OPTIMIZING ENERGY EFFICIENCY OF TDMA WITH FINITE RATE FEEDBACK*

Antonio G. Marques

Dept. of Signal Theory and Communications
Rey Juan Carlos University, Madrid, Spain

Xin Wang and Georgios B. Giannakis

Dept. of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN, USA

ABSTRACT

We deal with energy efficient time-division multiple access over fading channels with finite-rate feedback for use in the power-limited regime. Through FRF 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 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. Numerical results are presented to evaluate the energy savings.

***Index Terms*—** Quantization, Optimization methods, Multiuser channels, Minimum energy control, Resource management

1. INTRODUCTION

Resource allocation for fading channels has been studied in [1, 2] and energy-efficiency policies for time-division multi-access (TDMA) have been investigated from an information theoretic perspective in [3]. Assuming that both transmitters and receivers have available perfect (P-) channel state information (CSI), the approaches in [3] provide fundamental power limits when each user can support capacity-achieving codebooks, and also yield guidelines for practical designs where users can only support a finite number of adaptive modulation and coding (AMC) modes with prescribed bit error probabilities (BER). While the assumption of P-CSI renders analysis and design tractable, it may not be always realistic. It then motivates a *finite-rate* feedback (FRF) model, where only *quantized* (Q-) CSI is available at the transmitter through a finite number of bits of feedback from the receiver. Based on the FRF, [4] minimized transmit-power of orthogonal frequency-division multiplexing (OFDM) systems. In this paper, we consider energy efficiency issues for TDMA over fading channels with FRF. 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, the goal is to map channel quantization states to AMC modes and allocate optimally time slots to users so that *transmit*-power is minimized. To tackle it, 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 and then solve each of them in an optimal way; i.e., we resort to a coordinate descent [5] approach to come up with an iterative algorithm which assembles the different sub-solutions to solve the main problem.

*The work in this paper was supported by the USDoD ARO grant No. W911NF-05-1-0283 and by the C. A. Madrid grant No. P-TIC-000223-0505.

2. MODELING PRELIMINARIES

Consider K users linked wirelessly to an access point (AP). The input-output relationship is $y(n) = \sum_{k=1}^{K} \sqrt{h_k(n)}x_k(n) + z(n)$ 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 $\sigma^2 = 1$. We confine ourselves to TDMA; i.e., when $x_k(n) \neq 0$, 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 [1, 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 AP 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. Having perfect knowledge of $\{h_k\}_{k=1}^{K}$, the AP 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. Users then transmit with the indicated AMC modes at the assigned time fractions. FRF from the AP to users consists of a few bits indexing predetermined AMC modes and time slots.[1]

3. QUANTIZATION AND RESOURCE ALLOCATION WITH FINITE RATE FEEDBACK

We wish to minimize total power under individual average rate constraints in a TDMA system. Given a time allocation policy $\boldsymbol{\tau}(\cdot)$, let $\tau_k(\mathbf{h})$ denote the time *fraction* allocated to user k if fading $\mathbf{h}$ occurs. With FRF from the AP, 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

[1] Notation: ^{T} denotes transposition, $\lceil x \rceil$ the minimum integer $\geq x$, and $[x]^+ := \max(x, 0)$. 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, and $F(\mathbf{h})$ their joint cumulative distribution function (cdf).

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 $\rho_{k,l}$ of the AMC mode when $\mathbf{h} \in Q_{k,l}$. Note that with Q-CSI, user k is only allowed to use a *fixed transmit* power $\pi_{k,l}$ for its lth mode. We need to optimize $\pi_{k,l}$ in our FRF 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 $\boldsymbol{\tau}(\cdot)$. 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.

With $\bar{R}_k$ and $\bar{\epsilon}_k$ collecting the prescribed rate and BER requirements, power weights μ_k, and using the previous definitions, the energy-efficient quantization and resource allocation problem is

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

where the left-hand side of second and third constraint represents the average rate and BER per user. The problem (1) is 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 [5] 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 [6].

