

ENERGY-EFFICIENT TDMA WITH QUANTIZED CHANNEL STATE INFORMATION*

Antonio G. Marques

Department of Signal Theory and Communications

Rey Juan Carlos University, Madrid, Spain

Xin Wang and Georgios B. Giannakis (contact author)

Department of Electrical and Computer Engineering

University of Minnesota, Minneapolis, MN, USA

ABSTRACT

We deal with energy efficient time-division multiple access (TDMA) over fading channels with finite-rate feedback 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 (AMC) modes and allocate optimally time slots to users so that transmit-power is minimized. To this end, we develop two joint quantization and resource allocation approaches. In the first one, we rely on the quantization regions associated to each AMC mode and the time allocation policy inherited from the perfect CSI case to optimize the fixed transmit-power across quantization states. In the second approach, we pursue separable optimization and resort to coordinate descent algorithms to solve the following two sub-problems: (a) given a time allocation, we optimize the quantization regions and transmit-powers; and (b) with improved quantization regions, we optimize the time allocation policy. Numerical results are present to evaluate the energy savings and compare the novel approaches.

1. INTRODUCTION

Recently energy-efficient resource allocation has attracted growing attention [1, 2, 3]. Resource allocation for fading channels has been studied in [4, 5] and energy-efficiency policies for TDMA have been investigated from an information theoretic perspective in [6]. Assuming that both transmitters and receivers have available perfect (P-) channel state information (CSI), the approaches in [6] 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. It then motivates a *finite-rate* feedback model, where only *quantized* (Q-)

*Work in this paper was supported by the ARO Grant No. W911NF-05-1-0283 and was prepared through collaborative participation in the Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory 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.

CSI is available at the transmitter through a finite number of bits of feedback from the receiver. Based on the finite-rate feedback, [7] 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, 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 this end, we develop two joint quantization and resource allocation approaches. In the first one, we rely on the quantization regions associated to each AMC mode and the time allocation policy inherited from the perfect CSI case to optimize the fixed transmit-power across quantization states. In the second approach, we pursue separable optimization and resort to coordinate descent algorithms to solve the following two sub-problems: (a) given a time allocation, we optimize the quantization regions and transmit-powers; and (b) with improved quantization regions, we optimize the time allocation policy.

2. MODELING PRELIMINARIES

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

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

where $x_k(n)$ and $h_k(n)$ are the transmitted signal and fading process of the k th user, respectively, and $z(n)$ denotes AWGN with variance $\sigma^2 = 1$. As in [6], we confine ourselves to TDMA where each user transmits in a dedicated time fraction, not overlapping with other users. We assume that $\{h_k(n)\}_{k=1}^K$ are jointly stationary and ergodic with continuous stationary distribution. The joint fading process adheres to a block fading channel model. User transmissions to the access point are naturally frame-based. 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 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 [9]. 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.

Notation: Using boldface lower-case letters to denote column vectors, we let $\mathbf{h} := [h_1, \dots, h_K]^T$ 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.

3. RESOURCE ALLOCATION WITH FINITE AMC MODES AND PERFECT CSI

In this section, we review briefly the energy efficient resource allocation scheme in [6] 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 $\tau(\cdot)$, 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 k th 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 k th 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 $(\rho_{k,l}, p_{k,l})$, $l = 1, \dots, M_k$, 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 l th AMC mode, $p_{k,l}$ is the minimum receive-power required to maintain a prescribed BER. Although the k th 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 [6, Fig. 2])

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

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].

