Power-Efficient Resource Allocation and Quantization for TDMA Using Adaptive Transmission and Limited-Rate Feedback

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

***Abstract*—Power-efficient scheduling and resource allocation are critical tasks for wireless sensor networks as well as commercial and tactical radios relying on IEEE access standards for power-limited communications. Tailored for such applications, this paper formulates and solves analytically novel convex optimization problems offering globally optimal user scheduling, as well as rate and power allocation for time-division multiple access (TDMA) in time-division-duplex or frequency-division-duplex operation. Through a limited-rate feedback link the access point provides quantized channel state information to the transmitters (Q-CSIT) based on which users adapt their modulation and code choices to the intended fading channel. When the quantizer needed to form the Q-CSIT is not prescribed, a joint allocation-quantization scheme is devised to minimize average transmit power subject to average rate and bit error rate constraints. The novel design couples adaptive transmission modes with quantization regions which are constructed to attain at least a local minimum of the average transmit power. Fairness in resource allocation is guaranteed by design. Transmit power and quantization region books are efficiently obtained offline while the online Q-CSIT based operation turns out to entail only a few feedback bits. Analysis and simulations include a perfect CSIT benchmark and reveal substantial power savings (as high as 15 dB) with low-overhead feedback.**

***Index Terms*—Adaptive algorithms, convex optimization, multiple access, power control, quantization, resource management, scheduling, wireless sensor networks.**

I. Introduction

With power efficiency emerging as a critical issue to extend battery lifetime in both commercial as well as tactical radios and wireless sensors, power-efficient resource allocation has attracted growing attention for additive white Gaussian noise (AWGN) channels [17], [8], [21], and time division multi-access (TDMA) fading channels [3], [20]. Resource allocation for fading channels is studied in [15], [10] and power-efficient policies for TDMA are investigated from an information-theoretic perspective in [18]. Assuming that both transmitters and receivers have available perfect (P-) channel state information (CSI), the approaches in [18] 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 (exact values of fading channel coefficients) at the transmitters (P-CSIT) renders analysis and design tractable, it may not be always realistic due to possible channel estimation errors, feedback delay and jamming [9], [13]. These considerations motivate a *limited-rate* feedback model, where only *quantized* (Q-) CSI is available at the transmitters through a few of bits of feedback from the receiving access point. Based on limited-rate feedback, [12] minimized transmit power of single-antenna orthogonal frequency-division multiplexing (OFDM) systems; while a number of recent works capitalize on limited-rate feedback for multi-antenna systems; see, e.g., [14], [5, Ch. 13] and references therein.

In this paper, we deal with wireless TDMA systems in the power-limited regime where single-antenna users rely on Q-CSIT to transmit with AMC, as described in Section II. From a high level view, our contributions consist of formulating and solving analytically constrained optimization problems with the objective of minimizing the weighted average of the system's aggregate transmit power, under average rate and bit error rate (BER) constraints. In addition to algorithms, valuable insights are gained on the fundamental limits for power-efficient TDMA with limited-rate feedback. Our unifying framework incorporates fairness in its design variables and encompasses a globally optimal P-CSIT based resource allocation scheme for time-division-duplex systems (Section III). This scheme is used to initialize and benchmark the Q-CSIT based solution which is particularly attractive for frequency-division-duplex systems (Section IV). When the book of quantization regions required to form the Q-CSIT is prescribed, the user scheduling and power control schemes we develop based on convex optimization algorithms are computationally efficient and are guaranteed to be globally convergent (Sections IV-A and IV-B).

We further develop an iterative (block-coordinate descent) joint allocation-quantization algorithm to optimize the man-

Manuscript received November 22, 2006; revised October 17, 2007. Published August 13, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ananthram Swami. Work in this paper was supported by the ARO Grant 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. The work of A. G. Marques in this paper was partially supported by the C.A. Madrid Government Grant P-TIC-000223-0505. Parts of this paper were presented at the Military Communications Conference, Washington DC, October 2006, and the IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, HI, April 2007.

X. Wang is with the Department of Electrical Engineering, Florida Atlantic University, Boca Raton, FL 33431 USA (e-mail: xin.wang@fau.edu).

A. G. Marques is with the Department of Signal Theory and Communications, Rey Juan Carlos University, Madrid 28943, Spain (e-mail: antonio.garcia.marques@urjc.es).

G. B. Giannakis is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: georgios@ece.umn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2008.924635

agement of transmit power and rate resources together with the form of Q-CSIT used by the terminals (Sections IV-C and IV-D). Coupling an innovative construction of power-efficient quantization regions with the choice of AMC modes reduces complexity in the off-line design of the quantizer but also leads to low-overhead feedback requirements in the on-line operation of the TDMA system. Simulated test cases compare the novel schemes against a heuristic alternative and the fundamental limits derived in [18] (Section V). The conclusions (Section VI) contend that analysis and simulations corroborate the high potential for practical deployment in view of the sizeable power savings that result at affordable off-line complexity and low-overhead messages fed back during the on-line operation.

II. Preliminaries and Problem Statement

Consider K terminals (users) in uplink TDMA frame-based communication with an access point over wireless flat fading channels adhering to the following operating conditions:

(oc-1) *Each user $k = 1,\ldots,K$ transmits in a separate user-specific time slot relying on a finite set of adaptive modulation and coding (AMC) pairs (modes $m = 1,\ldots,M_k$) each with corresponding rate $\rho_{k,m}$.*

(oc-2) *Flat fading channel coefficients $\{\sqrt{h_k}e^{j\theta_k}, h_k \geq 0\}_{k=1}^K$ remain invariant over a frame of duration T_f but are allowed to vary from frame-to-frame (block fading model). With^T denoting transposition, the resultant $K \times 1$ vector of channel gains $\mathbf{h} := [h_1,\ldots,h_K]^T$ is ergodic with continuous joint cumulative distribution function (CDF) $F(\mathbf{h})$ assumed known; e.g., Rayleigh if $\{\sqrt{h_k}e^{j\theta_k}\}_{k=1}^K$ are jointly complex Gaussian.*

All users are allowed to transmit per frame over nonoverlapping nonnegative *fractions* $\{\tau_k(\mathbf{h})\}_{k=1}^K$ whose duration depends on the channel realization $\mathbf{h}$. If we suppose without loss of generality (w.l.o.g.) that each frame has duration $T_f = 1$, then clearly $\sum_{k=1}^K \tau_k(\mathbf{h}) \in [0,1]$. Notice that the latter allows all users, or at the other extreme no user, transmitting over a given frame. Furthermore, if $r_k(\mathbf{h})$ denotes rate in bits/sec/Hz, then the kth user's transmission rate per frame is $r_k(\mathbf{h})\tau_k(\mathbf{h})$. Likewise, we will let $p_k(\mathbf{h})\tau_k(\mathbf{h})$ denote the transmit power of user k per frame.

According to (oc-1), user k can select in each frame a modulation with rate $\rho_{k,m}^{(\mathrm{mod})}$ along with a channel code with rate $\rho_{k,m}^{(\mathrm{cod})}$ to transmit with AMC rate $r_k(\mathbf{h}) = \rho_{k,m} := \rho_{k,m}^{(\mathrm{mod})}\rho_{k,m}^{(\mathrm{cod})}$. In addition to these prespecified AMC rates $\rho_{k,m}$ (that can be different from user to user), it is also possible for each terminal to transmit with linear combinations of $\rho_{k,m}$ by time sharing their usage over its own slot. For instance, using the mode m over $\alpha_{k,m}$ percentage of the τ_k slot and the mode $m+1$ in the remaining $(1-\alpha_{k,m})$ time, user k can transmit over a frame with rate $r_k(\mathbf{h})\tau_k(\mathbf{h}) = \tilde{\tau}_{k,m}(\mathbf{h})\rho_{k,m} + \tilde{\tau}_{k,m+1}(\mathbf{h})\rho_{k,m+1}$, where $\tilde{\tau}_{k,m}(\mathbf{h}) := \alpha_{k,m}\tau_k(\mathbf{h})$ and $\tilde{\tau}_{k,m+1}(\mathbf{h}) := (1-\alpha_{k,m})\tau_k(\mathbf{h})$. In general, user k can transmit with rate

$$r_k(\mathbf{h})\tau_k(\mathbf{h}) = \sum_{m=0}^{M_k} \tilde{\tau}_{k,m}(\mathbf{h})\rho_{k,m} \tag{1}$$

where we note that the first mode $m = 0$ corresponds to zero rate (in which case the user defers since $\rho_{k,0} = 0$) and the last mode $m = M_k$ corresponds to the maximum rate ρ_{k,M_k} each user can transmit with.

To respect user-specific quality of service requirements, transmissions in our TDMA system will also adhere to *average* rate and bit error rate (BER) constraints. With $E_{\mathbf{h}}$ denoting expectation with respect to (w.r.t.) the vector of channel gains, the average rate of user k is given by

$$\bar{R}_k := E_{\mathbf{h}}[r_k(\mathbf{h})\tau_k(\mathbf{h})] := \int_{\mathbf{h}} r_k(\mathbf{h})\tau_k(\mathbf{h})dF(\mathbf{h}) \tag{2}$$

and must remain above a prescribed[1] and feasible average rate; i.e., $\bar{R}_k \geq \check{R}_k$ with $\check{R}_k \leq \rho_{k,M_k}\ \forall k$.

Likewise, the average BER constraint will satisfy $\bar{\epsilon}_k \in [0, \check{\epsilon}_k]$ $\forall k = 1,\ldots,K$, where

$$\epsilon_k = \epsilon_k(h_k p_k, r_k) \tag{3}$$

stands for the instantaneous BER function which naturally depends on the transmit rate r_k, receive-power $h_k p_k$ and the variance of the additive white Gaussian noise (AWGN) at the receiver which for notational brevity is fixed here to 1. As an example, it is known that the ϵ_k function for an M-ary QAM mode m can be well approximated as $\epsilon_{k,m} = \kappa_1 \exp[-\kappa_2 h_k p_k/(2^{\rho_{k,m}^{(\mathrm{mod})}} - 1)]$, where κ_1, κ_2 are modulation-dependent constants [6]. (In the coded case, the AMC rate $\rho_{k,m}^{(\mathrm{mod})} := \log_2(M)$ must be multiplied by $\rho_{k,m}^{(\mathrm{cod})}$ and the coding gain can be taken into account through the constant κ_2.)

The relationship (3) between transmit-rate, transmit power and BER will play an instrumental role in reaching our *objective* to minimize the *weighted average transmit power* (with weights $[w_1 \cdots w_K]^T := \boldsymbol{w}$)

$$\begin{aligned}\bar{P} &= \sum_{k=1}^K w_k \bar{P}_k := \sum_{k=1}^K w_k E_{\mathbf{h}}[p_k(\mathbf{h})\tau_k(\mathbf{h})] \\ &= \sum_{k=1}^K w_k \int_{\mathbf{h}} p_k(\mathbf{h})\tau_k(\mathbf{h})dF(\mathbf{h})\end{aligned} \tag{4}$$

given requirements $\{\check{R}_k, \check{\epsilon}_k\}_{k=1}^K$ and availability of channel state information at the transmitters (CSIT).

The forms of CSIT to be considered are perfect (P) and quantized (Q). Essentially perfect (P-)CSIT, i.e., each realization $\mathbf{h}$, can be acquired at the terminals with sufficiently long training sequences when the fading process is relatively slow and a reverse link is available as in time-division-duplex operation. On the other hand, Q-CSIT offers the only practical option with frequency-division-duplex systems where channel reciprocity does not hold; hence, CSI in e.g., the forward link cannot be obtained via training over the reverse link. The Q-CSIT in such systems is provided through a finite-rate feedback channel and is typically described by a codeword $\mathbf{c}(\mathbf{h})$ of the forward channel $\mathbf{h}$. If h_k falls in a quantization region $\mathcal{H}_{k,m}$ over which user k can support mode m, then in order to describe (i.e., index) all the possible $(\prod_{k=1}^K (M_k+1))$ vector quantization regions, the codeword $\mathbf{c}(\mathbf{h})$ must carry $\lceil \sum_{k=1}^K \log_2(M_k+1) \rceil$ bits, where

[1]All average (prescribed) quantities x in this paper will be denoted with $\bar{x}$ (respectively $\check{x}$).

$\lceil\ \rceil$ stands for the ceiling operator. This number of feedback bits can be prohibitive for K large; but as we will see in Sections III and IV, a significantly smaller number suffices for online operation of our AMC-based TDMA system.

Existing resource allocation approaches allow for general multiple access but assume infinite size capacity-achieving Gaussian codebooks, instantaneous or average rate constraints, AWGN channels and/or P-CSIT only operation [15], [10], [17], [3], [21]. Our TDMA setup on the other hand is tailored for CDMA2000 $1\times$EV-DO, WCDMA HSPDA and IEEE 802.16a standards [7], where the goal is to develop resource allocation schemes based on AMC modes under practical average rate and BER constraints. In addition to P-CSIT, we will develop optimal Q-CSIT based algorithms which besides resource allocation will be optimized w.r.t. the chosen quantization regions and for implementation purposes must entail affordable feedback.

Since fairness is a key issue in multi-access systems, we will close this section remarking that our framework guarantees fairness in resource allocation through two different mechanisms: i) the weights $\{w_k\}_{k=1}^K$ in the objective function and ii) the prescribed average rate constraints $\{\check{R}_k\}_{k=1}^K$.

III. Optimal Resource Allocation Based on P-CSIT

In this section, we derive power-efficient resource allocation for TDMA based on AMC and P-CSIT. The motivation is threefold: i) for time-division-duplex systems operating over slow fading channels where the duration of the aggregate downlink-uplink slot is considerably smaller than the channel coherence time, P-CSIT can be assumed available via sufficiently long training through the reverse link; ii) P-CSIT based allocation will guide the design steps and provide a feasible initialization of its Q-CSIT based counterpart we will develop in the next section for frequency-division-duplex systems; and iii) performance of P-CSIT based allocation will benchmark that of Q-CSIT. (This should not be surprising since P-CSIT is a limiting form of Q-CSIT with infinite number of feedback bits).