3.1. Initialization

We first use the resource allocation policies of [3] to initialize our coordinate descent method. Given AMC modes and P-CSI, [3, Theorem 6] yields the energy-efficient rate-power and time allocation policies $\boldsymbol{\tau}^*(\cdot)$ 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[2]:

Proposition 1 *With optimum rate allocation, the optimal region $Q_{k,l}^*$ for user $k \in [1, K]$ is given by $Q_{k,l}^* = \{\mathbf{h} :\ h_k \in [q_{k,l}^*, q_{k,l+1}^*)\}$, where $q_{k,l}^* = (p_{k,l} - p_{k,l-1})/(\rho_{k,l} - \rho_{k,l-1})\mu_k/\lambda_k^{P*}$ for $l \in [1, M_k]$ and $q_{k,M_k+1}^* = \infty$.* □

3.2. Optimal Transmit-Powers

It is clear from (1) that the rate constraints affect to $\boldsymbol{\tau}(\cdot)$ and $\mathbf{Q}$. In each iteration of our coordinate descent algorithm, we will descend the global objective within the feasible set. This guarantees that in this step we always start with a pair of $\mathbf{Q}$ and $\boldsymbol{\tau}(\cdot)$ already satisfying rate constraints to find the optimal $\boldsymbol{\pi}$. Therefore, given these $\mathbf{Q}$ and $\boldsymbol{\tau}(\cdot)$, finding the optimal $\boldsymbol{\pi}$ reduces to solve

$$\begin{cases} \min_{\boldsymbol{\pi}} \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall k,\ \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \le \bar{\epsilon}_k. \end{cases} \tag{2}$$

Let us define $A_{k,l} := \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})$. Since the functions $\epsilon_{k,l}(x)$ are convex, (2) is a convex optimization problem. Its solution can be analytically obtained as follows.

Proposition 2 *Given positives $\nu_k^{\pi*}, \forall k$, 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 $\int_{Q_{k,l}} \tau_k(\mathbf{h}) h_k \epsilon_{k,l}'(h_k \pi_{k,l}^*) dF(\mathbf{h}) = -\frac{\mu_k \bar{R}_k A_{k,l}}{\rho_{k,l} \nu_k^{\pi*}}$, or $\pi_{k,l}^* = 0$. And $\forall k \in [1, K]$, each Lagrange multiplier $\nu_k^{\pi*}$ is determined by satisfying the constraint $\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) / \bar{R}_k = \bar{\epsilon}_k$.* □

Notice that given $\tau_k(\mathbf{h})$, users are decoupled. Solving (2) is equivalent to solving K small problems. Given $\nu_k^{\pi*}$ and monotonically decreasing $\epsilon_{k,l}(\gamma)$, the solution to first equation of Proposition 2 is unique for $\pi_{k,l}^* > 0$ and we can use a one-dimensional. Then we can use another one-dimensional search to solve for $\nu_k^{\pi*}$ in the BER constraint. And the optimal transmit-powers $\boldsymbol{\pi}^*$ are in turn obtained.

3.3. 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 $\forall k$,

$$\begin{cases} \min_{\{Q_{k,l}\}_{k=1}^{M_k}} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \ge \bar{R}_k; \\ \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \le \bar{\epsilon}_k. \end{cases} \tag{3}$$

Similar to a constrained vector quantization [6], we derive:

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: $\forall l \in [1, M_k]$, $Q_{k,l}^* = \{\mathbf{h} :\ \psi_{k,l}(h_k) \le \psi_{k,j}(h_k);\ \forall j \ne l, j \in [0, M_k]\}$. Moreover, λ_k^{q*} and ν_k^{q*} are determined by satisfying slackness conditions $\lambda_k^{q*} \times (\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) dF(\mathbf{h}) - \bar{R}_k) = 0$ and $\nu_k^{q*} \times (\sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) - \bar{\epsilon}_k) = 0$.* □

To obtain the optimal $Q_{k,l}^*$, we need to find λ_k^{q*} and ν_k^{q*}. Since (3) 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 the constraints, we stop the search and return these values. After obtaining λ_k^{q*} and ν_k^{q*}, $\forall k$ (using K two-dimensional searches), we in turn determine $\mathbf{Q}^*$.

