad (33)
\end{aligned}
$$

To obtain the optimal $\boldsymbol{\tau}^*(\cdot)$, we need to find $\boldsymbol{\lambda}^{\tau *}$ and $\boldsymbol{\nu}^{\tau *}$. Instead of a $2K$-dimensional exhaustive search, we accomplish this by a sub-gradient ascend algorithm. First, it follows readily that the Lagrange dual function $g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)$ for (32) is given by (33), where for a given $(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)$, the time allocation $\tau_k(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau, \mathbf{h})$ is provided by Proposition 4 (without considering the rate and BER constraints). The dual of (32) is

$$
\max_{\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau} \; g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau), \qquad \text{s.t. } \boldsymbol{\lambda}^\tau \geq \mathbf{0}, \boldsymbol{\nu}^\tau \geq \mathbf{0}. \qquad (34)
$$

Since (32) is a convex problem, the duality gap is zero; and thus $(\boldsymbol{\lambda}^{\tau *}, \boldsymbol{\nu}^{\tau *}) = \arg \max_{\boldsymbol{\lambda}^\tau \geq \mathbf{0}, \boldsymbol{\nu}^\tau \geq \mathbf{0}} g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)$. Therefore, we can obtain $(\boldsymbol{\lambda}^{\tau *}, \boldsymbol{\nu}^{\tau *})$ via the following sub-gradient projection algorithm. Note that the dual function $g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)$ is concave since it is the point-wise infimum of a family of affine functions of $(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)$, and thus the convergence of our sub-gradient projection algorithm is guaranteed [17].

Algorithm 1: *T0) initialization:* Generate an arbitrary non-negative vector $(\boldsymbol{\lambda}^\tau(0), \boldsymbol{\nu}^\tau(0))$. Select tolerance $\varepsilon > 0$, calculate $g(\boldsymbol{\lambda}^\tau(0), \boldsymbol{\nu}^\tau(0))$ and let the iteration index $t = 1$.

T1) $\forall k \in [1, K]$, numerically evaluate the partial derivatives $\Delta\lambda_k^\tau := \frac{\partial g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)}{\partial \lambda_k^\tau}$ and $\Delta\nu_k^\tau := \frac{\partial g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)}{\partial \nu_k^\tau}$ at $(\boldsymbol{\lambda}^\tau(t-1), \boldsymbol{\nu}^\tau(t-1))$. Choose a step size δ by line search and then update $\lambda_k^\tau(t) = [\lambda_k^\tau(t-1) + \delta \Delta\lambda_k^\tau]^+$ and $\nu_k^\tau(t) = [\nu_k^\tau(t-1) + \delta \Delta\nu_k^\tau]^+$.

T2) Stopping criterion: Calculate the objective $g(\boldsymbol{\lambda}^\tau(t), \boldsymbol{\nu}^\tau(t))$ using $(\boldsymbol{\lambda}^\tau(t), \boldsymbol{\nu}^\tau(t))$. If

$$
\frac{g(\boldsymbol{\lambda}^\tau(t), \boldsymbol{\nu}^\tau(t)) - g(\boldsymbol{\lambda}^\tau(t-1), \boldsymbol{\nu}^\tau(t-1))}{g(\boldsymbol{\lambda}^\tau(t), \boldsymbol{\nu}^\tau(t))} < \varepsilon,
$$

return $(\boldsymbol{\lambda}^\tau(t), \boldsymbol{\nu}^\tau(t))$ and stop. Otherwise, increase t by 1 and go to *T1)*.

Once $\boldsymbol{\lambda}^{\tau *}$ and $\boldsymbol{\nu}^{\tau *}$ are calculated, the time allocation policy in Proposition 4 is in turn determined. Although we rely on Q-CSI, Proposition 4 employs continuous values of the channel realizations to find the optimal time allocation. This is not a contradiction because time allocation is carried out at the receiver where P-CSI is available. We will show in the sequel (Proposition 5) that the transmitters need only Q-CSI.

E. Joint Quantization and Resource Allocation Algorithm

For the global objective

$$
J := \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}),
$$

we propose based on Propositions 1-4 the following joint quantization and resource allocation (JQRA) algorithm:

Algorithm 2: *J0) initialization:* Produce initial time allocation $\boldsymbol{\tau}^{(0)}(\cdot)$ and quantization regions $\mathbf{Q}^{(0)}$ from Proposition 1 and Lemma 1. Select tolerance $\varepsilon > 0$, initialize objective at $J^{(0)} = \infty$ and set the iteration index $t = 1$.

J1) $\boldsymbol{\tau}^{(t-1)}(\cdot), \mathbf{Q}^{(t-1)} \rightarrow \boldsymbol{\pi}^{(t)}$: Given $\boldsymbol{\tau}^{(t-1)}(\cdot)$ and $\mathbf{Q}^{(t-1)}$, obtain $\boldsymbol{\pi}^{(t)}$ from Proposition 2.

J2) $\boldsymbol{\pi}^{(t)}, \boldsymbol{\tau}^{(t-1)}(\cdot) \rightarrow \mathbf{Q}^{(t)}$: Given $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t-1)}(\cdot)$, obtain $\mathbf{Q}^{(t)}$ from Proposition 3.

J3) $\mathbf{Q}^{(t)}, \boldsymbol{\pi}^{(t)} \rightarrow \boldsymbol{\tau}^{(t)}(\cdot)$: Given $\mathbf{Q}^{(t)}$ and $\boldsymbol{\pi}^{(t)}$, obtain $\boldsymbol{\tau}^{(t)}(\cdot)$ from Proposition 4 with Algorithm 1.

J4) Stopping criterion: Calculate the objective $J^{(t)}$ using $\mathbf{Q}^{(t)}$, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\cdot)$. If $(J^{(t-1)} - J^{(t)})/J^{(t)} < \varepsilon$, return the current quantization and resource allocation and stop. Otherwise, increase t by 1 and go to *J1)*.

Since the global objective J is decreasing in each step, it is easy to see that as $t \rightarrow \infty$, the JQRA algorithm converges. As we mentioned at the beginning of this section, this algorithm follows the coordinate descent principle [16] and the quantization design is also reminiscent of the well-known Lloyd algorithm [18].