Given P-CSIT $\mathbf{h}$, the average rate of user k can be expressed in terms of the known AMC modes $\{\rho_{k,m}\}$ and the unknown fractions $\{\tilde{\tau}_{k,m}\}$ as [cf. (1) and (2)] $\bar{R}_k = E_{\mathbf{h}}[\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})\rho_{k,m}]$. In addition, given h_k and $\check{\epsilon}_k$ it is possible for each rate r_k to solve (3) w.r.t. the transmit power to obtain $p_k = (1/h_k)\epsilon_k^{-1}(\check{\epsilon}_k, r_k)$, where $\epsilon_k^{-1}(\cdot)$ denotes the inverse function. Furthermore, time-sharing implies that any rate expressed by a linear combination of AMC modes $\{\rho_{k,m}\}$ as in (1), gives rise to the same linear combination of corresponding minimum transmit powers, call them $\{p_{k,m}(h_k) := (1/h_k)\epsilon_k^{-1}(\check{\epsilon}_k, \rho_{k,m})\}$, which meet the prespecified BER constraint $\check{\epsilon}_k$ for a given $\mathbf{h}$. Hence, the transmit power of user k per realization $\mathbf{h}$ can be expressed in terms of the powers $\{p_{k,m}(h_k)\}$ and the unknown fractions $\{\tilde{\tau}_{k,m}\}$ as $P_k(\mathbf{h}) = \sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})p_{k,m}(h_k)$; and therefore, $\bar{P}_k := E_{\mathbf{h}}[P_k(\mathbf{h})] = E_{\mathbf{h}}[\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})p_{k,m}(h_k)]$.

Recapitulating, transmit power and transmit rate per user are both expressible as a function of $\{\tilde{\tau}_{k,m}(\mathbf{h})\}$ and are coupled in a way that automatically satisfies the BER constraints per realization $\mathbf{h}$. Hence, the power-efficient allocation under P-CSIT amounts to finding the vector $\tilde{\boldsymbol{\tau}}(\mathbf{h})$ which comprises the fractions $\{\tilde{\tau}_{k,m}(\mathbf{h}),\ k=1,\ldots,K,\ m=1,\ldots,M_k\}$ so that ("s. to" stands for subject to)

$$\begin{cases}\min_{\tilde{\boldsymbol{\tau}}(\mathbf{h})\geq\mathbf{0}} & \sum_{k=1}^K w_k E_{\mathbf{h}}\left[\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})p_{k,m}(h_k)\right] \\ \text{s. to} & E_{\mathbf{h}}\left[\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})\rho_{k,m}\right]\geq\check{R}_k,\ k=1,\ldots,K \\ & \sum_{k=1}^K\ \sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})\leq 1\quad\forall\mathbf{h}\end{cases} \tag{5}$$

with $\boldsymbol{\lambda}^{P_\tau} := [\lambda_1^{P_\tau},\ldots,\lambda_K^{P_\tau}]^T$ denoting the Lagrange multipliers associated with the average rate constraints[2] and ignoring temporarily the constraint on time fractions, the Lagrangian of this constrained minimization problem can be written as

$$L(\boldsymbol{\lambda}^{P_\tau},\tilde{\boldsymbol{\tau}}(\mathbf{h})) = \sum_{k=1}^K\sum_{m=0}^{M_k}E_{\mathbf{h}}\left[\varphi_{k,m}^{P_\tau}\left(\mathbf{h},\lambda_k^{P_\tau}\right)\tilde{\tau}_{k,m}(\mathbf{h})\right] + \sum_{k=1}^K\lambda_k^{P_\tau}\check{R}_k \tag{6}$$

where the *instantaneous* cost function $\varphi_{k,m}^{P_\tau}(\mathbf{h},\lambda_k^{P_\tau}) := w_k p_{k,m}(h_k) - \lambda_k^{P_\tau}\rho_{k,m}$ depends on the kth entry of $\boldsymbol{\lambda}^{P_\tau}$ corresponding to the kth average rate constraint. The Lagrange dual function is then given by

$$g(\boldsymbol{\lambda}^{P_\tau}) = \min_{\tilde{\boldsymbol{\tau}}(\mathbf{h})\geq\mathbf{0}} L(\boldsymbol{\lambda}^{P_\tau},\tilde{\boldsymbol{\tau}}(\mathbf{h}))$$
$$\text{s. to}\sum_{k=1}^K\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})\leq 1\ \forall\mathbf{h}$$

and the solution for the dual problem of (5) is $g(\boldsymbol{\lambda}^{P_\tau *}) = \max_{\boldsymbol{\lambda}^{P_\tau}\geq\mathbf{0}} g(\boldsymbol{\lambda}^{P_\tau})$.

Since (5) is a strictly feasible and convex optimization problem (w.r.t. $\tilde{\boldsymbol{\tau}}(\mathbf{h})$), its optimum coincides with that of the dual problem $g(\boldsymbol{\lambda}^{P_\tau *})$ [2, pp. 226]. Consider now the user index and mode index for which the cost function in (6) is minimized per channel realization as

$$(k^*,m_{k^*}^*) := \arg\min_{(k,m)}\varphi_{k,m}^{P_\tau}\left(\mathbf{h},\lambda_k^{P_\tau *}\right)$$
$$\varphi_{k,m}^{P_\tau}\left(\mathbf{h},\lambda_k^{P_\tau *}\right) := w_k p_{k,m}(h_k) - \lambda_k^{P_\tau *}\rho_{k,m}. \tag{7}$$

Then it follows readily that $\forall\mathbf{h}$,

$$\sum_{k=1}^K\sum_{m=0}^{M_k}\varphi_{k,m}^{P_\tau}\left(\mathbf{h},\lambda_k^{P_\tau}\right)\tilde{\tau}_{k,m}(\mathbf{h}) \geq \varphi_{k^*,m_{k^*}^*}^{P_\tau}\left(\mathbf{h},\lambda_k^{P_\tau *}\right)\sum_{k=1}^K\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h}).$$

But for $\varphi_{k^*,m_{k^*}^*}^{P_\tau}(\mathbf{h},\lambda_k^{P_\tau *}) < 0$ this last lower bound is satisfied as an equality if we assign the entire frame to the terminal k^* and

[2]Throughout λ (respectively μ) will denote Lagrange multipliers associated with average rate (respectively BER) constraints. Superscript $^{P_\tau *}$ (also used in cost functions φ) will indicate that the corresponding quantity relies on P-CSIT to obtain the optimal (always denoted by *) user-time allocation τ. For Q-CSIT, we will use e.g., $\lambda^{Q_p *}$ when computing optimal power allocation p.

have it transmit with AMC mode $m^*_{k^*}$; hence, for the optimum allocation of time fractions per realization $\mathbf{h}$ we have

$$\text{If } \varphi^{P_\tau}_{k^*,m_{k^*}}\left(\mathbf{h},\lambda^{P_\tau *}_k\right)<0,\ \text{then}\tilde{\tau}^*_{k^*,m^*_{k^*}}(\mathbf{h})=1 \text{ and } \tilde{\tau}^*_{k,m}(\mathbf{h})=0 \text{ for } (k,m)\neq(k^*,m^*_{k^*}) \quad (8)$$

or the rare (and somewhat trivial) case where all users defer

$$\text{If } \varphi^{P_\tau}_{k^*,m^*_{k^*}}\left(\mathbf{h},\lambda^{P_\tau *}_k\right)=0,\ \text{then } \tilde{\tau}^*_{k,m}(\mathbf{h})=0\,\forall k,\ m. \quad (9)$$

The nontrivial allocation (8) clearly satisfies $\sum_{k=1}^{K}\sum_{m=0}^{M_k}\tilde{\tau}_{k,m}(\mathbf{h})=1$ (likewise the trivial one in (9) renders this sum 0); and since either one of the two is optimum per realization $\mathbf{h}$, they also minimize the average transmit power in (5) provided we have a means of obtaining the optimum Lagrange multiplier vector $\boldsymbol{\lambda}^{P_\tau *}$. Notice that in general $\varphi^{P_\tau}_{k,m}(\mathbf{h},\lambda^{P_\tau *}_k)<0\ \forall k$ since otherwise the trivial case is always in force. With the optimum time allocation in (8) the optimum average rate (power) can be found after retaining only the summand $\rho_{k^*,m^*_{k^*}}$ (respectively $p_{k^*,m^*_{k^*}}(h_k)$) in (5) and integrating over the known CDF of $\mathbf{h}$.

To find $\lambda^{P_\tau *}_k$ we will satisfy each average rate constraint in (5). Specifically, if the argument (i) denotes iteration index and $\beta_\lambda>0$ a stepsize, we rely on iterating until convergence the recursion

$$\lambda^{P_\tau}_k(i+1)=\left[\lambda^{P_\tau}_k(i)+\beta_\lambda\left(\check{R}_k-E_{\mathbf{h}}\left[\sum_{m=0}^{M_k}\rho_{k,m}\tilde{\tau}_{k,m}(\boldsymbol{\lambda}^{P_\tau}(i);\mathbf{h})\right]\right)\right]^+ \quad (10)$$

where $[x]^+:=\max(x,0)$ ensures that the Lagrange multipliers are always nonnegative as they should for our minimization problem. Recursion (10) represents a standard (sub-)gradient update whose fast convergence to the *unique* global optimum $\lambda^{P_\tau *}_k$ from any initial condition, say $\lambda^{P_\tau}_k(0)=0\ \forall k$, is always guaranteed because the objective as well as the constraints in (5) are convex w.r.t. $\{\{\tilde{\tau}_{k,m}\}_{k=1}^K\}_{m=0}^{M_k}$ ([1], Proposition 6.3.1).

In practice, $E_{\mathbf{h}}$ in (10) can be replaced by $N^{-1}\sum_{n=1}^{N}\sum_m\rho_{k,m}\tilde{\tau}_{k,m}(i;\mathbf{h}^{(n)})$, where $\{\mathbf{h}^{(n)}\}_{n=1}^N$ are realizations of $\mathbf{h}$ generated from $F(\mathbf{h})$, particularly easily when channels across users are independent, or, when $\{\sqrt{h_k}e^{j\theta_k}\}_{k=1}^K$ are (even correlated) complex Gaussian. Recalling the user-time allocation policy in (8) or (9), at most one user-mode pair (k,m) is nonzero for each $\mathbf{h}^{(n)}$ per iteration. Notice that $\lambda^{P_\tau}_k(i-1)$ is used now in (7) to compute the winner terminal $k^*(i;\mathbf{h}^{(n)})$ and mode $m^*_{k^*}(i;\mathbf{h}^{(n)})$ per realization $\mathbf{h}^{(n)}$, based on which we find $r_k(i;\mathbf{h}^{(n)})=\rho_{k^*(i;\mathbf{h}^{(n)}),m^*_{k^*}(i;\mathbf{h}^{(n)})}$ and $\tilde{\tau}_{k,m}(i;\mathbf{h}^{(n)})=1$ for $k=k^*(i;\mathbf{h}^{(n)})$ and $m=m^*_{k^*}(i;\mathbf{h}^{(n)})$, and 0 otherwise. As stopping rule for the iterations in (10), we check whether the relative difference $|g(\boldsymbol{\lambda}^{P_\tau}(i+1))-g(\boldsymbol{\lambda}^{P_\tau}(i))|/|g(\boldsymbol{\lambda}^{P_\tau}(i+1))|$ of the dual function [defined after (5)] drops below a preselected tolerance level, in which case we return $\lambda^{P_\tau *}_k=\lambda^{P_\tau}_k(i+1)\,\forall k$. The trivial case for which convergence occurs to $\lambda^{P_\tau *}_k=0$ for some k, implies that the corresponding average rate is satisfied as a strict inequality (cf. the Karush–Kuhn–Tucker conditions [2, pp. 243]).

Before summarizing our P-CSIT based allocation scheme, it is worth in this optimization problem (as well in those of the ensuing section) to pay attention on what can be obtained offline and what is needed during the online operation. Clearly, since the expected value in (10) requires only knowledge of the channel gain CDF $F(\mathbf{h})$, the optimum Lagrange multipliers can be found offline using long-term statistical information of the wireless fading channel. Interestingly, with $\boldsymbol{\lambda}^{P_\tau *}$ available off-line and perfect knowledge of h_{k^*} obtained from the reverse link, the access point needs to broadcast online only the index $\mathbf{c}(k^*)$ of the minimum-cost terminal k^* along with the index $\mathbf{c}(m^*_{k^*})$ of its minimum-cost mode $m^*_{k^*}$ found as in (7). Using those, the "winner user" will transmit with AMC rate $\rho_{k^*,m^*_{k^*}}$ and power $p_{k^*,m^*_{k^*}}(h_k)=(1/h_{k^*})\epsilon_k^{-1}(\check{\epsilon}_k,\rho_{k^*,m^*_{k^*}})$. This low overhead in the feedback results because we posed average (as opposed to instantaneous) rate constraints. Through the use of AMC modes, the BER constraints in the P-CSIT based operation are also automatically satisfied for each channel realization (and thus on average as well). In addition, the average rate constraints are decoupled across users which implies that the gain h_{k^*} needed for the winner-terminal to select its transmit power is readily available from the reverse link in time-division-duplex systems; i.e., it is not necessary for the access point to broadcast the realization $\mathbf{h}$. In summary, we have established the following proposition.

Proposition 1: Under (oc-1) and (oc-2), minimization of the weighted average transmit power under average rate and BER constraints reduces to the constrained minimization problem (5) over user-time fractions $\tilde{\tau}_{k,m}$. Its almost surely optimal solution corresponds to a greedy allocation [cf. (7) and (8) or (9)] where at most one minimum-cost user k^* transmits over the entire frame with a minimum-cost AMC rate $m^*_{k^*}$ and minimum power adapted to the P-CSIT h_{k^*} so that the prescribed BER $\check{\epsilon}_{k^*}$ is satisfied. The Lagrange multipliers $\{\lambda^{P_\tau *}_k\}_{k=1}^K$ required to obtain the minimum cost are computable at the access point offline using the channel gain CDF; while the online operation requires low overhead for feeding back $\mathbf{c}(\mathbf{h})=[\mathbf{c}^T(k^*),\mathbf{c}^T(m^*_{k^*})]^T$ carrying $\lceil\log_2(\sum_{k=1}^K M_k+1)\rceil$ bits from the access point to the terminals.