3.4. Optimal Time Allocation

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

$$\begin{cases} \min_{\boldsymbol{\tau}(\cdot)} \sum_{k=1}^{K} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall \mathbf{h},\ \sum_{k=1}^{K} \tau_k(\mathbf{h}) \le 1; \\ \quad \forall k,\ \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \ge \bar{R}_k; \\ \quad \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \le \bar{\epsilon}_k. \end{cases} \tag{4}$$

[2]Proofs for all Propositions can be found in [7].

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 (4) can be obtained as follows:*
1. *$\forall k \in [1, K]$, $\tilde{\varphi}_k(\mathbf{h}) \geq 0$, 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 $\{\tilde{\varphi}_{i_j}(\mathbf{h})\}_{j=1}^J < 0$, then $\tau_{i_j}^*(\mathbf{h}) = \tau_j^*$ with any $\sum_{j=1}^J \tau_j^* = 1$, and $\forall k \neq i_j$, $k \in [1, K]$, $\tau_k^*(\mathbf{h}) = 0$.*
Moreover, $\lambda_k^{\tau*}$ and $\nu_k^{\tau*}$ should satisfy the complementary slackness conditions $\forall k \in [1, K]$ similar to λ_k^{q*} and ν_k^{q*} in Proposition 3. □

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 cost 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. For cases where $\tilde{\varphi}_k(\mathbf{h}) \geq 0\ \forall k \in [1, K]$, we should let all users to defer. 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 (4) is given by

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

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 (4) is

$$\max_{\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau} g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau), \qquad \text{s.t. } \boldsymbol{\lambda}^\tau \geq \mathbf{0}, \boldsymbol{\nu}^\tau \geq \mathbf{0}. \tag{5}$$

Since (4) 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. As our problem is convex, the convergence of our sub-gradient projection algorithm is guaranteed [5]. Once $\boldsymbol{\lambda}^{\tau*}$ and $\boldsymbol{\nu}^{\tau*}$ are calculated, the time allocation policy in Proposition 4 is in turn determined.

Sub-gradient Algorithm: ***[T0]*** Generate an arbitrary positive 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] Numerically evaluate the partial derivatives $\Delta\lambda_k^\tau := \frac{\partial g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)}{\partial \lambda_k^\tau}$ and $\Delta\nu_k^\tau := \frac{\partial g(\boldsymbol{\lambda}^\tau, \boldsymbol{\nu}^\tau)}{\partial \nu_k^\tau}$ $\forall k$ at $(\boldsymbol{\lambda}^\tau(t-1), \boldsymbol{\nu}^\tau(t-1))$. Choose a step size δ 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] 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)*.

3.5. 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:

JQRA Algorithm: ***[J0]*** Produce initial $\boldsymbol{\tau}^{(0)}(\cdot)$ and $\mathbf{Q}^{(0)}$ from [3, Theorem 6] and Proposition 1. Select tolerance $\varepsilon > 0$, initialize objective at $J^{(0)} = \infty$ and set the iteration index $t = 1$.
[J1] $\boldsymbol{\tau}^{(t-1)}(\cdot), \mathbf{Q}^{(t-1)} \to \boldsymbol{\pi}^{(t)}$: Given $\boldsymbol{\tau}^{(t-1)}(\cdot)$ and $\mathbf{Q}^{(t-1)}$, obtain $\boldsymbol{\pi}^{(t)}$ from Proposition 2.
[J2] $\boldsymbol{\pi}^{(t)}, \boldsymbol{\tau}^{(t-1)}(\cdot) \to \mathbf{Q}^{(t)}$: Given $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t-1)}(\cdot)$, obtain $\mathbf{Q}^{(t)}$ from Proposition 3.
[J3] $\mathbf{Q}^{(t)}, \boldsymbol{\pi}^{(t)} \to \boldsymbol{\tau}^{(t)}(\cdot)$: Given $\mathbf{Q}^{(t)}$ and $\boldsymbol{\pi}^{(t)}$, obtain $\boldsymbol{\tau}^{(t)}(\cdot)$ from Proposition 4.
[J4] *Stopping criterion:* Calculate $J^{(t)}$ using $\mathbf{Q}^{(t)}$, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\cdot)$. If $(J^{(t-1)} - J^{(t)})/J^{(t)} < \varepsilon$, return $\mathbf{Q}^{(t)}$, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\cdot)$ and stop. Otherwise, $t = t + 1$ and go to *J1)*.