F. Optimal Feedback Bits

JQRA provides a quantizer design which is computed off-line and there is no need to execute it again if the statistics of the channel and the quality-of-service requirements of the users do not change. Once the quantization and resource allocation strategy is solved by JQRA, the access point quantizes each fading state and feeds back the user-AMC-mode selections per time block. Then users defer or transmit with the indicated AMC modes.

Proposition 5: *Given the quantizer design and user selection policy represented by* $\mathbf{Q}^*$, $\boldsymbol{\pi}^*$, $\boldsymbol{\lambda}^{\tau *}$ *and* $\boldsymbol{\nu}^{\tau *}$ *in JQRA,* $\forall \mathbf{h}$,

the access point sends to the users the codeword $c^*(\mathbf{h}) = [k^*(\mathbf{h}); l^*(\mathbf{h})]$ *which encodes the optimal resource allocation for the current fading state, so that:*

1) $k^*(\mathbf{h}) = \arg\min_k \{\tilde{\varphi}_k(\mathbf{h}, \mathbf{Q}^*, \boldsymbol{\pi}^*, \boldsymbol{\lambda}^{\tau*}, \boldsymbol{\nu}^{\tau*})\}_{k=1}^K$ *(pick any* k^* *when multiple minima occur), where* $\tilde{\varphi}_k(\mathbf{h}, \mathbf{Q}^*, \boldsymbol{\pi}^*, \boldsymbol{\lambda}^{\tau*}, \boldsymbol{\nu}^{\tau*}) := \mu_k \pi^*_{k,l_k(\mathbf{h})} - \lambda_k^{\tau*} \rho_{k,l_k(\mathbf{h})} + \nu_k^{\tau*} \rho_{k,l_k(\mathbf{h})} \epsilon_{k,l_k(\mathbf{h})} \left(h_k \pi^*_{k,l_k(\mathbf{h})}\right) / \bar{R}_k$.
2) $l^*(\mathbf{h}) = \{\ l;\ s.t.\ \mathbf{h} \in Q_{k^*(\mathbf{h}),l},\quad l = 1, \ldots, M_k\}$.

When the users receive the broadcasted codeword $c^*(\mathbf{h}) = [k^*(\mathbf{h}); l^*(\mathbf{h})]$*, the optimal multiple access to the channel consists of the* k^**th user transmitting its* l^**th mode using power* $\pi^*_{k^*(\mathbf{h}),l^*(\mathbf{h})}$ *while the rest of the users remaining inactive.*

Proof: This is a direct consequence of the optimal time allocation in Proposition 4 and the definitions of $\mathbf{Q}^*$ and $\boldsymbol{\pi}^*$. □

Proposition 5 implies the optimal resource allocation policy can be obtained by letting only one user to transmit per fading state. In other words, over all possible strategies, the optimal solution only allows to activate one AMC mode of one user per block. Therefore, we only need $\lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ feedback bits to index $\sum_{k=1}^K M_k$ different user-AMC-mode combinations and the case of all users deferring. For illustration, consider a system where 85-170 users are present, each supporting six different AMC modes, as in IEEE 802.16 systems [15]. To implement the derived resource allocation, in this case the access point only needs to feed back 9-10 bits per fading state. This is certainly affordable by most practical systems.

G. Users with Independent Fading Processes

Since the optimal $\{\tau_k(\mathbf{h})\}_{k=1}^K$ is not available analytically, in general multi-dimensional integrals are involved in solving (19), (25) and (32). To gain more insight, let us look at a special case where the channels $\{h_k\}_{k=1}^K$ are independent. Since the value of $\tilde{\varphi}_k(\mathbf{h})$ defined in Proposition 4 is only determined by the value of h_k, we henceforth interchangeably use $\tilde{\varphi}_k(\mathbf{h})$ and $\tilde{\varphi}_k(h_k)$. Let $q_{i,l}^{\min}$ and $q_{i,l}^{\max}$ denote the maximum and minimum values for $h_i \in Q_{i,l}$. Since $\tilde{\varphi}_i(h_i)$ is monotonically decreasing for $h_i \in Q_{i,l}$, we define

$$s_{i,k}^{(l)}(h_k) = \begin{cases} q_{i,l}^{\min}, & \tilde{\varphi}_i(q_{i,l}^{\min}) \le \tilde{\varphi}_k(h_k), \\ q_{i,l}^{\max}, & \tilde{\varphi}_i(q_{i,l}^{\max}) \ge \tilde{\varphi}_k(h_k), \\ s, & \tilde{\varphi}_i(s) = \tilde{\varphi}_k(h_k),\ q_{i,l} < s < q_{i,l+1}, \end{cases} \tag{35}$$

If $F_k(\cdot)$ stands for the cdf of user k's fading channel, we can establish that:

Lemma 2: *If the fading processes of users are independent and* $\mathbf{h}_{-k} := [h_1, \ldots, h_{k-1}, h_{k+1}, \ldots, h_K]^T$*, then we have*

$$\int_{\mathbf{h}_{-k}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \mathbf{I}_{\{\tilde{\varphi}_k(h_k)<0\}} \times \prod_{i \ne k} \Bigg[\Pr(Q_{i,0}) + \sum_{l=1}^{M_i} \Pr(Q_{i,l}) \Pr\left(h_i < s_{i,k}^{(l)}(h_k) \Big| Q_{i,l} \right) \Bigg] dF_k(h_k) \tag{36}$$

where $\Pr(Q_{i,l})$ *denote the probability measure of region* $Q_{i,l}$ *and* $\Pr\left(h_i < s_{i,k}^{(l)}(h_k) \Big| Q_{i,l} \right)$ *denote the probability of the event* $h_i < s_{i,k}^{(l)}(h_k)$ *when* $h_i \in Q_{i,l}$.