With power cost weights $\boldsymbol{\mu} := [\mu_1, \dots, \mu_K]^T$ and using $\Upsilon_k(x)$, the energy-efficient resource allocation policies with individual rate constraints $\{\bar{R}_k\}_{k=1}^K$ solve the optimization problem

$$\begin{cases} \min_{\mathbf{r}(\cdot), \tau(\cdot)} E_{\mathbf{h}} \left[\sum_{k=1}^K \mu_k \tau_k(\mathbf{h}) \Upsilon_k(r_k(\mathbf{h})) \right] \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \quad \forall k, E_{\mathbf{h}} [\tau_k(\mathbf{h}) r_k(\mathbf{h})] \geq \bar{R}_k. \end{cases} \quad (3)$$

Since every point of $\Upsilon_k(r_k(\mathbf{h}))$ can be achieved by time-sharing between points $(\rho_{k,l}, p_{k,l}/h_k)$, finding the optimal resource allocation

strategies for (3) is equivalent to solving

$$\begin{cases} \min_{\tilde{\tau}(\mathbf{h})} \sum_{k=1}^K E_{\mathbf{h}} \left[\sum_{l=0}^{M_k} \mu_k \frac{\tilde{\tau}_{k,l}(\mathbf{h})}{h_k} p_{k,l} \right] \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \quad \forall k, E_{\mathbf{h}} \left[\sum_{l=0}^{M_k} \tilde{\tau}_{k,l}(\mathbf{h}) \rho_{k,l} \right] \geq \bar{R}_k. \end{cases} \quad (4)$$

It turns out that the optimal resource allocation policies are obtained via greedy water-filling as summarized next (c.f. [6, 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 (4), and subsequently the optimal allocation $r_k^*(\mathbf{h})$ and $\tau_k^*(\mathbf{h})$ for (3) as follows. Given a positive $\boldsymbol{\lambda}^{P^*} := [\lambda_1^{P^*}, \dots, \lambda_K^{P^*}]^T$, for each fading state \mathbf{h} , let $l_k^* := \max \{l : \mu_k \gamma_{k,l}/h_k \leq \lambda_k^{P^*}\}$ ($l_k^* = 0$ 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,*

$$r_i^*(\mathbf{h}) = \rho_{i,l_i^*}, \quad \tau_i^*(\mathbf{h}) = 1; \quad (5)$$

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,*

$$r_{i_j}^*(\mathbf{h}) = \rho_{i_j,l_{i_j}^*}, \quad \tau_{i_j}^*(\mathbf{h}) = \tau_j^*, \quad (6)$$

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

In (5) and (6), $\boldsymbol{\lambda}^{P^*}$ and $\{\tau_j^*\}_{j=1}^J$ should satisfy the individual rate constraints

$$E_{\mathbf{h}} [\tau_k^*(\mathbf{h}) r_k^*(\mathbf{h})] = \bar{R}_k, \quad k = 1, \dots, K. \quad (7)$$

Moreover, $\boldsymbol{\lambda}^{P^*}$ is almost surely unique and can be iteratively computed by [6, 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.

4. QUANTIZATION AND RESOURCE ALLOCATION WITH FINITE-RATE FEEDBACK

With finite-rate feedback from the access point, 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 k th user's l th 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 l th 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 l th 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 k th user's l th 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, 8]. With $\bar{\boldsymbol{\epsilon}} := [\bar{\epsilon}_1, \dots, \bar{\epsilon}_K]^T$ collecting the prescribed BER requirements, the energy-efficient quantization and resource allocation problem is

$$\begin{cases} \min_{\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \forall k, \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k; \\ \forall k, \frac{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h})}{\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})} \leq \bar{\epsilon}_k. \end{cases} \quad (8)$$

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

$$\begin{cases} \min_{\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}(\cdot)} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \forall k, \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k; \\ \forall k, \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \leq \bar{\epsilon}_k. \end{cases} \quad (9)$$

If all rate requirements are met with equality in the optimal solution to (8), solving (9) yields the same optimal solution. However, if some rate requirements are over-satisfied in the optimum of (8), the solution of (9) will upperbound that of (8) since we impose stricter BER constraints. The problem (9) is still complicated and not convex. To solve it, we develop two simplified approaches.