Since the global objective J is decreasing in each step, it is easy to see that as $t \to \infty$, the JQRA algorithm converges.

3.6. Optimal Feedback Bits

JQRA provides a quantizer design which is computed off-line. After that, the AP 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 $\mathbf{Q}^*$, $\boldsymbol{\pi}^*$, $\boldsymbol{\lambda}^{\tau*}$ and $\boldsymbol{\nu}^{\tau*}$ from JQRA, $\forall \mathbf{h}$, the AP 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) $k^*(\mathbf{h}) = \arg_k \min\{\tilde{\varphi}_k(\mathbf{h}, \mathbf{Q}^*, \boldsymbol{\pi}^*, \boldsymbol{\lambda}^{\tau*}, \boldsymbol{\nu}^{\tau*})\}_{k=1}^K$ (pick any k^* if 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) $l^*(\mathbf{h}) = \{\, l;\ s.t.\ \mathbf{h} \in Q_{k^*(\mathbf{h}),l},\ \ l = 1, \ldots, M_k \}$. When the users receive the broadcasted $c^*(\mathbf{h})$, the optimal multiple access 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.* □

This 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 the different user-AMC-mode combinations and the case of all users deferring. Consider a system with 85-170 users, each supporting six different AMC modes. 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.

4. NUMERICAL RESULTS

In this section, we present numerical results of JQRA for a 2-user Rayleigh flat-fading TDMA channel. The system bandwidth is $B =$

Table 1. Power weighted cost (measured in dBw) $\overline{P}_{HEUR}$ for Q-CSIT heuristic approach, $\overline{P}_{JQRA}$ for Q-CSIT JQRA, and $\overline{P}_{PCSIT}$ for P-CSIT solution [3] in different test cases. (Reference case: $\mu_1 = 1/2$, $\mu_2 = 1/2$, $\bar{R}_1 = \bar{R}_2 = 100\ kpbs$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0\ dB$. In other cases, indicated changes are made while other parameters remaining the same as reference case.)

Variation	$\overline{P}_{HEUR}$	$\overline{P}_{JQRA}$	$\overline{P}_{PCSIT}$
Reference Case	15.23	8.79	8.21
$\mu_1 = 2/3,\ \mu_2 = 1/3$	15.22	8.76	8.03
$\mu_1 = 6/7,\ \mu_2 = 1/7$	15.08	8.51	7.98
$\bar{\gamma}_1 = 3,\ \bar{\gamma}_2 = 0$	14.83	7.71	7.15
$\bar{R}_1 = 100,\ \bar{R}_2 = 50$	13.01	6.59	6.22

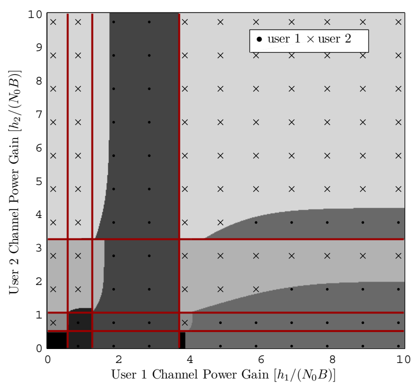

Fig. 1. Optimal time allocation policy and quantization regions obtained by the JQRA algorithm. Regions are indicated using different shades and quantization thresholds are represented with bold lines ($\mu_1 = 2/3$, $\mu_2 = 1/3$, $\bar{R}_1 = \bar{R}_2 = 100\ kpbs$, $\bar{\gamma}_1 = \bar{\gamma}_2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