Proof: Since $\epsilon_{k,l}(x)$ is decreasing, it is clear that $\tilde{\varphi}_k(h_k)$ decreases as $h_k \in Q_{k,l}$ increases; hence, for $h_i < s_{i,k}^{(l)}(h_k)$, we have $\tilde{\varphi}_i(h_i) > \tilde{\varphi}_k(h_k)$. With continuous fading distributions, in our optimal time allocation, $\tau_k(\mathbf{h}) = 1$ when $\tilde{\varphi}_k(h_k) < 0$ and $\forall i \ne k$, $\tilde{\varphi}_i(h_i) > \tilde{\varphi}_k(h_k)$; otherwise, $\tau_k(\mathbf{h}) = 0$. Together with the fact $dF(\mathbf{h}) = \prod_k dF_k(h_k)$ when $\{h_k\}_{k=1}^K$ are independent, we have

$$\begin{aligned} \int_{\mathbf{h}_{-k}} \tau_k(\mathbf{h}) dF(\mathbf{h}) &= \int_{\mathbf{h}_{-k}} \mathbf{I}_{\{\tilde{\varphi}_k(\mathbf{h})<0 \ \&\ \forall i, \tilde{\varphi}_i(\mathbf{h}) > \tilde{\varphi}_k(\mathbf{h})\}} dF(\mathbf{h}) \\ &= \mathbf{I}_{\{\tilde{\varphi}_k(h_k)<0\}} \left[\prod_{i \ne k} \int_{h_i} \mathbf{I}_{\{\tilde{\varphi}_i(h_i) > \tilde{\varphi}_k(h_k)\}} dF_i(h_i) \right] dF_k(h_k) \\ &= \mathbf{I}_{\{\tilde{\varphi}_k(h_k)<0\}} \prod_{i \ne k} \Bigg[\Pr(Q_{i,0}) + \sum_{l=1}^{M_i} \Pr(Q_{i,l}) \Pr\left(h_i < s_{i,k}^{(l)}(h_k) \Big| Q_{i,l} \right) \Bigg] dF_k(h_k). \end{aligned}$$

□

Define

$$f_k(h_k) := \mathbf{I}_{\{\tilde{\varphi}_k(h_k)<0\}} \prod_{i \ne k} \Bigg[\Pr(Q_{i,0}) + \sum_{l=1}^{M_i} \Pr(Q_{i,l}) \Pr\left(h_i < s_{i,k}^{(l)}(h_k) \Big| Q_{i,l} \right) \Bigg].$$

Then using Lemma 2, we can solve the optimal transmit-powers $\boldsymbol{\pi}^{(t)}$ from given $\boldsymbol{\tau}^{(t-1)}(\cdot)$ and $\mathbf{Q}^{(t-1)}$ as follows [c.f. Proposition 2].

Corollary 1: *If the fading processes of users are independent, given the positive* $\boldsymbol{\nu}^{\pi*}$*, the optimal* $\pi^*_{k,l}$ *is the unique value for which* $\pi^*_{k,l} = 0$ *or*

$$\int_{h_k \in Q_{k,l}} f_k(h_k) h_k \epsilon'_{k,l}(h_k \pi^*_{k,l}) dF_k(h_k) = -\frac{\mu_k \bar{R}_k A_{k,l}}{\rho_{k_l} \nu_k^{\pi*}}. \tag{37}$$

And $\forall k \in [1, K]$*,* $\nu_k^{\pi*}$ *is determined by satisfying the BER constraint*

$$\sum_{l=1}^{M_k} \rho_{k,l} \int_{h_k \in Q_{k,l}} f_k(h_k) \epsilon_{k,l}(h_k \pi^*_{k,l}) dF_k(h_k) / \bar{R}_k = \bar{\epsilon}_k. \tag{38}$$

Corollary 1 shows that to carry out the step *J1)* in the JQRA algorithm, we only need one-dimensional integrals to solve for $\boldsymbol{\pi}^{(t)}$. Similar to Corollary 1, we can also derive the counterparts of Propositions 3 and 4 for independently fading channels. In this case, steps *J2)* and *J3)* in the JQRA algorithm only require numerical calculations of one-dimensional integrals[1].

[1]For the initialization step, the derivation is slightly different but even more compact. It was established in [8, Corollary 2] that, letting $s_{i,k}(h_k)$ denote the solution to $\varphi_i(h_i) - \varphi_k(h_k) = 0$, i.e., $\varphi_i(s_{i,k}(h_k)) = \varphi_k(h_k)$, then $\int_{\mathbf{h}_{-k}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \prod_{i \ne k} F_i(s_{i,k}(h_k))\ dF_k(h_k)$ since $\varphi_k(h_k)$ defined in Proposition 1 is decreasing function of h_k.

V. Simplified JQRA Algorithm

In the previous section we solved (16) by an iterative algorithm and specified the resulting resource allocation policies as in Proposition 5. Although JQRA yields optimal quantization and time-power allocation based on Q-CSI, certain applications could benefit from a less complex algorithm. This motivates the simplified JQRA (S-JQRA) algorithm we derive in this section.

Recall that with P-CSI, each user can adapt its transmit-power to instantaneously achieve the required BER level. However, this is not feasible with Q-CSI since the transmit-power per quantization region per user is fixed. Nevertheless, we can mimic this strategy to simplify (16) as follows. Given the quantization regions and time allocation policy, we uniquely determine the transmit-power for each quantization region so that each user's average BER per region attains the BER target. Clearly, this is a suboptimal approach since we do not optimize transmit-powers once the quantization regions and time allocation policy are given. However, it simplifies the process because: i) BER constraints do not explicitly appear; and ii) the number of optimization variables is reduced to two subsets of quantization regions and the time allocation policy. With this simplification, the optimization problem to solve becomes

$$\begin{cases} \min_{\mathbf{Q},\boldsymbol{\tau}(\cdot)} \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^{K} \tau_k(\mathbf{h}) \leq 1; \\ \quad \forall k, \ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k. \end{cases} \tag{39}$$

where $\pi_{k,l}$ is uniquely determined by the equation $f_\epsilon(\pi_{k,l}, \tau_k(\mathbf{h}), Q_{k,l}) = 0$, and

$$\begin{aligned} & f_\epsilon(\pi_{k,l}, \tau_k(\mathbf{h}), Q_{k,l}) \\ & \quad := \frac{\int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h})}{\int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})} - \bar{\epsilon}_k. \end{aligned} \tag{40}$$

It is easy to check that with so-determined transmit-powers, the average BER constraints are satisfied. Similar to the JQRA algorithm, now we can divide the optimization process into two sub-problems and resort to a coordinate descent approach: i) given the time allocation, we calculate the optimal quantization regions; and ii) with the new quantization regions, we update the optimal time allocation policy.