4.1. Initialization

We first use the resource allocation policies of Section III to provide an initial point. 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*

$$Q_{k,l}^* = \{\mathbf{h} : h_k \in [q_{k,l}^*, q_{k,l+1}^*]\}, \quad (10)$$

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 l th 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 l th mode is picked, and thus (10) follows. \square

4.2. Approach I: Optimizing Transmit-Powers

In our energy-efficient quantization and resource allocation, we need to determine the optimal quantization regions \mathbf{Q} , transmit powers $\boldsymbol{\pi}$ and the time allocation policy $\boldsymbol{\tau}(\cdot)$. With P-CSI, Proposition 1 and Lemma 1 yield the optimal allocation of time slots specified by $\boldsymbol{\tau}^*(\cdot)$ and optimal quantization regions by \mathbf{Q}^* . Assuming these $\boldsymbol{\tau}^*(\cdot)$ and \mathbf{Q}^* also provide good approximations for optimal time allocation and fading regions in Q-CSI case, then we can only optimize over the transmit powers $\boldsymbol{\pi}$ to yield a energy-efficient Q-CSI solution. It is clear from (9) 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 $\boldsymbol{\tau}^*(\cdot)$. Now with a pair of \mathbf{Q}^* and $\boldsymbol{\tau}^*(\cdot)$ already satisfying rate constraints, finding the optimal $\boldsymbol{\pi}$ is to solve

$$\begin{cases} \min_{\boldsymbol{\pi}} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}^*} \tau_k^*(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall k, \sum_{l=1}^{M_k} \frac{\rho_{k,l}}{\bar{R}_k} \int_{Q_{k,l}^*} \tau_k^*(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) \leq \bar{\epsilon}_k. \end{cases} \quad (11)$$

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 (11) using a compact \mathbf{Q} containing AMC modes with non-zero measures. Since the functions $\epsilon_{k,l}(x)$ are convex, (11) is a convex optimization problem. Its solution can be analytically obtained as follows.

Proposition 2 *Given a positive $\boldsymbol{\nu}^{\pi^*} := [\nu_1^{\pi^*}, \dots, \nu_K^{\pi^*}]^T$, 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*

$$\int_{Q_{k,l}^*} \tau_k^*(\mathbf{h}) h_k \epsilon'_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) = -\frac{\mu_k \bar{R}_k A_{k,l}}{\rho_{k,l} \nu_k^{\pi^*}}. \quad (12)$$

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

$$\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}^*} \tau_k^*(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}^*) dF(\mathbf{h}) / \bar{R}_k = \bar{\epsilon}_k. \quad (13)$$

Proof: See [13, Sec. IV-B]. \square

Notice that given $\tau_k^*(\mathbf{h})$, users are decoupled. Solving (11) is equivalent to solving K small problems; i.e., $\min \mu_k \sum_{l=1}^{M_k} \pi_{k,l} A_{k,l}$, subject to $\sum_{l=1}^{M_k} \rho_{k,l} \times \int_{Q_{k,l}^*} \tau_k^*(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h}) / \bar{R}_k \leq \bar{\epsilon}_k$. Given $\nu_k^{\pi^*}$ and monotonically decreasing $\epsilon_{k,l}(\gamma)$, the solution to (12) 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 (13). And the optimal transmit-powers $\boldsymbol{\pi}^*$ are in turn obtained. Henceforth, we will refer this simple algorithm as optimizing transmit-power (OTP) algorithm.