100 KHz, and the AWGN has two-sided power spectral density N_0 Watts/Hz. Fading coefficients h_k, have mean $\bar{h}_k$ and are assumed independent. The average signal-to-noise ratio (SNR) is $\bar{\gamma}_k = \frac{\bar{h}_k}{(N_0B)}$. The transmission rates per symbol of AMC modes are: $\rho_{k,l} = 1,\ 3,\ 5$ $bits/sym$. The corresponding BER can be approximated as $\epsilon_{k,l}(\gamma)$ $= 0.2\exp(-\gamma/(2^{\rho_{k,l}} - 1)$; and we set $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$.

Supposing P-CSI at transmitters (P-CSIT) or Q-CSIT, we test the P-CSIT based resource allocation [3] and our Q-CSIT based JQRA. For comparison, we also test a widely employed 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. The AP selects for each user an AMC mode so that the instantaneous BER is less than or equal to the required level. With such a quantization, each user's transmit-power is then selected to ensure that its rate constraint is satisfied. Numerical results describing the behavior of our algorithm in different cases are summarized in Table I. We observe that: i) JQRA clearly outperforms the heuristic Q-CSIT approach (yielding around 6 dB savings); 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.

To gain more insight, let us take a closer look at our joint quantization and resource allocation solution when $\mu_1/\mu_2 = 2$. For this case the optimum powers (measured in dBw) are: $\pi^*_{1,1} = 8.6$, $\pi^*_{1,2} = 13.2$, $\pi^*_{1,3} = 15.6$, $\pi^*_{2,1} = 9.0$, $\pi^*_{2,2} = 13.8$, and $\pi_{2,3} = 16.3$. 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 channel is better, we use a higher rate with more transmit-power. The quantization regions and time allocation are depicted in Fig. 1 that reveals optimal quantization regions $\{\{Q^*_{k,l}\}^3_{l=1}\}^2_{k=1}$ are non-overlapping consecutive intervals which can be determined by a set of thresholds $\{q^*_{k,l}\}$, which are represented with bold lines. This implies that a simple quantization-region based time allocation approach may provide a good approximation to the optimal policy. Numerical results also reveal that 5-10 outer iterations of JQRA suffice to converge to the optimal solution.

5. CONCLUSIONS

Based on Q-CSI, we derived an energy-efficient JQRA strategy for TDMA fading channels which relies on coordinate descent principles to assemble the different sub-solutions of the decoupled subproblems to solve the main problem. Numerical results showed that with Q-CSIT only available, our JQRA algorithm achieve energy efficiency surprisingly close to that obtained with P-CSIT, and yield large energy-savings compared to a widely used Q-CSIT approach.

6. REFERENCES

[1] 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.

[2] L. Li and A. J. Goldsmith, "Capacity and Optimal Resource Allocation for Fading Broadcast Channels–Part I: Ergodic Capacity" *IEEE Trans. on Inf. Theory*, vol. 47, No.3, pp. 1083-1102, March 2001.

[3] 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.

[4] A. G. Marques, F. F. Digham, and G. B. Giannakis, "Power-Efficient OFDM via Quantized Channel State Information," *IEEE Journal on Selected Areas in Commun.*, vol. 57, No.5, pp. 522-544, July 2006.

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

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

[7] X. Wang, A. G. Marques, and G. B. Giannakis, "Power-Efficient Resource Allocation and Quantization for TDMA Using Adaptive Transmission and Limited-Rate Feedback," *IEEE Trans. on Signal Process.*, submitted, May 2006, available at http://spincom.ece.umn.edu/