To optimize the quantization regions, we need to solve $\forall k$,

$$\begin{cases} \min_{\{Q_{k,l}\}_{k=1}^{M_k}} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k. \end{cases} \tag{41}$$

The counterpart of Proposition 3 is now as follows:

Proposition 6: *Given a positive λ_k^{s*}, we define $\tilde{\psi}_{k,l}(h_k) := \mu_k \pi_{k,l} - \lambda_k^{s*} \rho_{k,l}$ for $l \in [1, M_k]$ and $\tilde{\psi}_{k,0}(h_k) = 0$. Then $\forall l \in [1, M_k]$, we can obtain the optimal $Q_{k,l}^*$ as:*

$$Q_{k,l}^* = \left\{ \mathbf{h} : \ \tilde{\psi}_{k,l}(h_k) \leq \tilde{\psi}_{k,j}(h_k); \ \forall j \neq l, j \in [0, M_k] \right\}. \tag{42}$$

Moreover, λ_k^{s*} is determined by satisfying the rate condition

$$\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \bar{R}_k. \tag{43}$$

Proof: We can mimic the proof of Proposition 3 to derive (42). We next show that the optimal solution achieves the required rate with equality as in (43). To see this, suppose we have obtained a set of quantization regions over-satisfying the rate requirement for user k. Then we can simply reduce one region; e.g., the region for the lowest-rate AMC mode, such that the achieved average rate is reduced but still satisfies the requirement. It is easy to see that the new regions are certainly better than the previous ones since the average power for the lowest-rate AMC mode, and thus the total average power, decreases. Therefore, (43) must hold and this implies that $\lambda_k^{s*} > 0$ by the complementary slackness conditions. □

To obtain $\mathbf{Q}^*$, we only need $\boldsymbol{\lambda}^{s*} := [\lambda_1^{s*}, \ldots, \lambda_K^{s*}]^T$, which can be simply calculated by K one-dimensional searches. Once we have updated quantization regions, the next step is to update the time allocation policy $\tau(\mathbf{h})$. And this can be accomplished with Proposition 4 and Algorithm 1. With denoting J the global objective and using Propositions 1, 4 and 5, we propose the following S-JQRA algorithm:

Algorithm 3: *S0) initialization:* Produce initial time allocation $\boldsymbol{\tau}^{(0)}(\cdot)$ and quantization regions $\mathbf{Q}^{(0)}$ from Proposition 1 and Lemma 1, and then $\boldsymbol{\pi}^{(0)}$ using (40). Select tolerance $\varepsilon > 0$, initial objective $J^{(0)} = \infty$ and let the iteration index $t = 1$.

- *S1)* $\boldsymbol{\tau}^{(t-1)}(\cdot), \boldsymbol{\pi}^{(t-1)} \rightarrow \mathbf{Q}^{(t)}, \boldsymbol{\pi}^{(t)}$: Given $\boldsymbol{\tau}^{(t-1)}(\cdot)$ and $\boldsymbol{\pi}^{(t-1)}$, obtain $\mathbf{Q}^{(t)}$ from Proposition 6, and $\boldsymbol{\pi}^{(t)}$ as a function of $\boldsymbol{\tau}^{(t-1)}$ and $\mathbf{Q}^{(t)}$ using (40).
- *S2)* $\mathbf{Q}^{(t)}, \boldsymbol{\pi}^{(t)} \rightarrow \boldsymbol{\tau}^{(t)}(\cdot)$: Given $\mathbf{Q}^{(t)}$ and $\boldsymbol{\pi}^{(t)}$, obtain $\boldsymbol{\tau}^{(t)}(\cdot)$ from Proposition 4 and Algorithm 1.
- *S3) Stopping criterion:* Calculate the objective $J^{(t)}$ using $\mathbf{Q}^{(t)}$, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\cdot)$. If $|J^{(t-1)} - J^{(t)}|/J^{(t)} < \varepsilon$; return the current quantization and resource allocation and stop. Otherwise, increase t by 1 and go to *S1)*.

Since the global objective J is decreasing in each step of the iterations, as $t \rightarrow \infty$, the S-JQRA algorithm converges. Once the S-JQRA algorithm is run, the instantaneous allocation policy and feedback information specified in Proposition 5 still apply.

Although S-JQRA is sub-optimal, it has less computational complexity than JQRA. In *S1)* of S-JQRA, we use K *one*-dimensional searches to obtain $\mathbf{Q}^{(t)}$ and then compute the corresponding $\boldsymbol{\pi}^{(t)}$. This computation is much less than that for *J2)* of JQRA, where K *two*-dimensional searches are used to obtain $\mathbf{Q}^{(t)}$. While *S2)* of S-JQRA is exactly the same as *J3)* of JQRA, in each iteration S-JQRA saves the computation load for *J1)* of JQRA. Moreover, with two steps per iteration, S-JQRA could converge considerably faster than the three-step JQRA. Therefore, compared to JQRA, S-JQRA largely reduces the overall computation load.

VI. Numerical Results

In this section, we present numerical results of our joint quantization and resource allocation for a two-user Rayleigh flat-fading TDMA channel. The available system bandwidth is $B = 100$ KHz, and the AWGN has two-sided power spectral density N_0 Watts/Hz. Fading coefficients h_k, $k = 1, 2$, have mean $\bar{h}_k$ and are assumed independent. The average signal-to-noise ratio (SNR) for user k is $\bar{\gamma}_k = \bar{h}_k/(N_0 B)$. Unless otherwise specified, we assume that each user supports three M-ary quadrature amplitude modulation (QAM) modes: 2-QAM, 8-QAM and 32-QAM; i.e., the transmission rates of AMC modes are: $\rho_{k,l} = 1,\ 3,\ 5\ bits/symbol$. The corresponding BER can be approximated as [14]

$$\epsilon_{k,l}(\gamma) = 0.2e^{-\frac{\gamma}{2^{\rho_{k,l}}-1}}. \tag{44}$$

In all simulations, we assume the BER constraints are given by $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$.