Although we allowed users to share each frame at the outset, the power-efficient allocation ended up being a greedy (or opportunistic) one. The opportunity per channel realization is given at most to a single user transmitting with a single AMC mode minimizing the functional $\varphi^{P_\tau}_{k,m}(\mathbf{h},\lambda^{P_\tau *}_k)$ which captures the smallest *net* transmit power cost (power spent minus rate rewarded) depending on the channel quality $\mathbf{h}$. Notice that the larger the "winner-user" channel gain h_{k^*}, the higher transmit rate $r_{k^*}(h_{k^*})$ can be afforded while meeting the prespecified BER constraint. At the other extreme, if all users experience a deep fade ($h_k\approx 0\ \forall k$) we will have $m=0\ \forall k$, in which case $r_{k^*}(h_{k^*})=0$ and all users will defer [cf. (9)]. These observations show that the optimum allocation asserted by Proposition 1 is in the spirit of the well known water-filling principle typically encountered when maximizing sum-capacity subject to power constraints.

When the winner pair entails the zeroth transmission mode, which terminal gains access to the channel is irrelevant since the transmit-configuration of all user terminals

is identical no matter who the winner is. This means that instead of indexing any mode-user combination (which requires $B = \lceil \log_2(\sum_{k=1}^{K}(M_k+1)) \rceil$ feedback bits) it suffices to index any *active* mode-user combination along with the all-users-defer case. This explains why only $B = \lceil \log_2(\sum_{k=1}^{K} M_k + 1) \rceil$ feedback bits suffice as per Proposition 1. A couple of remarks are now in order.

Remark 1: It is intriguing that the P-CSIT based "winner-takes-all" optimal allocation which has been derived for rate maximizing general multiple access systems under average power constraints assuming infinite-size capacity-achieving codebooks [10] carries over to the specific TDMA setup considered here which relies on a finite pool of AMC modes to minimize average transmit power under average BER and rate requirements. Note that in addition to differences in the criterion, constraints and operating conditions, the formulation here accommodates transmission modes used in the standards.

Remark 2: In finding the optimal user-mode pair in (7) it was tacitly assumed that the net costs $\varphi_{k,m}^{P_\tau}(\mathbf{h}, \lambda_k^{P_\tau *})$ have a unique maximizer $\forall \boldsymbol{h}$. However, this is not the case when the maximum $\varphi_{k,m}^{P_\tau}(\mathbf{h}, \lambda_k^{P_\tau *})$ is attained by multiple user-mode pairs for a given fading state $\boldsymbol{h}$. Interestingly, the event of having more that one winner per slot has Lebesgue measure zero for ergodic fading channels with continuous CDF—the typical case in wireless systems. As a result, the allocation implemented when multiple winning users tie has measure zero contribution to the average power cost; and thus any arbitrary (deterministic or random) assignment to a single user among those tied yields the same average cost; see also [15, Lemma 3.15] for related comments in a different context. This explains the almost sure optimality asserted in Proposition 1. For deterministic channels or random channels with discrete probability density functions however, these ties are not measure-zero events and have to be accounted for [22]. Specifically, the optimal time-sharing fractions among the multiple winners must be determined to ensure that the individual rate constraints are satisfied.

IV. Optimal Resource Allocation and Quantization Based on Q-CSIT

In this section we deal with power-efficient TDMA based on AMC and Q-CSIT, a setup particularly appealing for frequency-division-duplex systems in the power-limited regime. When quantization regions $\mathcal{H}_{k,m}$ are *a priori* specified, the schemes in this section yield the optimal allocation of user times and power allocations per frame. At least as important, the present section develops a systematic block-coordinate descent algorithm to jointly optimize the allocation as well as the selection of quantization regions that influence the form of Q-CSIT used.

But before specifying this form, it is instructive to point out the differences emerging when one replaces P-CSIT with Q-CSIT. Those appear in at least two facets:

(d1) with the Q-CSIT vector $\mathbf{c}(\mathbf{h})$ containing a finite number of (say B) bits, only a finite number of choices (namely 2^B) are available for the terminals to adapt; and

(d2) in lieu of $\mathbf{h}$, the coupling of transmit rate with transmit power that allows one to be computed from the other for a prescribed BER (via (3)) is no longer possible. This in turn implies three things: (i) an extra set of power variables $\boldsymbol{p} := [\boldsymbol{p}_1^T \cdots \boldsymbol{p}_K^T]^T$, where $\{\boldsymbol{p}_k := [p_{k,1} \cdots p_{k,M_k}]^T\}_{k=1}^K$, in the optimization problem; along with (ii) an extra set of Lagrange multipliers associated with average BER constraints that must now be explicitly accounted for; and (iii) impossibility to express transmit power in terms of the fractions $\tilde{\boldsymbol{\tau}}(\mathbf{h})$ which allow for linear combinations through time sharing; as a result, henceforth only the original TDMA fractions $\tau_k(\mathbf{h})$ will be used per user k to form the user-time allocation vector $\boldsymbol{\tau}(\mathbf{h}) := [\tau_1(\mathbf{h}) \cdots \tau_K(\mathbf{h})]^T$.

Since a quantizer is basically a classifier, to define the Q-CSIT we need to specify the input, output, type and number of classes as they relate to our problem at hand. These are as follows:

(oc-3) *For each terminal k we consider $M_k + 1$ classes, as many as its AMC modes; the input is $\mathbf{h}$ and the output is the vector $\mathbf{c}(m_k^*; \mathbf{h})$ indexing the selected mode m_k^*; if $\mathbf{h} \in \mathcal{H}_{k,m^*}$ and terminal k is selected by the allocation scheme, then it transmits with rate $r_k = \rho_{k,m^*}$ and power $p_k = p_{k,m^*}$ both of which are assumed constant over the region $\mathcal{H}_{k,m^*}$. If not a priori specified, the regions $\mathcal{H}_{k,m}$ are found using a suitable distance criterion as elaborated in Section IV-C. We will show that the winner-takes-all policy is also optimum when Q-CSIT is used. This implies that the Q-CSIT vector $\mathbf{c}(\mathbf{h})$ does not need to index each and every quantization region but only the one corresponding to the winner user, namely $\mathbf{c}(\mathbf{h}) := [\mathbf{c}^T(k^*; \mathbf{h}), \mathbf{c}^T(m_{k^*}^*; \mathbf{h})]^T$, i.e., the index of the selected terminal $\mathbf{c}(k^*; \mathbf{h})$ and its selected mode index $\mathbf{c}(m_{k^*}^*; \mathbf{h})$.*

The novel coupling of AMC modes with quantization regions introduced by (oc-3) is motivated by the P-CSIT based setup where even though a linear combination of rates was allowed, the nontrivial solution ended up entailing only a single AMC rate for the "winner terminal." This coupling will prove beneficial not only in formulating tractable convex optimization problems for optimal resource allocation but further in reducing the feedback overhead. Contributing to the overhead reduction is also the fact that quantization in (oc-3) is decoupled across users. In essence, we have K quantization problems each entailing $M_k + 1$ classes; hence the Q-CSIT $\mathbf{c}(\mathbf{h})$ under (oc-3) will only need to carry $B = \lceil \log_2(\sum_{k=1}^{K}(M_k+1)) \rceil$ bits (we will see that this number can be further reduced to $B = \lceil \log_2(\sum_{k=1}^{K} M_k + 1) \rceil$). The quantization regions per user are nonoverlapping across modes but for a given mode they overlap across users; i.e., with $\cap(\cup)$ denoting set intersection (union), we have $\mathcal{H}_{k,m} \cap \mathcal{H}_{k,m'} = \emptyset$ for $m \neq m'$, but $\mathcal{H}_{k,m} \cap \mathcal{H}_{k',m} \neq \emptyset$ for $k \neq k'$; and certainly $\cup_{m=0}^{M_k} \mathcal{H}_{k,m} = \mathcal{D}_{\mathbf{h}} \subseteq \mathbb{R}_0^+ \forall k$, where $\mathcal{D}_{\mathbf{h}}$ denotes the domain of $\mathbf{h}$ in the CDF which is a subset of the nonnegative K-dimensional real vectors $\mathbb{R}_0^+$. Had we coupled quantization also across users, we would have a single classification problem with nonoverlapping regions but with more classes $(\prod_{k=1}^{K}(M_k+1))$.

One could be tempted to rely on a conventional channel quantizer with output the index of the region and centroids $\{\mathbf{h}_Q\}$ lying "close" to $\mathbf{h}$ in e.g., the minimum mean-square error (MMSE) sense, as in the scalar or vector Lloyd quantizer [11]. However, this approach is clearly suboptimum in two counts: first, the average BER constraints would be impossible

to guarantee due to the aforementioned (d2); and second, the MMSE optimal quantizer would not ensure optimality in the minimum transmit power sense sought here.

Having specified the Q-CSIT form, we proceed to express the average power, rate, BER and variables involved in the ensuing optimization problems. Since the adaptive configurations are finite as per (d1), the transmit power $p_k(\mathbf{c}(\mathbf{h}))$ of user k per channel realization $\mathbf{h}$ is a discrete random variable; while the corresponding time fraction $\tau_k(\mathbf{h})$ is a continuous one over $[0,1]$. Hence, $\bar{P}_k := E_{\mathbf{h}}[p_k(\mathbf{c}(\mathbf{h}))\tau_k(\mathbf{h})] = \int_{\mathcal{D}_{\mathbf{h}}} p_k(\mathbf{c}(\mathbf{h}))\tau_k(\mathbf{h})dF(\mathbf{h}) = \sum_{m=0}^{M_k}\int_{\mathcal{H}_{k,m}} p_k(\mathbf{c}(\mathbf{h}))\tau_k(\mathbf{h})dF(\mathbf{h})$, where the last equality follows after splitting the integral over the nonoverlapping quantization regions. But since under (oc-3) the transmit power per region is constant, we arrive at (recall that $p_{k,0}=0$, and thus the defer-mode $m=0$ need not be included)

$$\bar{P}_k = \sum_{m=1}^{M_k} p_{k,m}\int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h})dF(\mathbf{h}). \tag{11}$$

Arguing along the same lines it follows readily that the average transmit rate can be expressed as (recall also that $\rho_{k,0}=0$)

$$\bar{R}_k = \sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h})dF(\mathbf{h}) \geq \check{R}_k, \quad k=1,\ldots,K \tag{12}$$

where the integral can be interpreted as the probability that terminal k uses mode m (i.e., r_k takes the value $\rho_{k,m}$).

The fact that the rate per user is a random variable in adaptive transmissions prevents one from expressing the average BER by simply integrating (3) over $F(\mathbf{h})$, as when the rate r_k is deterministically constant. Instead, we need to express $\bar{\epsilon}_k$ as the ratio of the average number of erroneously received bits over the average number of transmitted bits per frame. The denominator is simply the average rate in (12). Arguing as in (11) and (12), the numerator of this ratio is $\sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})\epsilon_{k,m}(h_k p_{k,m})dF(\mathbf{h})$, where $\epsilon_{k,m}$ is the instantaneous BER of user k using mode m. (Notice that since the rate is constant over $\mathcal{H}_{k,m}$, different from (3), $\rho_{k,m}$ does not appear in $\epsilon_{k,m}$.) Hence, the average BER constraint under (oc-1)–(oc-3) can be expressed as: $k=1,\ldots,K$,

$$\bar{\epsilon}_k = \frac{\sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})\epsilon_{k,m}(h_k p_{k,m})dF(\mathbf{h})}{\sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})dF(\mathbf{h})} \leq \check{\epsilon}_k. \tag{13}$$

To simplify (13), we will replace the denominator with $\check{R}_k$ from (12). This could render the solution of the ensuing optimization problem more conservative (if the rate constraints in (12) were not met tightly), because we impose stricter average BER constraints. With this replacement, the constrained optimization we seek to solve in this section is (14), at the bottom of the page.

Clearly, if regions $\boldsymbol{\mathcal{H}}$ are given *a priori* (using e.g., uniform, Lloyd or entropy-based quantizers per h_k), then the problem variables reduce from $(\boldsymbol{\mathcal{H}}, \boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h}))$ to $(\boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h}))$.

In what follows, we will first solve (14) w.r.t. $\boldsymbol{\tau}(\mathbf{h})$ when $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{p}$ are given (Section IV-A). Part of this user time allocation sub-problem will be all one needs for the *on-line* Q-CSIT based optimal resource allocation that utilizes the optimal power-book $\boldsymbol{p}^*$ and possibly the optimal quantization regions $\boldsymbol{\mathcal{H}}^*$, both of which are obtained *off-line*. Construction of $\boldsymbol{p}^*$ (Section IV-B) and $\boldsymbol{\mathcal{H}}^*$ (Section IV-C) will also rely on part of the time allocation solution. For their off-line optimization we will follow a block-coordinate approach where during the $(t+1)$st block iteration two of the three vectors $(\boldsymbol{\mathcal{H}}^{(t)}), \boldsymbol{p}^{(t)}, \boldsymbol{\tau}(\mathbf{h})^{(t)})$ from the (t)th iteration will be used to obtain the third one for the $(t+1)$st iteration, until convergence (Section IV-D).

A. Optimal User-Time Allocation

Here we solve (14) for $\boldsymbol{\tau}(\mathbf{h})$ when $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{p}$ are given. The resultant solution will be useful in three cases: (i) for resource allocation when $\boldsymbol{\mathcal{H}} = \check{\boldsymbol{\mathcal{H}}}$ and/or $\boldsymbol{p} = \check{\boldsymbol{p}}$ are prescribed by the system setup; (ii) for the on-line phase where $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}^*$ and $\boldsymbol{p} = \boldsymbol{p}^*$ are fixed to their optimal values obtained in the off-line phase; and (iii) for the off-line phase to obtain $\boldsymbol{\tau}^{(t+1)}(\mathbf{h})$ when $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}^{(t)}$ and $\boldsymbol{p} = \boldsymbol{p}^{(t)}$ are available from the previous block-coordinate iteration.