4.3. Approach II: Two-Step Coordinate Descend Algorithm

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

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. With this simplification, the optimization problem to solve becomes

$$\begin{cases} \min_{\mathbf{Q}, \tau(\cdot)} \sum_{k=1}^K \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \forall \mathbf{h}, \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1; \\ \quad \forall k, \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k. \end{cases} \quad (14)$$

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

$$f_\epsilon(\pi_{k,l}, \tau_k(\mathbf{h}), Q_{k,l}) := \frac{\int_{Q_{k,l}} \tau_k(\mathbf{h}) \epsilon_{k,l}(h_k \pi_{k,l}) dF(\mathbf{h})}{\int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h})} - \bar{\epsilon}_k. \quad (15)$$

It is easy to check that with so-determined transmit-powers, the average BER constraints are satisfied. Now we can divide the optimization process into two separate sub-problems and then solve each of them in an optimal way; i.e., we resort to a coordinate descent [10] approach to come up with an iterative algorithm which assembles the different sub-solutions to solve the main problem. Our algorithm will run as follows: 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. Notice that this is a well appreciated strategy in the field of quantization theory, and a good example is the Lloyd algorithm

First, given a time allocation policy, users are decoupled. To optimize the quantization regions, we need to solve $\forall k$,

$$\begin{cases} \min_{\{Q_{k,l}\}_{l=1}^{M_k}} \mu_k \sum_{l=1}^{M_k} \pi_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s.t. } \sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \bar{R}_k. \end{cases} \quad (16)$$

Using the distortion measure argument and the *nearest-neighbor rule* as the constrained vector quantization in [12], we can establish:

Proposition 3 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:

$$Q_{k,l}^* = \left\{ \mathbf{h} : \tilde{\psi}_{k,l}(h_k) \leq \tilde{\psi}_{k,j}(h_k); \forall j \neq l, j \in [0, M_k] \right\}. \quad (17)$$

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

$$\sum_{l=1}^{M_k} \rho_{k,l} \int_{Q_{k,l}^*} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \bar{R}_k. \quad (18)$$

Proof: See [13, Sec. IV-D]. \square

With $\tau(\cdot)$ and π , to obtain \mathbf{Q}^* , we only need $\lambda^{s*} := \{\lambda_k^{s*}\}_{k=1}^K$, 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 $\tau(\mathbf{h})$. With \mathbf{Q} and π given, finding the optimal time allocation policy is to solve

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

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

Proposition 4 Given $\lambda^{\tau*} := [\lambda_1^{\tau*}, \dots, \lambda_K^{\tau*}]^T \geq \mathbf{0}$ and $\nu^{\tau*} := [\nu_1^{\tau*}, \dots, \nu_K^{\tau*}]^T \geq \mathbf{0}$, 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 $\tilde{\varphi}_k(\mathbf{h}) := \mu_k \pi_{k,l_k(\mathbf{h})} - \lambda_k^{\tau*} \rho_{k,l_k(\mathbf{h})} + \nu_k^{\tau*} \rho_{k,l_k(\mathbf{h})} \epsilon_{k,l_k(\mathbf{h})} (h_k \pi_{k,l_k(\mathbf{h})}) / \bar{R}_k$. Then the optimal solution $\tau^*(\cdot)$ to (19) 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 $\{\tilde{\varphi}_{i_j}(\mathbf{h})\}_{j=1}^J < 0$, 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, $\lambda^{\tau*}$ and $\nu^{\tau*}$ 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 [13, Appendix]. \square

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. For cases where $\tilde{\varphi}_k(\mathbf{h}) \geq 0 \forall k \in [1, K]$, 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 .

To obtain the optimal $\tau^*(\cdot)$, we need to find $\lambda^{\tau*}$ and $\nu^{\tau*}$. 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 $g(\lambda^\tau, \nu^\tau)$ for (19) is given by

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

where for a given (λ^τ, ν^τ) , the time allocation $\tau_k(\lambda^\tau, \nu^\tau, \mathbf{h})$ is provided by Proposition 4 (without considering the rate and BER constraints). The dual of (19) is

$$\max_{\lambda^\tau, \nu^\tau} g(\lambda^\tau, \nu^\tau), \quad \text{s.t. } \lambda^\tau \geq \mathbf{0}, \nu^\tau \geq \mathbf{0}. \quad (20)$$