Supposing P-CSI at transmitters (P-CSIT) or Q-CSIT and $\bar{\gamma}_k = 0\ dB$, $k = 1, 2$, we test the P-CSIT based resource allocation [8] and our Q-CSIT based JQRA and S-JQRA algorithms. For comparison, we also test a heuristic Q-CSIT based approach, where each user is assigned equal time fraction and transmits with equal power for all its AMC modes per block. With a fixed transmit-power, the access point selects for each user an AMC mode so that the instantaneous BER is less than or equal to the required level. Clearly, this quantization is a conservative one; i.e., except for the boundaries of the quantization regions, the BER requirement is always over-satisfied. With such a quantization, each user's transmit-power is then selected to ensure that its rate constraint is satisfied. Notice that due to its simplicity, the quantization in this heuristic scheme is actually widely employed in practical systems with adaptive transmissions; e.g., the CDMA2000 1xEVDO and WCDMA HSDPA. We first consider individual rate constraints: $\bar{R}_1 = 100\ kbps$ and $\bar{R}_2 = 100\ kbps$. With different power weights, Fig. 1 shows the weighted total power consumptions for these four approaches; while Fig. 2 depicts the performance loss of the three different Q-CSIT based approaches with respect to the P-CSIT solution to gauge the price paid for finite-rate feedback. We observe that: i) both JQRA and S-JQRA clearly outperform the heuristic Q-CSIT approach (yielding around 6 dB savings); and ii) the gap between JQRA and P-CSIT solution is very small. Since the P-CSIT solution lower bounds all Q-CSIT based approaches, this indicates that our coordinate descend algorithms are near-optimal.

The power region with P-CSIT has been derived in [8]. As stated in Section IV, we can use the solutions of (17) for all $\boldsymbol{\mu} \geq \mathbf{0}$ as the "boundary points" to approximately characterize the power region with Q-CSIT. We test two different sets of individual rate constraints: i) $\bar{R}_1 = 100\ kbps$, $\bar{R}_2 = 100\ kbps$, and ii) $\bar{R}_1 = 100\ kbps$, $\bar{R}_2 = 50\ kbps$. Fig. 3 depicts the power regions of the Rayleigh fading TDMA channels provided by the P-CSIT solution and the Q-CSIT based JQRA, where $\bar{P}_1$ and $\bar{P}_2$ represent the average transmit-power for the first and second user, respectively. With identical individual rate constraints, power regions are symmetric respect to the bisector $\bar{P}_1 = \bar{P}_2$; while they are not symmetric for $R_2 = R_1/2$. Clearly, the Q-CSIT regions are always contained in the P-CSIT regions. Notice that the Q-CSIT regions in Fig. 3 are actually conservative estimates. Even though it is seen that these Q-CSIT regions are very close to the P-CSIT regions. This again confirms the energy efficiency of the proposed JQRA algorithm. Numerical results describing the behavior of our algorithm in different cases are summarized in Table I. These show that the constraints are tightly satisfied and corroborate that our Q-CSIT based iterative coordinate descent algorithms achieve energy efficiency close to the optimal P-CSIT based one.

To gain more insight, let us take a closer look at our joint quantization and resource allocation solution when $\bar{R}_1 = \bar{R}_2 = 100\ kbps$ and $\mu_1/\mu_2 = 2$. Using JQRA, the power and rate loadings are listed in Table II, whereas the quantization regions and time allocation are depicted in Fig. 4, where different shades are used to represent which user is selected to access the channel. From Table II, we deduce that $\forall l_1 > l_2$, $\pi_{k,l_1} > \pi_{k,l_2}$. This indicates that for the simulated scenarios, the water-filling principles still hold in the Q-CSIT based optimal power loading, as in the P-CSIT case; i.e., when the

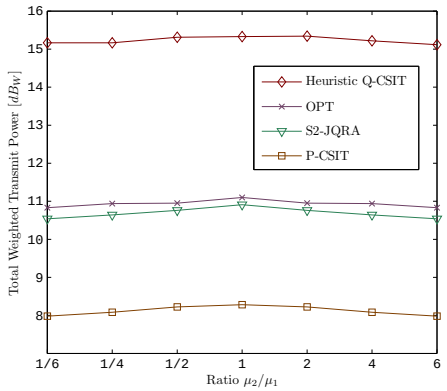

Fig. 1. Total power consumption for different resource allocation approaches with different power weight ratio μ_2/μ_1 and $\sum_{k=1}^{2}\mu_k = 1$ when $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100\ kbps$, and $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$.

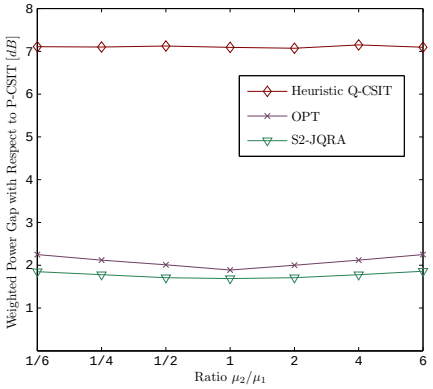

Fig. 2. Performance loss of Q-CSIT based approaches with respect to the P-CSIT solution when $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100\ kbps$, and $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$.

TABLE I

JQRA PERFORMANCE FOR DIFFERENT TEST CASES (FOR CASES II AND IV THE RATE REQUIREMENT FOR USER 2 IS 50 $kbps$).