Since $\boldsymbol{\mathcal{H}}$ is known, for a given realization $\mathbf{h} \in \mathcal{H}_{k,m}$ the mode $m_k(\mathbf{h})$ for the transmit rate $\rho_{k,m_k(\mathbf{h})}$ and power $p_{k,m_k(\mathbf{h})}$, if user k is selected, are also known [cf. (oc-3)]. As in Section III, the dual function evaluated at the optimum value of the multipliers can be written as $g(\boldsymbol{\lambda}^{Q_\tau *}, \boldsymbol{\mu}^{Q_\tau *}) = \sum_{k=1}^K\sum_{m=1}^{M_k}\int_{\mathcal{H}_{k,m}}\varphi_k^{Q_\tau}(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *})\tau_k^*(\mathbf{h})dF(\mathbf{h}) + \sum_{k=1}^K(\lambda_k^{Q_\tau *}\check{R}_k - \mu_k^{Q_\tau *}\check{\epsilon}_k)$, where the instantaneous cost function

$$\begin{aligned}\varphi_k^{Q_\tau}\left(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *}\right) &:= w_k p_{k,m_k(\mathbf{h})} - \lambda_k^{Q_\tau *}\rho_{k,m_k(\mathbf{h})} \\ &\quad + \mu_k^{Q_\tau *}\rho_{k,m_k(\mathbf{h})}\epsilon_{k,m_k(\mathbf{h})}\left(h_k p_{k,m_k(\mathbf{h})}\right)/\check{R}_k\end{aligned} \tag{15}$$

depends on the optimum Lagrange multiplier $\lambda_k^{Q_\tau *}$ $(\mu_k^{Q_\tau *})$ corresponding to the kth average rate (BER) constraint. Upon defining the minimizing user index for this cost as $k^* := \arg\min_k \varphi_k^{Q_\tau}(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *})$, and repeating the arguments we followed to derive (8) and (9) we find, with

$$\begin{cases} \min_{\boldsymbol{\mathcal{H}}, \boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h})} & \sum_{k=1}^K w_k \sum_{m=1}^{M_k} p_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})dF(\mathbf{h}) \\ \text{s. to} & \sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})dF(\mathbf{h}) \geq \check{R}_k,\ k=1,\ldots,K; \\ & \sum_{m=1}^{M_k}\rho_{k,m}\int_{\mathcal{H}_{k,m}}\tau_k(\mathbf{h})\epsilon_{k,m}(h_k p_{k,m})dF(\mathbf{h})/\check{R}_k \leq \check{\epsilon}_k,\ k=1,\ldots,K; \\ & 0 \leq \sum_{k=1}^K \tau_k(\mathbf{h}) \leq 1,\ \forall \mathbf{h}. \end{cases} \tag{14}$$

$\boldsymbol{\lambda}^{Q_\tau *} := [\lambda_1^{Q_\tau *} \dots \lambda_K^{Q_\tau *}]^T$ and $\boldsymbol{\mu}^{Q_\tau *} := [\mu_1^{Q_\tau *} \dots \mu_K^{Q_\tau *}]^T$ given, that

$$\text{If } \varphi_{k^*}^{Q_\tau}\left(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *}\right) < 0, \text{ then } \tau_{k*}^*(\mathbf{h}) = 1, \quad \text{and} \quad \tau_k^*(\mathbf{h}) = 0 \text{ for } k \neq k^* \tag{16}$$

or, the trivial case where all users defer:

$$\text{If } \varphi_{k^*}^{Q_\tau}\left(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *}\right) = 0, \text{ then } \tau_k^*(\mathbf{h}) = 0 \,\forall\, k. \tag{17}$$

The time allocations in (16) and (17) clearly satisfy the constraint $\sum_{k=1}^K \tau_k(\mathbf{h}) \in [0,1]$; and since either one holds true per $\mathbf{h}$ they also minimize the average transmit power in (14) provided we have a means of obtaining the optimum Lagrange multiplier vectors $\boldsymbol{\lambda}^{Q_\tau *}$ and $\boldsymbol{\mu}^{Q_\tau *}$.

Based on $F(\mathbf{h})$, the wanted multipliers are computed off-line to satisfy the average rate and BER constraints $\forall k = 1, \dots, K$ through the sub-gradient iterations (18) and (19) [cf. (10)].

$$\lambda_k^{Q_\tau}(i+1) = \left[\lambda_k^{Q_\tau}(i) + \beta_\lambda(\check{R}_k - \sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k^*(i;\mathbf{h}) dF(\mathbf{h}))\right]^+ \tag{18}$$

$$\mu_k^{Q_\tau}(i+1) = \left[\mu_k^{Q_\tau}(i) + \beta_\mu \left(\sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k^*(i;\mathbf{h}) \times \epsilon_{k,m}(h_k p_{k,m}) dF(\mathbf{h})/\check{R}_k - \check{\epsilon}_k\right)\right]^+. \tag{19}$$

As in Section III, recursions (18) and (19) are guaranteed to converge to the optimum pair $(\boldsymbol{\lambda}^{Q_\tau *}, \boldsymbol{\mu}^{Q_\tau *})$ starting from any initial condition because the problem (14) is convex in the variables $\{\tau_k(\mathbf{h})\}_{k=1}^K$.

Similar to $E_{\mathbf{h}}$ in (10), the integrals in (18) and (19) are replaced in practice by sums over realizations $\{\mathbf{h}^{(n)}\}_{n=1}^N$. During iteration (i), if $h_k^{(n)}$ falls in region $m(i; h_k^{(n)})$ the multipliers $\lambda_k^{Q_\tau}(i-1)$ and $\mu_k^{Q_\tau}(i-1)$ are used in (15) to compute the winner terminal $k^*(i; \mathbf{h}^{(n)})$ which contributes a summand with known rate $\rho_{k^*(i;\mathbf{h}^{(n)}),m(i;h_k^{*(n)})}$ to the integral (now sum) in (18) and a summand with known power $p_{k^*(i;\mathbf{h}^{(n)}),m(i;h_k^{*(n)})}$ to the integral (now sum) in (19). Notice that the two (sub-)gradient updates per user should be run and terminated jointly. Again, we stop the iterations by checking the relative difference $|g(\boldsymbol{\lambda}^{Q_\tau}(i+1), \boldsymbol{\mu}^{Q_\tau}(i+1)) - g(\boldsymbol{\lambda}^{Q_\tau}(i), \boldsymbol{\mu}^{Q_\tau}(i))|/|g(\boldsymbol{\lambda}^{Q_\tau}(i+1), \boldsymbol{\mu}^{Q_\tau}(i+1))|$ of the dual function (defined after (14)) against a preselected tolerance; and if smaller, then we return $\lambda_k^{Q_\tau *} = \lambda_k^{Q_\tau}(i+1)$ and $\mu_k^{Q_\tau *} = \mu_k^{Q_\tau}(i+1)\ \forall k$.

With the nonnegative multipliers $\{\lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *}\}_{k=1}^K$ computed off-line using the channel gain CDF, for each $\mathbf{h}$ realization during online operation the access point only needs to: (i) select per user k from the given quantization and power books $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{p}$ the region $\mathcal{H}_{k,m}$ that h_k falls into, and thus the rate $\rho_{k,m_{k(\mathbf{h})}}$ and corresponding power $p_{k,m_{k(\mathbf{h})}}$; (ii) evaluate the cost in (15) for all $k = 1, \dots, K$ to select the winner (minimum cost) terminal k^*; and (iii) feedback $[\mathbf{c}^T(k^*; \mathbf{h}), \mathbf{c}^T(m_{k^*}; \mathbf{h})]^T$ for the winner terminal to select rate and power from its known quantization and power books (or for all users to defer if the trivial case in (17) is active). Summarizing, we have proved that:

Proposition 2: Under (oc-1)–(oc-3), if $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{p}$ are given, then Q-CSIT based almost surely optimal user-time allocation is uniquely given by (16) or (17) and is solely determined by the cost $\varphi_{k^*}^{Q_\tau}(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *})$ in (15). The optimum Lagrange multiplier vectors $\boldsymbol{\lambda}^{Q_\tau *}$ and $\boldsymbol{\mu}^{Q_\tau *}$ are computable at the access point off-line using the channel gain CDF; while the on-line operation requires low-overhead feedback of $\mathbf{c}(\mathbf{h}) = [\mathbf{c}^T(k^*), \mathbf{c}^T(m_{k^*})]^T$ carrying $\lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits from the access point to the terminals.

Except for the difference in the cost, it is worth noting that when the quantization regions and the power book are specified the optimal user-time allocation based on Q-CSIT ends up being also greedy, similar to the P-CSIT based one we saw in Proposition 1. The nonpositive cost $\varphi_k^{Q_\tau}(\mathbf{h}, \lambda_k^{Q_\tau *}, \mu_k^{Q_\tau *})$ can be viewed as a link quality indicator of user k (the smaller the better), based on which at most one (minimum-cost) user is allowed to transmit per frame. If there are multiple users $\{k_j^*\}_{j=1}^J, J \leq K$, attaining the same minimum cost $\varphi_{k_j^*}^{Q_\tau}$, arbitrary time sharing of the frame or assignment of the entire frame to one of them at random will be equally optimal. The dependence of $\varphi_k^{Q_\tau}$ on the channel h_k but also on the multipliers $\lambda_k^{Q_\tau *}$ and $\mu_k^{Q_\tau *}$ shows that the terminal with "best" link quality actually incurs the smallest net cost in terms of the fulfilled average power and BER requirements minus the average rate it is rewarded. As with P-CSIT where all users defer when $\mathbf{h} \approx \mathbf{0}$, the deep-fading condition with Q-CSIT corresponds to having $\varphi_k^{Q_\tau}(\mathbf{h}) \geq 0, \forall k$. Because we seek to minimize average power under average rate and BER constraints, letting users to transmit during deep fades only gains small rate rewards at high power and BER costs; i.e., the intuition behind the optimal user-time allocation solution in Proposition 2 is to save transmit power for better channel instantiations which entail smaller net costs with higher rewards, and thus gain power efficiency.

Remark 3: We already pointed out in Remark 1 the differences between this paper's allocation schemes and the P-CSIT based ones in [15], [10] with regards to the operating conditions and the feedback overhead. Albeit greedy, our optimal scheduling policies based on either P-CSIT or Q-CSIT are also fair in the sense that they satisfy the average individual rate and BER requirements even if at most one winner terminal takes all resources per channel realization. In addition to $\{\check{R}_k, \check{\epsilon}_k\}_{k=1}^K$, fairness is imposed explicitly through the weights $\{w_k\}_{k=1}^K$ employed in the objective function. Indeed, if average transmit power is more critical for terminal k, it suffices to assign to it a larger w_k so that the optimal solution weighs more reduction of its power consumption. If on the other hand transmit rate is more critical for terminal k, its average rate requirement $\check{R}_k$ would be naturally higher and the optimal policy will fairly present more chances for terminal k to transmit even if it has worse average SNR than others. In a nutshell, the optimal user-time allocations of Propositions 1 and 2 not only minimize average transmit power but are also fair by construction.

B. Off-Line Construction of the Optimal Transmit-Power Book

In this subsection, we are given quantization regions $\boldsymbol{\mathcal{H}}$ (prescribed as $\boldsymbol{\mathcal{H}} = \check{\boldsymbol{\mathcal{H}}}$ or provided by the previous block iteration as $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}^{(t-1)}$) and the optimal user-time mapping $\boldsymbol{\tau}(\mathbf{h})$ obtained as in the previous subsection. The goal is to solve off-line (14) w.r.t. $\boldsymbol{p}$ in order to obtain the transmit-power book $\boldsymbol{p}^*$, when all other variables have been optimized, or to produce $\boldsymbol{p}^{(t)}$ for the next block iteration.

Because transmit-power variables under Q-CSIT are optimized independently from transmit-rate modes which are coupled with quantization regions [cf. (oc-3)], the average rate constraints are not present here. The same is true for the time fraction constraints since $\boldsymbol{\tau}(\mathbf{h})$ is given. Hence, the optimal power allocation problem reduces to

$$\begin{cases} \min_{\boldsymbol{p}} & \sum_{k=1}^{K} w_k \sum_{m=1}^{M_k} p_{k,m} \int\limits_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s. to} & \sum_{m=1}^{M_k} \rho_{k,m} \int\limits_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \epsilon_{k,m}(h_k p_{k,m}) dF(\mathbf{h}) / \check{R}_k \leq \check{\epsilon}_k \\ & k = 1, \ldots, K. \end{cases} \tag{20}$$

Since $\check{\epsilon}_k \geq 0\ \forall k$ and we can always increase $p_{k,m}$ sufficiently to render $\epsilon_{k,m}(h_k p_{k,m})$ arbitrarily small, it is possible to meet any prescribed average BER constraint; thus, the minimization problem in (20) is always feasible. In fact, with the objective being convex in $p_{k,m}$ and likewise for the constraint (since $\epsilon_{k,m}$ is a convex function of the transmit-power) the problem in (20) is also convex. This implies that similar to the previous subsection fast algorithms are available to find its unique global optimum.

To this end, we again follow the Lagrange multiplier method and equate the derivative of the Lagrangian $L(p_{k,m}, \mu_k^{Q_p})$ w.r.t. $p_{k,m}$ to zero, to obtain $\forall k$ the KKT conditions

$$\mu_k^{Q_p *}(\rho_{k,m}/\check{R}_k) \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \dot{\epsilon}_{k,m}(h_k p_{k,m}^*) dF(\mathbf{h}) + w_k \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) - \nu_{k,m}^{Q_p *} = 0 \tag{21}$$

where: $\dot{\epsilon}_{k,m}$ stands for the derivative of the BER function w.r.t. $p_{k,m}$; $\mu_k^{Q_p *} \geq 0$ denotes the optimum multiplier associated with the kth BER constraint; and $\nu_{k,m}^{Q_p *} \geq 0$ is the same for the (k,m)th implicit constraint $p_{k,m} \geq 0$. Since (20) is convex, strong duality holds and the KKT conditions are sufficient and necessary for global optimality [2]. Interestingly, when the set $\{\tau_k(\mathbf{h})\}_{k=1}^K$ is given, transmit-power optimization is decoupled across users. Thus, solving (20) is equivalent to solving K small problems; i.e., $\forall k$, it suffices to: $\min_{\boldsymbol{p}_k} w_k \sum_{m=1}^{M_k} p_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h})$ subject to $\sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \epsilon_{k,m}(h_k p_{k,m}) dF(\mathbf{h}) / \check{R}_k \leq \check{\epsilon}_k$.