Since (19) is a convex problem, the duality gap is zero; and thus $(\lambda^{\tau*}, \nu^{\tau*}) = \arg \max_{\lambda^\tau \geq 0, \nu^\tau \geq 0} g(\lambda^\tau, \nu^\tau)$. Therefore, we can obtain $(\lambda^{\tau*}, \nu^{\tau*})$ via the following sub-gradient projection algorithm. Note that the dual function $g(\lambda^\tau, \nu^\tau)$ is concave since it is the point-wise infimum of a family of affine functions of (λ^τ, ν^τ) , and thus the convergence of our sub-gradient projection algorithm is guaranteed [11].

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

[T1] $\forall k \in [1, K]$, numerically evaluate the partial derivatives $\Delta\lambda_k^\tau := \frac{\partial g(\lambda^\tau, \nu^\tau)}{\partial \lambda_k^\tau}$ and $\Delta\nu_k^\tau := \frac{\partial g(\lambda^\tau, \nu^\tau)}{\partial \nu_k^\tau}$ at $(\lambda^\tau(t-1), \nu^\tau(t-1))$. 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 $g(\lambda^\tau(t), \nu^\tau(t))$ using $(\lambda^\tau(t), \nu^\tau(t))$. If*

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

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

Once $\lambda^{\tau*}$ and $\nu^{\tau*}$ are calculated, the time allocation policy in Proposition 4 is in turn determined. 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, 3 and 4 the following 2-step joint quantization and resource allocation (2S-JQRA) algorithm. Since the global objective J is decreasing in each step of the iterations, as $t \rightarrow \infty$, the 2S-JQRA algorithm converges.

Algorithm 2 [S0] *Initialization: Produce initial time allocation $\tau^{(0)}(\cdot)$ and quantization regions $\mathbf{Q}^{(0)}$ from Proposition 1 and Lemma 1, and then $\pi^{(0)}$ using (15). Select tolerance $\varepsilon > 0$, initial objective $J^{(0)} = \infty$ and let the iteration index $t = 1$.*

[S1] $\tau^{(t-1)}(\cdot), \pi^{(t-1)} \rightarrow \mathbf{Q}^{(t)}, \pi^{(t)}$: Given $\tau^{(t-1)}(\cdot)$ and $\pi^{(t-1)}$, obtain $\mathbf{Q}^{(t)}$ from Proposition 3, and $\pi^{(t)}$ as a function of $\tau^{(t-1)}$ and $\mathbf{Q}^{(t)}$ using (15).

[S2] $\mathbf{Q}^{(t)}, \pi^{(t)} \rightarrow \tau^{(t)}(\cdot)$: Given $\mathbf{Q}^{(t)}$ and $\pi^{(t)}$, obtain $\tau^{(t)}(\cdot)$ from Proposition 4.

[S3] *Stopping criterion: Calculate the objective $J^{(t)}$ using $\mathbf{Q}^{(t)}$, $\pi^{(t)}$ and $\tau^{(t)}(\cdot)$. 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].*

5. 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 $\bar{\gamma}_k = \bar{h}_k/(N_0B)$. 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 [8]

$$\epsilon_{k,l}(\gamma) = 0.2e^{-\frac{\gamma}{2^{\rho_{k,l}}-1}}. \quad (21)$$

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 [6] and our Q-CSIT based OTP and 2S-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. 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 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 OTP and 2S-JQRA clearly outperform the heuristic Q-CSIT approach (yielding around 5 dB savings); and ii) while the gap between 2S-JQRA and P-CSIT solution is small, even the simple OTP algorithm provides a good solution not far away from the P-CSI solution. Since the P-CSIT solution lower bounds all Q-CSIT based approaches, this indicates that our coordinate descend algorithm is near-optimal. The convergence of 2S-JQRA is illustrated in Fig. 3, where the average total weighted power evolves with the inner iteration steps. We can see that 2S-JQRA converges after a small number of iterations (around 6 inner steps or 3 outer iterations). The variations through the curve are due to the finite resolution in the numerical integrations involved. As numerical results have demonstrated that the global performance of both 2S-JQRA and OPT algorithms is comparable, the final selection might take into account the trade-off between complexity and power savings.