Test Case	User (k)	μ_k	$\bar{\gamma}_k[dB]$	$\bar{R}_k[kbps]$	$\bar{\epsilon}_k$	$P_k[dBw]$	P-CSIT $P_k[dBw]$
I	1	1	0	99	10^{-3}	8.75	8.21
	2	1	0	100	10^{-3}	8.80	8.21
II	1	1	0	100	10^{-3}	8.13	7.75
	2	1	0	50	10^{-3}	4.15	3.80
III	1	1	3	100	10^{-3}	6.60	6.03
	2	1	0	100	10^{-3}	8.58	8.03
IV	1	1	3	100	10^{-3}	6.64	5.93
	2	1	0	50	10^{-3}	3.52	3.31
V	1	4/3	0	99	10^{-3}	8.46	7.88
	2	2/3	0	99	10^{-3}	9.07	8.32

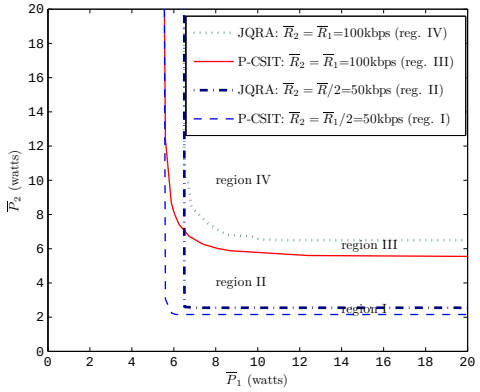

Fig. 3. Power regions for P-CSIT and JQRA policies ($\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$).

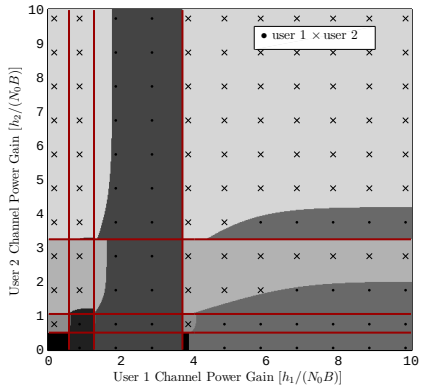

Fig. 4. Optimal time allocation policy and quantization regions obtained by the JQRA algorithm, where user selections are indicated using different shades and quantization thresholds are represented with bold lines ($\mu_1 = 2/3$, $\mu_2 = 1/3$, $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100\ kpbs$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$).

TABLE II

POWER ($\pi_{k,l}$) AND RATE ($\rho_{k,l}$) LOADING PER QUANTIZATION STATE (AMC MODE) RETURNED BY JQRA ($\mu_1 = 2/3$, $\mu_2 = 1/3$, $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100\ kpbs$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$).

	User 1			**User 2**		
Quantization Region	$Q_{1,1}$	$Q_{1,2}$	$Q_{1,3}$	$Q_{2,1}$	$Q_{2,2}$	$Q_{2,3}$
Tx-power $[dBw]$	8.56	13.23	15.60	8.99	13.84	16.29
Rate $[bits/sym]$	1	3	5	1	3	5

TABLE III

AVERAGE WEIGHTED POWER FOR JQRA WITH DIFFERENT NUMBER OF FEEDBACK BITS ($\mu_1 = \mu_2 = 1$, $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100\ kpbs$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$).

Algorithm	JQRA	JQRA	JQRA	JQRA	P-CSIT
# of bits	1	2	3	4	∞
Average Power $[dBw]$	23.05	11.98	8.52	8.43	8.10

channel is better, we use a higher rate with more transmit-power. As shown in Fig. 4, simulation results reveal optimal quantization regions $\{Q^*_{k,l}\}_{l=1}^3$, $k = 1, 2$, are non-overlapping consecutive intervals which can be determined by a set of thresholds $\{q^*_{k,l}\}$, which are represented with bold lines. Fig. 4 reveals that the boundaries between different users' allocation follow the quantization threshold values. This implies that a simple quantization-region based time allocation approach may provide a good approximation to the optimal policy, and motivates future research in developing simplified (or even distributed) time-allocation approaches with Q-CSIT.

We have seen that with three AMC modes, the JQRA and S-JQRA algorithms provide energy efficiency close to the optimal P-CSIT solution. To achieve this, we need $\lceil \log_2(3+3+1) \rceil = 3$ bits for the Q-CSI per time block [c.f. Proposition 5]. We next show how the number of feedback bits affects the performance of JQRA. When $\bar{R}_1 = \bar{R}_2 = 100\ kbps$ and $\mu_1/\mu_2 = 1$, Table III shows the total power cost for the two-user Rayleigh flat-fading TDMA channel with different number of feedback bits. When only one bit is available, the feedback information only indicates user selection, and once the user is picked, it transmits regardless of $\mathbf{h}$. It is shown in Fig. 4 that for some $\mathbf{h}$ where channel is in deep fade, we should let both users defer. Even though the region for this case is small, much power (23.05 dBw) is required to compensate these "bad" channels. As the number of feedback bits increases, the number of active AMC modes per user increases. It is seen from Table III that surprisingly, even for two feedback bits case where one user supports two modes and the other supports only one mode, JQRA provides a power cost not far away from the P-CSIT solution. This reveals that the time allocation policy plays an important role in energy efficiency. Numerical results also reveal that a few (2-4) AMC modes per user, and thus a few feedback bits suffice to close the gap between Q-CSIT and P-CSIT.

The convergence of JQRA and S-JQRA is illustrated in

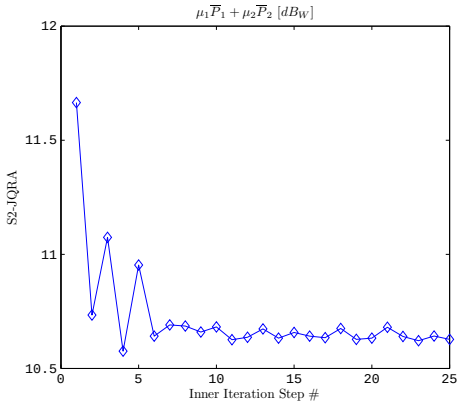

Fig. 5. Average weighted power evolution for JQRA and S-JQRA algorithms ($\mu_1 = 2/3$, $\mu_2 = 1/3$, $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100$ *kpbs*, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ *dB*).