Furthermore, we prove in the Appendix that i) at the optimum $p_{k,m}^*$ we have $\mu_k^{Q_p *} > 0\ \forall k$, which implies that all the average BER constraints are satisfied as strict equalities; ii) if $\int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) > 0$, then $p_{k,m}^* > 0$ in which case $\nu_{k,m}^{Q_p *} = 0$; and iii) if $\int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = 0$, the optimum transmit power for certain (k,m) pair(s) is $p_{k,m}^* = 0$, in which case $\nu_{k,m}^{Q_p *} > 0$. We will henceforth exclude case iii) by just removing zero-power AMC modes from the given $\boldsymbol{\mathcal{H}}$ and reformulating (20) with the more compact $\boldsymbol{\mathcal{H}}$ containing AMC modes with nonzero optimum power values. With this reduction, we ensure that $\nu_{k,m}^{Q_p *} = 0\ \forall k$; but since $\mu_k^{Q_p *} > 0\ \forall k$, we can simplify (21) and seek $p_{k,m}^*$ as the solution of the nonlinear equation

$$\dot{L}\left(p_{k,m}, \mu_k^{Q_p}\right) = \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \dot{\epsilon}_{k,m}(h_k p_{k,m}) dF(\mathbf{h}) + \frac{w_k \check{R}_k}{\rho_{k,m} \mu_k^{Q_p}} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = 0 \tag{22}$$

where the second summand is known so long as the multiplier $\mu_k^{Q_p} > 0$ is available ($\dot{L}$ denotes derivative of L w.r.t. $p_{k,m}$).

Multiplier $\mu_k^{Q_p *}$ can be determined by satisfying the average BER constraint as a strict equality; i.e., upon defining

$$f\left(\boldsymbol{p}_k\left(\mu_k^{Q_p}\right)\right) := -\check{\epsilon}_k + \sum_{m=1}^{M_k} \frac{\rho_{k,m}}{\check{R}_k} \int\limits_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \epsilon_{k,m}\left[h_k p_{k,m}\left(\mu_k^{Q_p}\right)\right] dF(\mathbf{h}) \tag{23}$$

we need to find for each k the root of the nonlinear equation $f(\boldsymbol{p}_k(\mu_k^{Q_p})) = 0$. Note that in order to stress the dependence of $p_{k,m}$ on the multiplier $\mu_k^{Q_p}$ in (22), we explicitly wrote the powers in (23) as $\boldsymbol{p}_k(\mu_k^{Q_p}) := [p_{k,1}(\mu_k^{Q_p}), \ldots, p_{k,M_k}(\mu_k^{Q_p})]^T$. Because $\epsilon_{k,m}$ is convex and monotonic, the roots of $\dot{L}(p_{k,m}, \mu_k^{Q_p})$ and $f(\boldsymbol{p}_k(\mu_k^{Q_p}))$ in (22) and (23) can be found efficiently with scalar sub-gradient iterations which will have guaranteed convergence to the unique pair $(\boldsymbol{p}_k^*, \mu_k^{Q_p *})$ even with arbitrary initialization.

Specifically, consider the ith iteration and suppose that $\mu_k(i)$ is available. Using it in (22) instead of $\mu_k^{Q_p *}$, we iterate the recursion

$$p_{k,m}(j_i + 1) = [p_{k,m}(j_i) + \beta_p \dot{L}(p_{k,m}(j_i), \mu_k(i))]^+ \tag{24}$$

until convergence, and set $p_{k,m}(i)$ equal to the limit. Based on $p_{k,m}(i)$, we next update the multiplier to closer satisfy (23) using

$$\mu_k(i+1) = [\mu_k(i) + \beta_\mu f(\boldsymbol{p}_k(\mu_k(i))]^+ \tag{25}$$

and then go back to rerun (24) with $i+1$ replacing i to find $p_{k,m}(i+1)$, and so on.

The stopping rules in these two nested recursions are similar to those in the previous subsection. As detailed before, the integrals involved in (22) and (23) are replaced in practice with averages over channel gain realizations, drawn from the CDF $F(\mathbf{h})$, and the approximation can be made arbitrarily accurate since all these are generated and computed off-line. Especially for the power allocation optimization where decoupling across users allows for one-dimensional iterations per user, it is possible instead of the sub-gradient updates (24) and (25) to resort to simpler e.g., bisection based alternatives which also exhibit fast convergence (geometric in the error) while bypassing the need to find appropriate stepsizes β_p and β_μ.

Summing up the results of this subsection we have:

Proposition 3: Under (oc-1)-(oc-3), if $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\tau}(\mathbf{h})$ are given, then Q-CSIT based power allocation is provided by the unique global optimum of (20). The optimal $p^*_{k,m}$ for each (k,m) pair is either $p^*_{k,m}=0$, or, the unique positive root of (22) which is computed efficiently (along with the associated optimum Lagrange multiplier) using a pair of nested one-dimensional iterations off-line.

Per user k, Proposition 3 asserts what could be interpreted as quantized power water-filling across the quantization regions $\{\mathcal{H}_{k,m}\}_{m=1}^{M_k}$. Indeed, as $\epsilon_{k,m}(h_k p_{k,m})$ is convex and monotonic, it is clear for the optimal power $p^*_{k,m}$ that: i) it increases as $\rho_{k,m}$ increases [cf. (22)]; and ii) it decreases as $\check{\epsilon}_k$ increases [cf. (23)].

C. Off-Line Construction of the Optimal Quantizer

Here the user-time allocation mapping $\boldsymbol{\tau}(\mathbf{h})$ and the power book $\boldsymbol{p}$ obtained as described in the last two subsections are given, and the goal is to optimize (14) w.r.t. $\boldsymbol{\mathcal{H}}$. Clearly, this step is not necessary when the quantization regions are available beforehand as $\boldsymbol{\mathcal{H}}=\check{\boldsymbol{\mathcal{H}}}$. It is present only in the optimization of the joint allocation-quantization task to obtain $\boldsymbol{\mathcal{H}}=\boldsymbol{\mathcal{H}}^{(t)}$ when $\boldsymbol{\tau}(\mathbf{h})=\boldsymbol{\tau}^{(t-1)}(\mathbf{h})$ and $\boldsymbol{p}=\boldsymbol{p}^{(t-1)}$ are available from the previous block iteration. Upon convergence of the block iteration, the off-line solution of this subsection will return the optimal book of quantization regions $\boldsymbol{\mathcal{H}}^*$ for use during the on-line operation.

Since $\boldsymbol{\tau}(\mathbf{h})$ is given, the time fraction constraints in (14) are not present here. And because (oc-3) entails K quantization problems decoupled across users, the pertinent optimization problem per user k is to

$$\begin{cases} \min_{\{\mathcal{H}_{k,m}\}_{m=1}^{M_k}} \; w_k \sum_{m=1}^{M_k} p_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \\ \text{s. to} \quad \sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) \geq \check{R}_k; \\ \quad \sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \epsilon_{k,m}(h_k p_{k,m}) dF(\mathbf{h})/\check{R}_k \leq \check{\epsilon}_k. \end{cases} \tag{26}$$

Notice that the quantization regions define regions of integration in (26). Since (26) is generally nonconvex w.r.t. these limits of integration, similar to *all* quantizer design problems, guaranteeing the global optimum is challenging and finding it with efficient algorithms is generally impossible.

Nonetheless, pursuing again a Lagrange multiplier approach, we form the Lagrangian $L(\mathcal{H}_{k,m}, \lambda_k^{Q_{\mathcal{H}}*}, \mu_k^{Q_{\mathcal{H}}*}) = \sum_{m=1}^{M_k} \varphi_{k,m}^{Q_{\mathcal{H}}}(h_k) \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) + \lambda_k^{Q_{\mathcal{H}}*} \check{R}_k - \mu_k^{Q_{\mathcal{H}}*} \check{\epsilon}_k$, where $\lambda_k^{Q_{\mathcal{H}}*}$ and $\mu_k^{Q_{\mathcal{H}}*}$ denote respectively the optimum multipliers corresponding to the average rate and BER constraints, and the instantaneous cost function is given by: for $m=1,\ldots,M_k$,

$$\varphi_{k,m}^{Q_{\mathcal{H}}}(h_k) := w_k p_{k,m} - \lambda_k^{Q_{\mathcal{H}}*} \rho_{k,m} + \mu_k^{Q_{\mathcal{H}}*} \frac{\rho_{k,m}}{\check{R}_k} \epsilon_{k,m}(h_k p_{k,m}) \tag{27}$$

with $\varphi_{k,0}^{Q_{\mathcal{H}}}(h_k) := 0$. To find the regions minimizing this cost, we should clearly assign each realization $\mathbf{h}$ to $\mathcal{H}^*_{k,m}$ if and only if its kth entry satisfies $\varphi_{k,m}^{Q_{\mathcal{H}}}(h_k) \leq \varphi_{k,m'}^{Q_{\mathcal{H}}}(h_k)\, \forall m' \neq m, m' = 0, 1, \ldots, M_k$. But this readily yields the quantization regions optimizing (26) for user k as

$$\mathcal{H}^*_{k,m} = \left\{ \mathbf{h} : \varphi_{k,m}^{Q_{\mathcal{H}}}(h_k) \leq \varphi_{k,m'}^{Q_{\mathcal{H}}}(h_k); \forall m' \neq m \right\}. \tag{28}$$

These regions can be constructed as soon as the optimum nonnegative multipliers $(\lambda_k^{Q_{\mathcal{H}}*}, \mu_k^{Q_{\mathcal{H}}*})$ involved in the cost are found (note that $\varphi_{k,m}^{Q_{\mathcal{H}}}(h_k)$ depends on $\lambda_k^{Q_{\mathcal{H}}*}$ and $\mu_k^{Q_{\mathcal{H}}*}$, and therefore $\mathcal{H}^*_{k,m}$ in (28) are in fact functions of $\lambda_k^{Q_{\mathcal{H}}*}$ and $\mu_k^{Q_{\mathcal{H}}*}$). If $\lambda_k^{Q_{\mathcal{H}}*} = \mu_k^{Q_{\mathcal{H}}*} = 0$ in (27), then no extra computation is needed since (28) is directly implementable, and the constraints are satisfied as strict inequalities (cf. the complementary slackness [2]).

But if $\lambda_k^{Q_{\mathcal{H}}*} > 0$ and $\mu_k^{Q_{\mathcal{H}}*} > 0$, then the corresponding constraints are satisfied at the optimum as equalities. This implies that writing the dual function in (26) as $g(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}}) := L(\mathcal{H}^*_{k,m}(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}}), \lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}})$, the optimum multipliers can be found as roots of the nonlinear (29) and (30)

$$\dot{g}_\lambda\left(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}}\right) = \check{R}_k - \sum_{m=1}^{M_k} \rho_{k,m} \times \int_{\mathcal{H}_{k,m}(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}})} \tau_k(\mathbf{h}) dF(\mathbf{h}) = 0 \tag{29}$$

$$\dot{g}_\mu\left(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}}\right) = \sum_{m=1}^{M_k} \rho_{k,m} \int_{\mathcal{H}_{k,m}(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}})} \tau_k(\mathbf{h}) \times \epsilon_{k,m}(h_k p_{k,m}) dF(\mathbf{h})/\check{R}_k - \check{\epsilon}_k = 0 \tag{30}$$

where $\dot{g}_\mu$ ($\dot{g}_\lambda$) denotes the partial derivative of g w.r.t. μ (respectively λ), and the dependence of the quantization regions on the wanted Lagrange multipliers is indicated in the limits of the integrals. To solve these equations one could be tempted to use nested iterations similar to those in (24) and (25). Unfortunately, since the nonlinear functions now are not guaranteed to be convex, sub-gradient iterations can only assure convergence to a (possibly nonunique) local optimum which also depends on the chosen initialization.

For this reason, we instead advocate to solve (29) and (30) by searching exhaustively over the two-dimensional space $(\lambda_k^{Q_{\mathcal{H}}}, \mu_k^{Q_{\mathcal{H}}})$. For each candidate pair of multipliers, the integrals involved are evaluated as discussed in the previous subsections by generating realizations $\{\mathbf{h}^{(n)}\}$ drawn from the CDF $F(\mathbf{h})$. Even though a two-dimensional search is not as efficient as a sub-gradient iteration or a line search, it is still manageable since it is performed off-line and yields the optimum solution regardless of the initialization. One more feature of this particular search is that it can be terminated (very) early in some cases. Indeed, as soon as we find multipliers at which the nonlinear function values satisfy $|\dot{g}_\lambda| < \varepsilon$ and $|\dot{g}_\mu| < \varepsilon$ for a preselected tolerance ε, we stop searching. At this point we should also reflect back on (oc-3) to appreciate the importance of decoupling the design of quantization regions across users. Had we jointly optimized $\mathcal{H}_{k,m}$ across all (k,m) pairs, the need would arise to search over a $2K$-dimensional space for obtaining the associated optimum Lagrange multiplier vectors

which is computationally more burdensome when compared to the K two-dimensional searches involved under (oc-3).

To recap the results of this subsection, we have established that:

Proposition 4: Under (oc-1)-(oc-3), if $\boldsymbol{\tau}(\mathbf{h})$ and $\boldsymbol{p}$ are fixed, the M_k+1 quantization regions optimizing (26) are given by (28) for each user $k=1,\ldots,K$. Computing the optimal Lagrange multipliers $(\lambda_k^{Q_{\mathcal{H}^*}}, \mu_k^{Q_{\mathcal{H}^*}})$ needed to obtain these optimal regions amounts to solving (29) and (30) per user, which requires a total of K two-dimensional exhaustive searches that are computed off-line.