6. CONCLUSIONS

Based on Q-CSI, we derived two energy-efficient joint quantization and resource allocation strategies for TDMA fading channels. Nu-

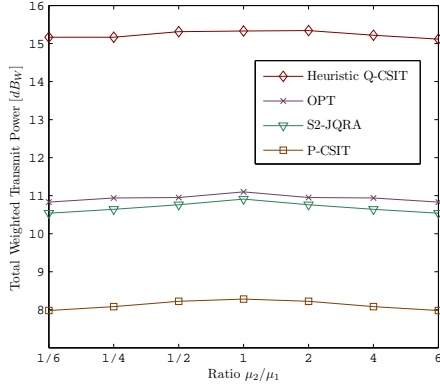


Fig. 1. Total power consumption for different resource allocation approaches with different power weight ratio μ_2/μ_1 and $\sum_{k=1}^2 \mu_k = 1$ when $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100$ kbps, and $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB.

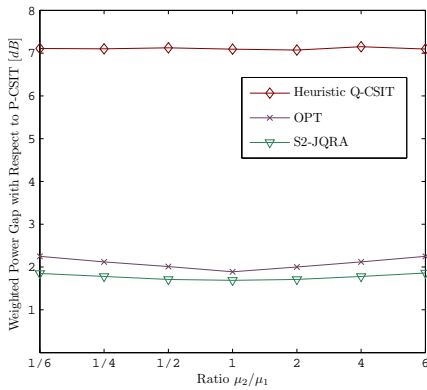


Fig. 2. Performance loss of Q-CSIT based approaches with respect to the P-CSIT solution when $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100$ kbps, and $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB.

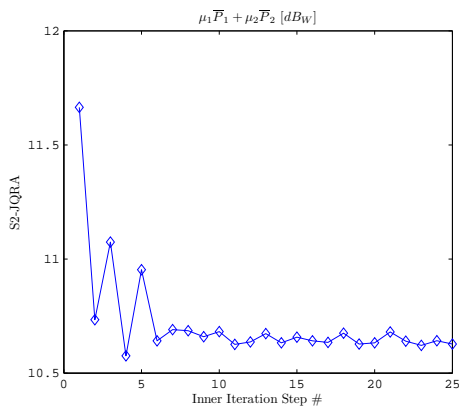


Fig. 3. Average weighted power evolution for S2-JQRA algorithm ($\mu_1 = 2/3$, $\mu_2 = 1/3$, $\bar{\epsilon}_1 = \bar{\epsilon}_2 = 10^{-3}$, $\bar{R}_1 = \bar{R}_2 = 100$ kbps, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB).

merical results showed that with Q-CSIT only available, both algorithms achieve energy efficiency surprisingly close to that obtained with P-CSIT, and yield large energy-savings compared to a heuristic and widely used Q-CSIT approach.¹

7. 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. 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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/>
- [7] 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/>.
- [8] A. J. Goldsmith and S. G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. on Commun.*, vol. 46, pp. 595602, May 1998.
- [9] IEEE 802.16 WG, *Air interface for fixed broadband wireless access systems*, IEEE Std. 802.16, April. 2002.
- [10] D. Bertsekas, *Nonlinear Programming: 2nd Ed.*, Athena Scientific, 1999.
- [11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [12] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Norwell, MA: Kluwer, 1992.
- [13] X. Wang, A. G. Marques, and G. B. Giannakis, "Energy-efficient quantization and resource allocation for TDMA with finite-rate feedback," *IEEE Trans. on Signal Processing*, submitted, May 2006, available at <http://spincom.ece.umn.edu/>

¹The 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.