Fig. 5, where the average total weighted power evolves with the inner iteration steps. We can see that JQRA converges after a small number of iterations (around 24 inner steps or 8 outer iterations). The small variations through the curve are due to the finite resolution in the numerical integrations involved. Another interesting observation is that even the first inner iteration step provides a good solution. Convergence of S-JQRA is even faster but also rougher, partly due to the fact that each iteration of S-JQRA entails only 2 inner iteration steps. And the converged solution of S-JQRA is around 1dB away from the JQRA solution. Fig. 5 clearly demonstrates the fast convergence of our JQRA and S-JQRA algorithms.

VII. Conclusions

With finite-rate feedback from the access point, users can only acquire Q-CSI and thus adopt a finite number of resource allocation configurations. Based on Q-CSI, we derived two energy-efficient joint quantization and resource allocation strategies for TDMA fading channels. The resulting JQRA and S-JQRA algorithms decouple the complex optimization task into three or two tractable minimization sub-problems. We relied on coordinate descent principles to derive iterative algorithms which assemble the different sub-solutions of the decoupled sub-problems to solve the main problem. Numerical results showed that with Q-CSIT only available, both JQRA and S-JQRA algorithms achieve energy efficiency surprisingly close to that obtained with P-CSIT, and yield large energy-savings compared to a heuristic Q-CSIT approach. While JQRA yields the optimal CSI quantizer and time-power allocation, the suboptimal S-JQRA has reduced computational complexity which can be attractive in practice.[2]

VIII. Appendix: Proof of Proposition 4

Let us define $Q_{k,0}$ $\forall k \in [1,K]$ as the set complement of $\bigcup_{l\in[1,M_k]} Q_{k,l}$, and set $\tilde{\varphi}_k(\mathbf{h})$ to a large number when $\mathbf{h} \in Q_{k,0}$ such that user k will not be selected. Then $\forall \boldsymbol{\tau} \neq \boldsymbol{\tau}^*$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\left[\lambda_k^{\tau^*}\left(\sum_{l=1}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h})-\bar{R}_k\right)\right.\\
&\qquad\left.+\nu_k^{\tau^*}\left(\bar{\epsilon}_k-\sum_{l=1}^{M_k}\frac{\rho_{k,l}}{\bar{R}_k}\int_{Q_{k,l}}\tau_k(\mathbf{h})\epsilon_{k,l}(h_k\pi_{k,l})dF(\mathbf{h})\right)\right]\\
&\overset{(a)}{=}\sum_{k=1}^{K}\lambda_k^{\tau^*}\left(\sum_{l=0}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h})\right.\\
&\qquad\left.-\sum_{l=0}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k^*(\mathbf{h})dF(\mathbf{h})\right)\\
&\qquad+\frac{\nu_k^{\tau^*}}{\bar{R}_k}\left(\sum_{l=0}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k^*(\mathbf{h})\epsilon_{k,l}(h_k\pi_{k,l})dF(\mathbf{h})\right.\\
&\qquad\left.-\sum_{l=0}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})\epsilon_{k,l}(h_k\pi_{k,l})dF(\mathbf{h})\right)\\
&=\sum_{k=1}^{K}\sum_{l=0}^{M_k}\int_{Q_{k,l}}(\tau_k(\mathbf{h})-\tau_k^*(\mathbf{h}))\times\\
&\qquad\left(\lambda_k^{\tau^*}\rho_{k,l}-\nu_k^{\tau^*}\frac{\rho_{k,l}}{\bar{R}_k}\epsilon_{k,l}(h_k\pi_{k,l})\right)dF(\mathbf{h})\\
&\overset{(b)}{=}\sum_{k=1}^{K}\sum_{l=0}^{M_k}\int_{Q_{k,l}}(\tau_k(\mathbf{h})-\tau_k^*(\mathbf{h}))(\mu_k\pi_{k,l}-\tilde{\varphi}_k(\mathbf{h}))dF(\mathbf{h})\\
&=\sum_{k=1}^{K}\mu_k\sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h})\\
&\qquad-\sum_{k=1}^{K}\mu_k\sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k^*(\mathbf{h})dF(\mathbf{h})\\
&\qquad-\sum_{k=1}^{K}\sum_{l=0}^{M_k}\int_{Q_{k,l}}(\tau_k(\mathbf{h})-\tau_k^*(\mathbf{h}))\tilde{\varphi}_k(\mathbf{h})dF(\mathbf{h});
\end{aligned}\tag{45}$$

where equality (a) is due to the complementary slackness conditions for rate and BER constraints, and equality (b) is due to the definition of $\tilde{\varphi}_k(\mathbf{h})$. Let

$$\begin{aligned}
C&:=\sum_{k=1}^{K}\sum_{l=0}^{M_k}\int_{Q_{k,l}}(\tau_k(\mathbf{h})-\tau_k^*(\mathbf{h}))\tilde{\varphi}_k(\mathbf{h})dF(\mathbf{h})\\
&=\int_{\mathbf{h}}\left(\sum_{k=1}^{K}\tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h})-\sum_{k=1}^{K}\tau_k^*(\mathbf{h})\tilde{\varphi}_k(\mathbf{h})\right)dF(\mathbf{h})
\end{aligned}\tag{46}$$

Since the definition of $\tilde{\varphi}_k(\mathbf{h})$ already accounts for quantization, (46) follows readily. Now $\forall \mathbf{h}$,

1) If $\forall k \in [1,K]$, $\tilde{\varphi}_k(\mathbf{h}) \geq 0$, then

$$\begin{aligned}
&\sum_{k=1}^{K}\tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h})-\sum_{k=1}^{K}\tau_k^*(\mathbf{h})\tilde{\varphi}_k(\mathbf{h})\\
&\qquad=\sum_{k=1}^{K}\tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h})\geq 0.
\end{aligned}\tag{47}$$

[2]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.