As we will also confirm by simulations, Proposition 4 yields for each terminal k a set of nonoverlapping consecutive intervals, i.e., $\{\mathcal{H}_{k,m} \equiv [H_{k,m}, H_{k,m+1})\}_{m=0}^{M_k}$, by assigning h_k to the region minimizing the Lagrangian of (26). Intuitively, $\lambda_k^{Q_{\mathcal{H}}*}$ can be interpreted as the utility of each transmitted bit and $\mu_k^{Q_{\mathcal{H}}*}$ as the cost of each erroneously received bit. With this interpretation, the Lagrangian function of assigning $\mathbf{h}$ to the region $\mathcal{H}_{k,m}$ represents the average of the net cost $\varphi_{k,m}^{Q_{\mathcal{H}}}(h_k)$ which takes into account transmit power and BER expenditures minus transmit-rate rewards. Note also that if $\mathbf{h} \in \mathcal{H}_{k,0}^*$ the kth user should clearly defer because the cost of assigning h_k to an active region $(m_k > 0)$ is higher than the cost of assigning it to the inactive region $(m_k = 0)$ which incurs zero cost. A remark is now in order on the relationship of this subsection's design with the vast literature on vector quantization; see e.g., [4].

Remark 4: Let us consider the distortion metric

$$d(\mathbf{h}, \mathcal{H}_{k,m}) := w_k p_{k,m} \tau_k(\mathbf{h}) - \lambda_k^{Q_{\mathcal{H}}*} \rho_{k,m} \tau_k(\mathbf{h}) + \mu_k^{Q_{\mathcal{H}}*} \rho_{k,m} \epsilon_{k,m}(h_k p_{k,m}) \tau_k(\mathbf{h}) / \check{R}_k \quad (31)$$

which coincides with $\varphi_{k,m}^{Q_{\mathcal{H}}*}(h_k)$ in (27) when $\tau_k(\mathbf{h}) = 1$, and equals 0 when $\tau_k(\mathbf{h}) = 0$. (Recall that in the optimum user-time allocation $\tau_k(\mathbf{h}) = 1$ or 0.) Based on (31), we can view minimization of the Lagrangian as minimizing the *average* distortion metric

$$E_{\mathbf{h}}[d(\mathbf{h}, \mathcal{H}_{k,m})] = \sum_{m=1}^{M_k} \int_{\mathcal{H}_{k,m}} \varphi_{k,m}^{Q_{\mathcal{H}}}(h_k) \tau_k(\mathbf{h}) dF(\mathbf{h}). \quad (32)$$

This in turn implies that finding $\mathcal{H}_{k,m}^*$ as in (28) can be interpreted as a *nearest-neighbor rule* [4] with the nonstandard metric in (31), according to which

$$\mathbf{h} \in \mathcal{H}_{k,m}, \quad \text{iff } d(\mathbf{h}, \mathcal{H}_{k,m}) \leq d(\mathbf{h}, \mathcal{H}_{k,m'}),\ \forall m' \neq m. \quad (33)$$

Casting the design of single-user multi-antenna systems with limited-rate feedback in a vector quantization framework has gained popularity recently; and variants of Lloyd's algorithm [11] have been put forth in this context; see e.g., [5], [19] and references therein. The main contribution of Proposition 4 to quantization is the adoption of a distortion metric minimizing weighted average transmit power in a TDMA setup where users rely on AMC modes that are coupled with the quantization regions per user [cf. (oc-3)]. The impact of this coupling in practice is major since it allows online usage of the quantizer designed off-line with low overhead in the feedback channel.

D. Joint Allocation-Quantization Algorithm

In this subsection, we combine our results from Sections IV-A–C to tackle the joint resource allocation-quantization (JRAQ) problem in (14). The resultant JRAQ block-coordinate descent algorithm minimizes the global objective function $J(\boldsymbol{\tau}(\mathbf{h}), \boldsymbol{p}, \boldsymbol{\mathcal{H}}) := \sum_{k=1}^{K} w_k \sum_{m=1}^{M_k} p_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h})$, by fixing two of the three sets of variables $(\boldsymbol{\tau}(\mathbf{h}), \boldsymbol{p}, \boldsymbol{\mathcal{H}})$ and minimizing w.r.t. the third one as summarized next:

Algorithm 1: *JRAQ*: Produce using the CDF $F(\mathbf{h})$ initial quantization regions $\boldsymbol{\mathcal{H}}^{(0)}$ and transmit-power variables $\boldsymbol{p}^{(0)}$ which are feasible. Select tolerance $\varepsilon > 0$, initialize objective at $J^{(0)} = \infty$ and set the iteration index $t = 1$.

- J1) Given $\boldsymbol{\mathcal{H}}^{(t-1)}$ and $\boldsymbol{p}^{(t-1)}$, obtain $\boldsymbol{\tau}^{(t)}(\mathbf{h})$ from Proposition 2.
- J2) Given $\boldsymbol{\tau}^{(t)}(\mathbf{h})$ and $\boldsymbol{\mathcal{H}}^{(t-1)}$, obtain $\boldsymbol{p}^{(t)}$ from Proposition 3.
- J3) Given $\boldsymbol{p}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\mathbf{h})$, obtain $\boldsymbol{\mathcal{H}}^{(t)}$ from Proposition 4.
- J4) *Stopping criterion*: Calculate the objective $J^{(t)}$ using $\boldsymbol{\mathcal{H}}^{(t)}$, $\boldsymbol{p}^{(t)}$ and $\boldsymbol{\tau}^{(t)}(\mathbf{h})$. If $|(J^{(t)} - J^{(t-1)})/J^{(t)}| < \varepsilon$, return the (t)th resource allocation and quantization variables and stop. Otherwise, increase t by 1 and go to *J1*).

In each of the four steps J1–J4, the global objective J is guaranteed not to increase and is lower bounded by the corresponding optimal P-CSIT based solution. In addition, Propositions 2–4 assert that iterations in each step will converge to the unique global optimum of each individual sub-problem and these solutions can be found analytically (but not in closed form except for $\boldsymbol{\tau}^*(\mathbf{h})$). Hence, we are ensured that the JRAQ algorithm will converge in a finite number of iterations at least to a local minimum [1, Theorem 2.7.1].

One reason that the block-coordinate descent iterations of JRAQ may not reach the global optimum is the fact that the minimization problem in (14) is nonconvex w.r.t. the quantization region variables $\boldsymbol{\mathcal{H}}$. Of course, this lack of convexity is inherent to all optimization problems dealing with vector quantization including the celebrated Lloyd algorithm [11]; i.e., it is not unique to our formulation. Nevertheless, in accordance with the widespread success of Lloyd's algorithm in practice, our simulations too confirm that the resulting Q-CSIT based JRAQ algorithm always achieves power efficiency close to the optimal P-CSIT solution that we developed in Section IV-A. And since P-CSIT is the limit of Q-CSIT as the number of feedback bits grows large, the P-CSIT solution lower bounds the Q-CSIT based one. This then corroborates that our JRAQ algorithm indeed attains near-global optimality.

Initialization: Critical to this claim of near-global optimality but also to the speed of convergence is the choice of the initial block variables $\boldsymbol{\mathcal{H}}^{(0)}$ and $\boldsymbol{p}^{(0)}$ needed to start up step J1. As a word of caution, if JRAQ is initialized randomly not only convergence speed may suffer severely but more importantly the block component iterations may fail to yield even a feasible solution. Recognizing the importance of initializing the JRAQ algorithm with a feasible set of variables, our idea is to rely on

the results of the P-CSIT based solution to initialize our Q-CSIT based one.

These considerations prompt us to replace $\varphi_{k,m}^{Q_{\mathcal{H}}}(h_k)$ in (28) with the P-CSIT based cost $\varphi_{k,m}^{P_\tau}(h_k, \lambda_k^{P_\tau *}) := w_k p_{k,m}(h_k) - \lambda_k^{P_\tau *}\rho_{k,m}$ from (7), and initialize the quantization regions per user k using

$$\mathcal{H}_{k,m}^{(0)} = \Big\{ \mathbf{h} :\ \varphi_{k,m}^{P_\tau}\left(h_k, \lambda_k^{P_\tau *}\right) \leq \varphi_{k,m'}^{P_\tau}\left(h_k, \lambda_k^{P_\tau *}\right);\ \forall m' \neq m \Big\}. \tag{34}$$

As we commented earlier, $\lambda_k^{P_\tau *}$ requires only knowledge of the channel gain CDF; and for each h_k, the transmit power also required in $\varphi_{k,m}^{P_\tau}(h_k, \lambda_k^{P_\tau *})$ is computed by inverting (3) as $p_{k,m} = (1/h_k)\epsilon_{k,m}^{-1}(\check{\epsilon}_k, \rho_{k,m})$, where $\rho_{k,m}$ is available from the pool of AMC modes.

To initialize the transmit-power vector, we take a conservative approach and adopt $\forall k = 1, \ldots, K$, $m = 1, \ldots, M_k$

$$p_{k,m}^{(0)} = \max_{\mathbf{h} \in \mathcal{H}_{k,m}^{(0)}} (1/h_k)\epsilon_{k,m}^{-1}(\check{\epsilon}_k, \rho_{k,m}). \tag{35}$$

Because all quantities from the P-CSIT based solution have been derived to meet the average rate (and implicitly also the average BER) constraints, it follows readily that the initialization provided by (34) and (35) yields feasible variables. In fact, since the maximum transmit power is considered in (35), the average BER constraints will be over-satisfied.

Practical Considerations: Having clarified its initialization, we will close this section with practical considerations pertaining to the off-line and on-line operation of the JRAQ algorithm. Given $(\boldsymbol{\lambda}^{Q_\tau *}, \boldsymbol{\mu}^{Q_\tau *})$ in J1, we only need to compare K instantaneous cost functions to find the winner user as in (16). The sub-problems in J2 and J3 are decoupled across users. Given $\mu_k^{Q_p *}$ per user k in J2, we only need to solve a nonlinear problem (22) to obtain the transmit power $p_{k,m}^*$, $m = 1, \ldots, M_k$; while the optimal quantization (28) per user k in J3 can be determined by comparing M_k cost functions in (27). Therefore, the optimization in each step exhibits linear complexity in the number of users K if the optimal Lagrange multipliers are available. Furthermore, for the problems of the form encountered in steps J1 and J2, recent convex optimization solvers are very fast and can reliably find Lagrange multipliers for problems of much larger dimension than those expected in our TDMA setting; see e.g., ([2], Chapter 1) where problems involving 100 or more variables are tackled with affordable complexity, requiring as few as 10–100 sub-gradient iterations. Efficient algorithms are also available in [4] to solve vector quantization problems similar to the one in J3.

As far as the online operation, recall that only part of step J1 is involved. Indeed, with inputs $\boldsymbol{\mathcal{H}}^*$, $\boldsymbol{p}^*$, $\boldsymbol{\lambda}^{Q_\tau *}$ and $\boldsymbol{\mu}^{Q_\tau *}$ all computed off-line, it is possible for the access point to find the most power-efficient user-mode pair and feed back to the terminals the corresponding Q-CSI vector $\mathbf{c}(\mathbf{h})$ per channel realization carrying only $\lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits (cf. Proposition 2). Except for the trivial case (where all users defer) the winner terminal "awarded" the frame also keeps its own transmit-power books $\boldsymbol{p}_{k^*}^*$ and based on the instantaneous feedback it transmits over the scheduled time frame with the scheduled power and AMC rate. (Since the optimal power book $\boldsymbol{p}^* = [\boldsymbol{p}_1^{*T}, \ldots, \boldsymbol{p}_K^{*T}]$ is calculated by the AP and broadcasted to all users during the initialization phase, the assumption that each user k knows its own book $\boldsymbol{p}_k^*$ is not restrictive.) We reiterate that only $\mathbf{c}(\mathbf{h})$ must be fed back per channel realization while recalculation of the books $\boldsymbol{\mathcal{H}}^*$ and $\boldsymbol{p}^*$ is necessary only when the fading channel CDF, user rate requirements or user population changes.

To appreciate the practical merits of the JRAQ algorithm consider a pragmatic example of $K = 5 - 10$ active TDMA users, each supporting $M_k = 6$ different AMC modes $\forall k$, as in the IEEE 802.16 standard. During the online operation of JRAQ, the access point in principle only needs to feed back 5–6 bits per fading state (i.e., coherence time). The overall feedback is included in the UL-MAP message which is encapsulated in a downlink frame to schedule the subsequent uplink frame; see [7, Sec. 6.2.7]. Q-CSIT based feedback operation is also standardized in other systems, e.g., via the data rate control (DRC) channel in CDMA 2000 1xEV-DO; and via the channel quality indicator (CQI) reporting in WCDMA HSPDA. In short, the feedback overhead for on-line operation of the JRAQ algorithm is certainly affordable by most practical systems.

V. Simulated Tests

In this section, we first test the JRAQ algorithm for a two-user ($K = 2$) Rayleigh flat-fading TDMA channel[3] and then check performance with a higher number of users.

The available system bandwidth is $W = 100$ KHz, and the AWGN has two-sided power spectral density N_0 Watts/Hz. The fading channel gains h_k have mean $\bar{h}_k$, $k = 1, 2$, and are assumed uncorrelated. The average signal-to-noise ratio (SNR) for user k is $\bar{\gamma}_k = \bar{h}_k/(N_0 W)$. Unless otherwise specified, we suppose that each user supports $M_1 = M_2 = 3$ quadrature amplitude modulation (QAM) modes, namely BPSK, 8-QAM and 32-QAM; hence, the corresponding (here uncoded) transmission rates are $\rho_{k,m} = 1, 3$, and 5 bits/symbol. The instantaneous BER in this case can be well approximated as [6] $\epsilon_{k,m}(\gamma) = 0.2\exp[-\gamma/(2^{\rho_{k,m}} - 1)]$. In all simulations, the prescribed average BER requirements are $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$.

With $\bar{\gamma}_k = 0$ dB for $k = 1, 2$, we test the P-CSIT based resource allocation scheme of Section III as well as the Q-CSIT based JRAQ algorithm of Section IV-D. To assess the role of optimal user-time and power allocation, we also test a heuristic Q-CSIT based approach, where all users are assigned equal time fractions per frame (i.e., $\tau_k = 1/K\ \forall k, h$) and each terminal transmits with fixed power p_k^* regardless of the AMC rate it adopts. For every candidate p_k and channel gain h_k, the access point selects an AMC mode $m_k(p_k, h_k)$ per user k so that the instantaneous BER meets the prescribed average BER; i.e., so that $\epsilon(h_k p_k, \rho_{k,m_k(p_k,h_k)}) \leq \check{\epsilon}_k$. This choice corresponds to a conservative quantizer since except for the boundaries of the resultant quantization regions, the average BER requirement is always over-satisfied. With such a quantization, each terminal's transmit power is then optimized to also meet the average rate

[3]Note that a two-dimensional channel space $\mathcal{D}$ ($K = 2$) facilitates visualization of e.g., the power region [18], the quantization regions and the user-time allocation. On the other hand, since JRAQ implements a winner-takes-all strategy it collects the maximum multiuser diversity that the channel provides. Hence, $K = 2$ represents a worse case scenario and the validity of the performance claims holds for a higher number of users.

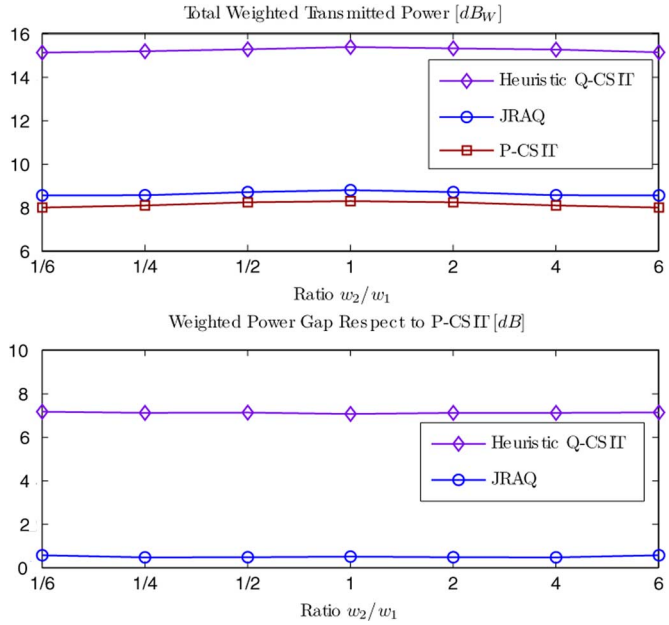


Fig. 1. Total power consumption for different resource allocation schemes with different power weight ratios w_2/w_1 and $\sum_{k=1}^{2} w_k = 1$ when $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\check{R}_1 = \check{R}_2 = 100$ kb/s, and $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB.

constraint; i.e., p_k^* is found using an one-dimensional search (e.g., bisection) as the root of the nonlinear equation $f(p_k) := E_h[(1/K)\rho_{k,m_k(p_k,h_k)}] - \check{R}_k = 0$ (cf. (5) with $\tilde{\tau}_{k,m_k(p_k,h_k)} = 1/K$ and $\tilde{\tau}_{k,m'} = 0\ \forall m' \neq m_k(p_k,h_k)$). Notice that due to its simplicity, the quantization approach of this heuristic scheme is actually widely employed in practical systems with adaptive transmissions, including those in the CDMA2000 1xEVDO and the WCDMA HSDPA standards.

We first consider individual average rate requirements: $\check{R}_1 =$ 100 kb/s and $\check{R}_2 = 100$ kb/s. With different power weights, Fig. 1 (top) depicts the weighted total power consumption for these three schemes; while Fig. 1 (bottom) depicts the performance loss of the two Q-CSIT based TDMA systems w.r.t. the P-CSIT solution in order to gauge the price paid by the limited-rate feedback operation. We observe that: i) the JRAQ algorithm clearly outperforms the heuristic Q-CSIT scheme (yielding around 6 dB savings); and ii) the gap between JRAQ and P-CSIT solutions is very small. Since the P-CSIT solution lower bounds all Q-CSIT based alternatives, this confirms that our block component iterative algorithms are indeed near-globally optimal. Recall that the P-CSIT based scheme is only suitable for TDD operation in relatively slow fading, where the channel can be accurately estimated and channel reciprocity holds; whereas the near-optimal Q-CSIT based one applies to both TDD and FDD systems.

To assess the fundamental limits of average power-efficiency in our Q-CSIT and P-CSIT based TDMA systems, it is possible to define regions where average power vectors must lie for prescribed sets of average rate and BER requirements. If convex, the boundaries of these regions provide the lowest average transmit powers attainable by different weight vectors $\boldsymbol{w} \geq \mathbf{0}$ in our weighted average power minimization framework. Notice that these power regions can be thought of as duals of the capacity regions in [15] and [10].

Given a feasible triplet $(\boldsymbol{\tau}(\mathbf{h}), \boldsymbol{p}, \boldsymbol{\mathcal{H}})$, i.e., a triplet satisfying the constraints in (14), we define the region of achievable average transmit-power vectors $\bar{\mathbf{p}}$ for the Q-CSIT based setup as

$$\bar{\mathcal{P}}^{Q}_{\text{TDMA}}(\boldsymbol{\mathcal{H}}, \boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h})) : = \left\{ \bar{\mathbf{p}} : \bar{P}_k \geq \sum_{m=1}^{M_k} p_{k,m} \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}), \quad \forall k \right\}. \quad (36)$$

For a prescribed pair of vectors $\check{\boldsymbol{r}} := [\check{R}_1, \ldots, \check{R}_K]^T$ and $\check{\boldsymbol{\epsilon}} := [\check{R}_1, \ldots, \check{R}_K]^T$, we take the union of achievable regions in (36) over the feasible set $\mathcal{F}$ of $(\boldsymbol{\tau}(\mathbf{h}), \boldsymbol{p}, \boldsymbol{\mathcal{H}})$ triplets and define the region of average powers in Q-CSIT based TDMA systems as

$$\bar{\mathcal{P}}_{\text{Q-CSIT}}(\check{\boldsymbol{r}}, \check{\boldsymbol{\epsilon}}) := \bigcup_{(\boldsymbol{\mathcal{H}}, \boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h})) \in \mathcal{F}} \bar{\mathcal{P}}^{Q}_{\text{TDMA}}(\boldsymbol{\mathcal{H}}, \boldsymbol{p}, \boldsymbol{\tau}(\mathbf{h})). \quad (37)$$

Likewise, upon replacing the triplet $(\boldsymbol{\tau}(\mathbf{h}), \boldsymbol{p}, \boldsymbol{\mathcal{H}})$ with the pair $(\tilde{\boldsymbol{\tau}}(\mathbf{h}), \boldsymbol{p})$ and the right-hand side of the inequality in (36) with $E_{\mathbf{h}}[\sum_{m=0}^{M_k} \tilde{\tau}_{k,m}(\mathbf{h}) \rho_{k,m}]$, we can define the region $\bar{\mathcal{P}}^{P}_{\text{TDMA}}(\tilde{\boldsymbol{\tau}}(\mathbf{h}), \boldsymbol{p})$ of achievable average transmit-power vectors $\bar{\mathbf{p}}$ for the P-CSIT based minimization in (5); and after taking the union as in (37), the corresponding region of $\bar{\mathcal{P}}_{P-\text{CSIT}}(\check{\boldsymbol{r}}, \check{\boldsymbol{\epsilon}})$ of all average transmit-power vectors for the P-CSIT based TDMA system [18]. As the latter is convex, its boundary points are attained by solving (5) for all possible weight vectors $\boldsymbol{w} \geq \mathbf{0}$ [18].

In the case of $K = 2$ users, we plot these power regions for two sets of individual average rate requirements: i) $\check{R}_1 =$ 100 kb/s, $\check{R}_2 = 100$ kb/s, and ii) $\check{R}_1 = 100$ kb/s, $\check{R}_2 = 50$ kb/s. Fig. 3 depicts on the $(\bar{P}_1, \bar{P}_2)$ plane the regions $\bar{\mathcal{P}}_{\text{P-CSIT}}(\check{\boldsymbol{r}}, \check{\boldsymbol{\epsilon}})$ and $\bar{\mathcal{P}}_{\text{Q-CSIT}}(\check{\boldsymbol{r}}, \check{\boldsymbol{\epsilon}})$ lying on the north-east side of their boundaries defined by the lines which correspond to the different average rate requirements (average BER requirements are fixed to $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$). For any fixed pair $(\bar{P}_1, \bar{P}_2)$ of average transmit powers within each power region, there always exists a corresponding resource allocation policy achieving the required average rate and BER. Furthermore, any transmit-power optimal solution yielding the smallest weighted sum of average transmit powers corresponds to a boundary point. Notice that with $\check{R}_1 = \check{R}_2$, the power regions are symmetric w.r.t. the line $\bar{P}_1 = \bar{P}_2$; while they are nonsymmetric for $\check{R}_2 = \check{R}_1/2$. Recall that P-CSIT can be seen as the limiting form of Q-CSIT as the feedback rate goes to infinity. With more information at the transmitters, we can perform more intelligent resource allocation. This explains why the Q-CSIT based power regions are always contained within the P-CSIT based ones. Note however that when the inequality present in the definition (36) is strict, the Q-CSIT regions in Fig. 3 are actually conservative estimates. Nonetheless the resultant average power regions for the Q-CSIT based TDMA are very close to their P-CSIT based counterparts. This implies that also with regards to fundamental limits, the JRAQ algorithm is near-globally optimal.

To assess performance of the JRAQ algorithm as the number of users increases, Fig. 2 depicts the total weighted average transmit power for different values of K with $\check{R}_k = 50$ kb/s, $w_k = 1$ and $\overline{\gamma}_k = 3$ dB remaining fixed $\forall k$. Two major conclusions can be drawn from Fig. 2: i) the gap between P-CSIT

TABLE I
PERFORMANCE OF THE JRAQ ALGORITHM FOR DIFFERENT TEST CASES (FOR CASES II AND IV THE RATE REQUIREMENT FOR USER 2 IS 50 KB/S)

Test Case	User (k)	w_k	$\bar{\gamma}_k[dB]$	$\check{R}_k[kbps]$	$\check{\epsilon}_k$	$\bar{P}_k[dBw]$	P-CSIT $\bar{P}_k[dBw]$
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
VI	1	4/3	0	100	10^{-3}	9.26	8.64
	2	4/3	0	49	10^{-3}	5.27	4.74
	3	2/3	0	99	10^{-3}	10.25	9.71
	4	2/3	0	50	10^{-3}	6.47	5.95

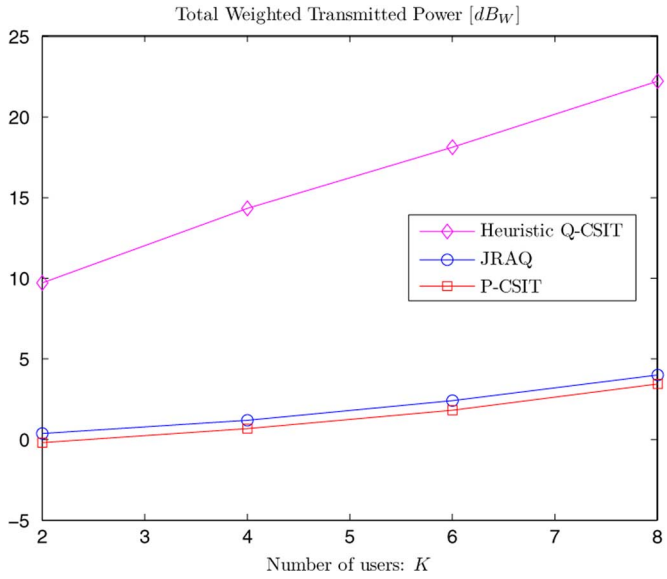


Fig. 2. Total power consumption for different resource allocation schemes with different number of users ($w_k = 1$, $\bar{\gamma}_k = 3$ dB, $\check{\epsilon}_k = 10^{-3}$, and $\check{R}_k = 50$ kb/s $\forall k$).

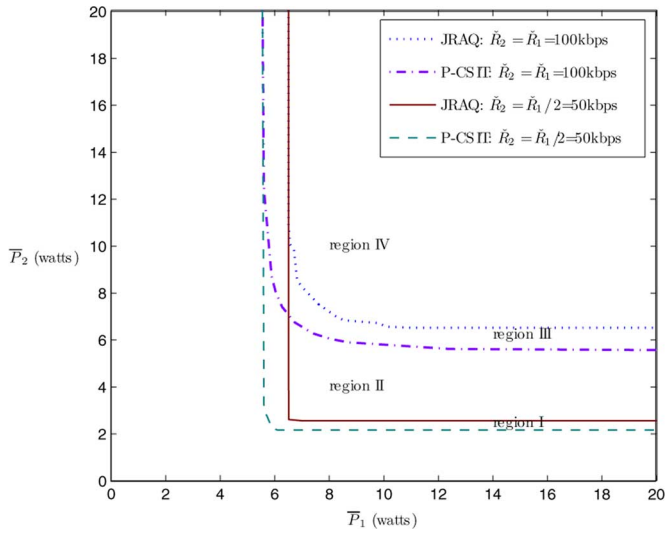


Fig. 3. Power regions for P-CSIT and JRAQ policies ($\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB).

based resource allocation and JRAQ remains small for all tested configurations; and ii) the gap between JRAQ and the heuristic Q-CSIT allocation widens as the number of users increases. This is because the opportunistic channel access in JRAQ exploits the multi-user diversity provided by the uncorrelated fading channels, whereas the fixed channel assignment implemented by the heuristic Q-CSIT allocation does not.

Additional numerical tests of the JRAQ algorithm are summarized in Table I. Its entries illustrate that the constraints are tightly met and corroborate that our Q-CSIT based block component iterations converge to average power efficiency close to that attained by the P-CSIT based benchmark.

To gain more insight, let us take a closer look at the JRAQ algorithm when $\check{R}_1 = \check{R}_2 = 100$ kb/s and $w_1/w_2 = 2$. The power and rate loadings are listed in Table II, whereas the quantization regions and user-time schedules are depicted in Fig. 4 (different shades represent the user selected to access the channel). From Table II, we deduce that $p_{k,m_1} > p_{k,m_2}$ $\forall m_1 > m_2$. This illustrates the water-filling principle which holds for both the Q-CSIT based optimal power loading of Section IV-B as well as for the P-CSIT one of Section III. Indeed, when the channel is more reliable, higher transmit rate can be afforded at lower transmit power. Fig. 4 also confirms that the optimal regions $\{\mathcal{H}^*_{k,m}\}_{m=1}^{3}$, $k = 1, 2$, are nonoverlapping consecutive intervals and can thus be determined by the set of thresholds $\{H^*_{k,m}\}$ represented with boldface lines.