2) If functions $\{\tilde{\varphi}_k(\mathbf{h})\}_{k=1}^K$ have a single minimum $\tilde{\varphi}_i(\mathbf{h}) < 0$, then

$$\begin{aligned}\sum_{k=1}^K \tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) - \sum_{k=1}^K \tau_k^*(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) \\ = \sum_{k=1}^K \tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) - \tilde{\varphi}_i(\mathbf{h}) \\ \geq \left(\sum_{k=1}^K \tau_k(\mathbf{h}) - 1\right)\tilde{\varphi}_i(\mathbf{h}) \geq 0. \end{aligned} \quad (48)$$

3) If functions $\{\tilde{\varphi}_k(\mathbf{h})\}_{k=1}^K$ have multiple minima $\left\{\tilde{\varphi}_{i_j}(\mathbf{h})\right\}_{j=1}^J < 0$, then letting $\tilde{\varphi}_m$ denote this minimum value, we have

$$\begin{aligned}\sum_{k=1}^K \tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) - \sum_{k=1}^K \tau_k^*(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) \\ = \sum_{k=1}^K \tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) - \sum_{j=1}^J \tau_j^*\tilde{\varphi}_m \\ \geq \left(\sum_{k=1}^K \tau_k(\mathbf{h}) - 1\right)\tilde{\varphi}_m \geq 0. \end{aligned} \quad (49)$$

Hence, we have $\sum_{k=1}^K \tau_k(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) - \sum_{k=1}^K \tau_k^*(\mathbf{h})\tilde{\varphi}_k(\mathbf{h}) \geq 0$, $\forall \mathbf{h}$, and thus $C \geq 0$. Then from (45),

$$\begin{aligned}\sum_{k=1}^K \Bigg[\lambda_k^{\tau*}\left(\sum_{l=1}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h}) - \bar{R}_k\right) \\ +\nu_k^{\tau*}\left(\bar{\epsilon}_k - \sum_{l=1}^{M_k}\frac{\rho_{k,l}}{\bar{R}_k}\int_{Q_{k,l}}\tau_k(\mathbf{h})\epsilon_{k,l}(h_k\pi_{k,l})dF(\mathbf{h})\right)\Bigg] \\ \leq \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h}) \\ -\sum_{k=1}^K \mu_k \sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k^*(\mathbf{h})dF(\mathbf{h}). \end{aligned} \quad (50)$$

Therefore, $\forall \boldsymbol{\tau} \neq \boldsymbol{\tau}^*$, if $\boldsymbol{\tau}$ satisfies the individual rate and BER constraints

$$\sum_{l=1}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h}) \geq \bar{R}_k,$$

$$\sum_{l=1}^{M_k}\rho_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})\epsilon_{k,l}(h_k\pi_{k,l})dF(\mathbf{h})/\bar{R}_k \leq \bar{\epsilon}_k,$$

we have $\sum_{k=1}^K \mu_k \sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k(\mathbf{h})dF(\mathbf{h}) \geq \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k}\pi_{k,l}\int_{Q_{k,l}}\tau_k^*(\mathbf{h})dF(\mathbf{h})$, which implies the optimality of $\boldsymbol{\tau}^*$.

REFERENCES

[1] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. on Networking*, vol. 10, no. 4, pp. 487-499, Aug. 2002.

[2] M. A. Khojastepour and A. Sabharwal, "Delay-constrained scheduling: Power efficiency, filter design, and bounds," *Proc. of INFOCOM Conf.*, vol. 3, pp. 1938-1949, Hong Kong, China, March 7-11, 2004.

[3] M. Zafer and E. Modiano, "A calculus approach to minimum energy transmission policies with quality of service guarantees," *Proc. of INFOCOM Conf.*, vol. 1, pp. 548-559, Miami, FL, March 13-17, 2005.

[4] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal energy allocation for delay-constrained data transmission over a time-varying channel," *Proc. of INFOCOM Conf.*, vol. 2, pp. 1095-1105, San Francisco, CA, March 3 - April 4, 2003.

[5] Y. Yao and G. B. Giannakis, "Energy-efficient scheduling for wireless sensor networks," *IEEE Trans. on Commun.*, vol. 53, no. 8, pp. 1333-1342, August 2005.

[6] D. Tse and S. V. Hanly, "Multiaccess fading channels–Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. on Inf. Theory*, vol. 44, No.7, pp. 2796-2815, Nov. 1998.

[7] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels–Part I and Part II" *IEEE Trans. on Inf. Theory*, vol. 47, No.3, pp. 1083-1102 and pp. 1103-1127, March 2001.

[8] X. Wang and G. B. Giannakis, "Energy-efficient resource allocation in TDMA over fading channels," *Proc. of the Intl. Symp. on Info. Theory*, Seattle, Washington, July 9-14, 2006, available at http://spincom.ece.umn.edu/

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

[10] M. Medard, "The effect upon channel capacity in wireless communnications of perfect and imperfect knowledge of the channel," *IEEE Trans. on Inf. Theory*, vol. 46, no. 3, pp. 933-946, May 2000.

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

[12] A. G. Marques, F. F. Digham, and G. B. Giannakis, "Power-efficient OFDM via quantized channel state information," *Proc. of Intl. Conf. on Commun.*, Istanbul, Turkey, June 11-15, 2006, available at http://spincom.ece.umn.edu/.

[13] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*, Cambridge Uiversity Press, 2005.

[14] A. J. Goldsmith and S. G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. on Commun.*, vol. 46, pp. 595602, May 1998.

[15] IEEE 802.16 WG, *Air interface for fixed broadband wireless access systems*, IEEE Std. 802.16, April. 2002.

[16] D. Bertsekas, *Nonlinear Programming: 2nd Ed.*, Athena Scientific, 1999.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Uiversity Press, 2004.

[18] S. Lloyd, "Least-squares quantization in PCM," *IEEE Trans. on Inf. Theory*, vol. 28, no. 2, pp. 129-137, Mar. 1982.

[19] P. Xia, S. Zhou, and G. B. Giannakis, "Multiantenna adaptive modulation with beamforming based on bandwidth-constrained feedback," *IEEE Trans. on Commun.*, vol. 53, no. 3, pp. 526-535, Mar. 2005.

[20] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Norwell, MA: Kluwer, 1992.