We have seen that with three AMC modes the JRAQ algorithm provides average power efficiency approaching the P-CSIT based benchmark, while requiring only 3 bits of Q-CSIT per frame. We next test how the number of feedback bits affects the performance of JRAQ. For $\check{R}_1 = \check{R}_2 = 100$ kb/s and $w_1/w_2 = 1$, Table III lists the total average transmit-power cost for the two-user Rayleigh flat-fading TDMA channel with a variable number of feedback bits. When one bit is available, the feedback information only indicates user selection; and once the winner terminal is picked, it transmits regardless of $\mathbf{h}$. Fig. 4 illustrates that for those $\mathbf{h}$ realizations that the channel experiences a deep fade, both users defer as suggested by our analytical results. Even though the region $\bar{\mathcal{P}}_{Q-CSIT}(\check{\mathbf{r}}, \check{\boldsymbol{\epsilon}})$ for this case is small, the power required to compensate for these "bad" channels is high (23.05 dBw). Compared to the power required with 4 feedback bits (8.43 dBw) this represents

TABLE II
TRANSMIT-POWER ($p_{k,m}$)) AND AMC RATE ($\rho_{k,m}$) LOADING PER QUANTIZATION STATE RETURNED BY THE JRAQ ALGORITHM ($w_1 = 2/3$, $w_2 = 1/3$, $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\check{R}_1 = \check{R}_2 = 100$ KB/S, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ DB)

	User 1			User 2		
Quantization Region	$\mathcal{H}_{1,1}$	$\mathcal{H}_{1,2}$	$\mathcal{H}_{1,3}$	$\mathcal{H}_{2,1}$	$\mathcal{H}_{2,2}$	$\mathcal{H}_{2,3}$
Transmit-power $p_{k,m}$ $[dBw]$	8.56	13.23	15.60	8.99	13.84	16.29
Rate $\rho_{k,m}$ $[bits/sym]$	1	3	5	1	3	5

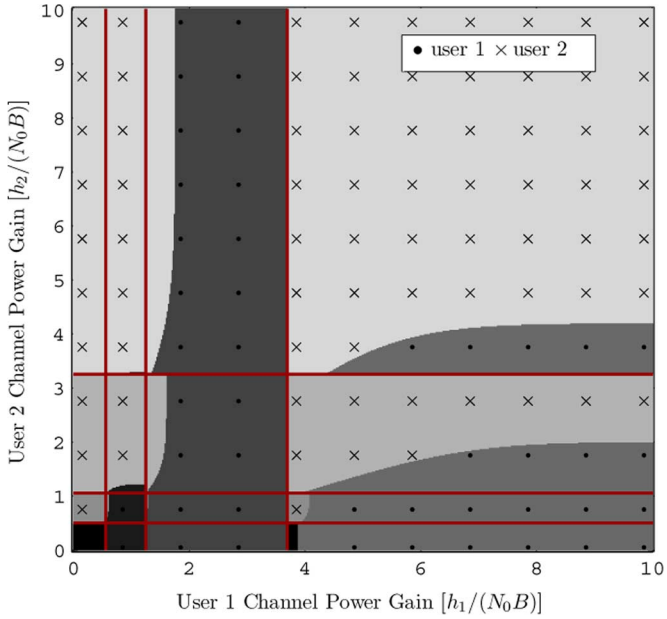


Fig. 4. Optimal user-time allocation policy and quantization regions obtained by the JRAQ algorithm, where user selections are indicated using different shades and quantization thresholds are represented with boldface lines ($w_1 = 2/3$, $w_2 = 1/3$, $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\check{R}_1 = \check{R}_2 = 100$ kb/s, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB).

TABLE III
AVERAGE WEIGHTED POWER FOR JRAQ WITH A VARIABLE NUMBER OF FEEDBACK BITS ($w_1 = w_2 = 1$, $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\check{R}_1 = \check{R}_2 = 100$ KB/S, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ DB)

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

about 15 dBw in average power savings. As the number of feedback bits increases, the number of active AMC modes per user increases too. Somewhat surprisingly, even with only two feedback bits (when $M_1 = 2$ and $M_2 = 1$), JRAQ provides an average power cost not too different from the P-CSIT benchmark (cf. Table III). This confirms that the user-time allocation policy plays a major role in power 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 based TDMA systems.

The convergence of JRAQ is illustrated in Fig. 5, where the average total weighted power is plotted as a function of the inner iteration steps. (Recall that for each outer iteration (t), the JRAQ algorithm implements three inner steps.) We observe that JRAQ

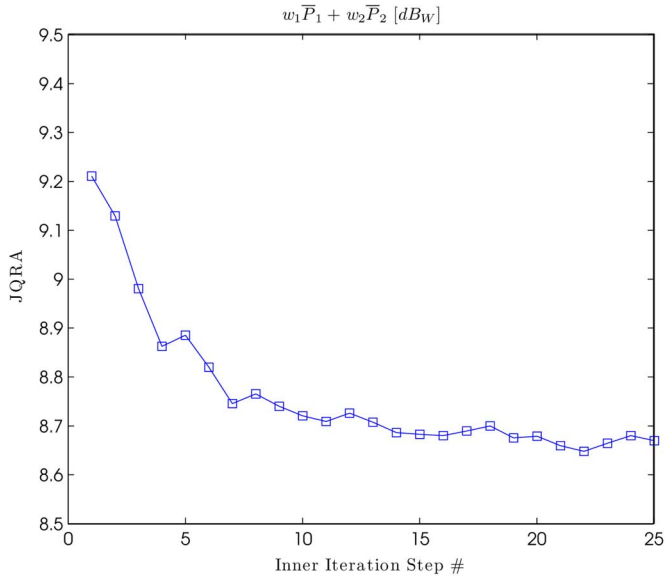


Fig. 5. Average weighted power evolution of the JRAQ algorithm ($w_1 = 2/3$, $w_2 = 1/3$, $\check{\epsilon}_1 = \check{\epsilon}_2 = 10^{-3}$, $\check{R}_1 = \check{R}_2 = 100\ kb/s$, $\bar{\gamma}_1 = \bar{\gamma}_2 = 0$ dB).

converges after a small number of iterations ($t = 5$ outer iterations which means $(5 \times 3 + 1) = 16$ inner steps). The curve is not perfectly smooth due to the finite resolution in the numerical integrations involved. Another interesting observation is that even the first inner iteration step returns a relatively reliable solution. Although JRAQ is implemented once off-line based on the long-term channel statistics, Fig. 5 further demonstrates the fast convergence of the off-line JRAQ algorithm that relies on the CDF of the channel gains.

VI. CONCLUSION

We developed a framework to minimize weighted average transmit power subject to average rate and BER constraints in wireless TDMA where terminals rely on a pool of AMC modes and adjust their transmissions according to the limited-rate feedback they receive from the access point. When P-CSIT is available, optimal scheduling and resource allocation policies turned out to be almost surely opportunistic (at most one user transmitting per frame), as with existing rate-maximizing schemes which rely on theoretically infinite-size capacity-achieving codebooks. Incorporation of average (as opposed to instantaneous) constraints in our practical P-CSIT setup enabled efficient off-line computation of the analytical (and globally convergent) solution based on the continuous CDF of the channel gains, and led to low overhead in the feedback link.

Interestingly, power-efficient policies turned out to be also opportunistic even for challenging setups where only Q-CSIT is available. For the case where quantization regions are prespecified by the system design, we formulated and solved analytically Q-CSIT based convex optimization problems yielding power-efficient scheduling and transmit-power allocation schemes that are globally optimum. We further tackled the joint resource allocation-quantization (JRAQ) problem which necessitated formulation and construction of a nonconventional power-efficient quantizer and simultaneous optimization of the resource management policies with the form of Q-CSIT used. The novel design linked quantization regions with AMC modes on a per user basis, which reduced complexity and lowered the required feedback overhead. The JRAQ problem was solved using an efficient block-component iterative algorithm with guaranteed convergence to at least a local optimum. (Recall that this is the best one can expect when nonconvex vector quantization problems are involved.) Requiring just sub-gradient recursions and two-dimensional searches, all of which are performed off-line, complexity of the optimization is certainly affordable. What is more, the on-line opportunistic scheduling requires a surprisingly small number of bits in the feedback (in the order of 3–7 per channel instantiation for TDMA systems with 2–16 users, each with 4–8 AMC modes).

We finally relied on simulated tests to compare our user scheduling, resource allocation, and JRAQ schemes based on Q-CSIT against a heuristic alternative and against the P-CSIT benchmark. These tests confirmed our analytical findings and demonstrated considerable savings in transmit power relative to sub-optimum allocation schemes. They also suggest that the Q-CSIT based JRAQ algorithm holds great potential for practical deployment since it fits the specs of current access standards and with simple enough on-line overhead it can come surprisingly close to the benchmark P-CSIT based performance.[4]

APPENDIX
PROOF OF (22) AND (23)

We will first show that $\mu_k^{Q_p*} \neq 0, \forall k$. Arguing by contradiction, we suppose that $\mu_k^{Q_p*} = 0$ for a certain k and deduce that [cf. (21)]

$$w_k \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = \nu_{k,m}^{Q_p*}, \quad m \in \{1, \ldots, M_k\}. \tag{38}$$

Since $\forall m$, $\int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) > 0$, we have that $\forall m$, $\nu_{k,m}^{Q_p*} > 0$. By complementary slackness [2] between $\nu_{k,m}^{Q_p*}$ and constraint $p_{k,m}^* \geq 0$, this implies that $p_{k,m}^* = 0, \forall m$. But excluding the trivial case where $\check{\epsilon}_k \geq 0.5$, the average BER corresponding to user k becomes 0.5 which is a contradiction. Hence, we have $\mu_k^{Q_p*} > 0 \ \forall k$. Also by the complementary slackness, we have from (21) that either $p_{k,m}^* = 0$, or, if $p_{k,m}^* > 0$ (thus $\nu_{k,m}^{Q_p*} = 0$), then

$$\mu_k^{Q_p*} \rho_{k,m} / \check{R}_k \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) \dot{\epsilon}_{k,m}(h_k p_{k,m}^*) dF(\mathbf{h}) + w_k \int_{\mathcal{H}_{k,m}} \tau_k(\mathbf{h}) dF(\mathbf{h}) = 0 \tag{39}$$

which readily leads to (22). Since $\mu_k^{Q_p*} > 0, \forall k$, the complementary slackness between $\mu_k^{Q_p*}$ and the BER constraints imply readily that the average BER constraints are achieved with equality as in (23).

[4]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.

REFERENCES

[1] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, M: Athena Scientific, 1999.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrdge, U.K.: Cambridge Univ. Press, 2004.

[3] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal energy allocation for delay-constrained data transmission over a time-varying channel," in *Proc. INFOCOM Conf.*, San Francisco, CA, Mar. 30–Apr. 3, 2003, vol. 2, pp. 1095–1105.

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

[5] G. B. Giannakis, Z. Liu, X. Ma, and S. Zhou, *Space-Time Coding for Broadband Wireless Communications*. New York: Wiley, 2007.

[6] A. J. Goldsmith and S. G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, pp. 595–602, May 1998.

[7] *Air interface for fixed broadband wireless access systems*, IEEE Std. 802.16, Apr. 2002.

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

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

[10] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—Part I and Part II," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1127, Mar. 2001.

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

[12] A. G. Marques, F. F. Digham, and G. B. Giannakis, "Power-efficient OFDM via quantized channel state information," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1581–1592, Aug. 2006.

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

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

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

[16] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

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

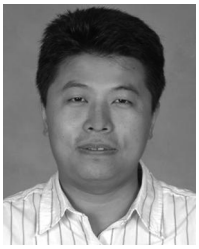
[18] X. Wang and G. B. Giannakis, "Power-efficient resource allocation for time-division multiple access over fading channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1225–1240, Mar. 2008.

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

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

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

[22] A. G. Marques, G. B. Giannakis, and F. J. Ramos, "Optimum scheduling for orthogonal multiple access over fading channels using quantized channel state information," in *Proc. Workshop on Signal Processing Advances in Wireless Communication*, Recife, Brazil, Jul. 4–6, 2008.



Xin Wang (M'04) received the B.Sc. degree and the M.Sc. degree from Fudan University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree from Auburn University, Auburn, IL, in 2004, all in electrical engineering.

From September 2004 to August 2006, he was a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. Since September 2006, he has been an Assistant Professor in the Department of Electrical Engineering, Florida Atlantic University, Boca Raton. His research interests include medium access control, cross-layer design, resource allocation, and signal processing for communication networks.

Antonio G. Marques (M'07) received the Telecommunications Engineering (B.S. and M.Sc.) degree and the Ph.D. degree in electrical engineering from the Universidad Carlos III de Madrid, Madrid, Spain, in 2002 and 2007 respectively, both with highest honors.

In 2003 he joined the Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain, where he currently develops his research and teaching activities as an Assistant Professor. Since 2005 he is also a Visiting Researcher at the Department of Electrical Engineering, University of Minnesota, MN. His research interests lie in the areas of communication theory, signal processing, and networking. His current research focuses on channel state information designs, energy-efficient resource allocation, and wireless ad hoc and sensor networks.

Dr. Marques' work brought him several awards in distinctive international conferences including the Inernational Conference on Acoustics, Speech and Signal Processing (ICASSP) 2007.

G. B. Giannakis (F'97) received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1981. From 1982 to 1986, he was with the University of Southern California (USC), where he received the M.Sc. degree in electrical engineering in 1983, the M.Sc. degree in mathematics in 1986, and the Ph.D. degree in electrical engineering in 1986.

Since 1999, he has been a Professor with the Electrical and Computer Engineering Department at the University of Minnesota, where he now holds an ADC Chair in wireless telecommunications. His general interests span the areas of communications, networking and statistical signal processing—subjects on which he has published more than 275 journal papers, 450 conference papers, two edited books, and two research monographs. Current research focuses on complex-field and network coding, multicarrier, cooperative wireless communications, cognitive radios, cross-layer designs, mobile ad hoc networks, and wireless sensor networks.

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