STOCHASTIC RESOURCE ALLOCATION FOR ORTHOGONAL ACCESS BASED ON QUANTIZED CSI: OPTIMALITY, CONVERGENCE AND DELAY ANALYSIS

Antonio G. Marques[*] *Georgios B. Giannakis*[†] *Javier Ramos*[*]

[*]Dept. of Signal Theory and Communications, Rey Juan Carlos University, Madrid, Spain
[†]Dept. of Electrical Engineering, University of Minnesota, Minneapolis, MN, USA

ABSTRACT

Dynamic allocation of power, rate and channel access is a critical task in wireless networks. Capitalizing on convex optimization and stochastic approximation tools, this paper develops a stochastic resource allocation algorithm that minimizes average transmit power under individual average rate constraints. Focus is placed on networks where users transmit orthogonally over a set of parallel channels and transmissions are adapted based on quantized channel state information (CSI) allowing even channel statistics to be unknown. Convergence of the developed stochastic scheme is characterized and the average queue delays are obtained in closed form.

Index Terms— Resource management, cross-layer design, stochastic approximation, multiuser channels, delay effects.

1. INTRODUCTION

The benefits of implementing adaptive scheduling and resource allocation for multiple access fading channels are well documented; see e.g., [1] for a tutorial treatment. Early efforts were focused on adaptation of rate, power and channel access per fading state so that a specific performance measure (e.g., ergodic capacity) is optimized while satisfying quality of service (QoS) requirements (e.g., average power consumption). This optimization was typically carried out assuming that both channel probability density function (PDF) and the instantaneous channel realization are perfectly known. However, in many practical wireless scenarios those assumptions are unrealistic. On one hand, errors in estimating the channel, feedback delay and the asymmetry between forward and reverse links render acquisition of perfect CSI impossible. For such cases, only quantized (Q-) CSI that can be pragmatically obtained through finite-rate feedback. On the other hand, channel statistics of the full system may be difficult to acquire due to a large number of users or non-stationarities. Stochastic approximation algorithms naturally emerge as a means of bypassing this problem and accounting for channel variations [2].

In response to these challenges, the present paper investigates scheduling and resource allocation for orthogonal multiaccess transmissions over fading channels with unknown statistics when only Q-CSI is available both at the receiving and transmitting ends. Using as starting point our results in [3], we minimize an average power cost subject to QoS constraints on average rate and bit error rate (BER). The main contribution of this work is twofold: (i) we develop an adaptive stochastic algorithm capable of learning the intended channels on-the-fly and converging to the optimal value; and (ii) we characterize the average delay of the proposed algorithm.[1]

2. DYNAMIC RESOURCE ALLOCATION

Consider a wireless network with M users, indexed by $m \in \{1, \ldots, M\}$, who transmit over K flat-fading orthogonal channels, indexed by $k \in \{1, \ldots, K\}$, to a common destination (e.g., base station, access point or ad-hoc network). Assuming zero-mean additive white Gaussian noise (AWGN) with unit variance, let $g_{m,k}$ represent the instantaneous power fading gain of the kth channel between the mth user and the destination; and $\mathbf{G}$ the $M \times K$ matrix with entries $[\mathbf{G}]_{m,k} := g_{m,k}$. Let the domain of each $g_{m,k}$ be divided into different regions. Instead of assuming that sources and destinations know $g_{m,k}$ (perfect CSI), here only the index of the region $g_{m,k}$ falls into (represented by $j_{m,k}$) is known. Let $\mathbf{J}$ denote the $M \times K$ matrix with entries $[\mathbf{J}]_{m,k} := j_{m,k}$, which represent the Q-CSI of the entire system. Since $g_{m,k}$ is random, the index $j_{m,k}$ is a discrete random variable (and $\mathbf{J}$ is a random matrix with $\mathcal{J}$ representing the set of possible realizations of $\mathbf{J}$).

As in [4], users are allowed to access simultaneously any of the channels. Let matrix $\mathbf{W}$ represent the allocation policy whose $[\mathbf{W}]_{m,k}$ entry corresponds to the portion of the kth channel dedicated to the mth user with $\sum_{m=1}^{M}[\mathbf{W}]_{m,k} \leq$

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

[1] *Notation*: We use boldface upper (lower) face letters to denote matrix (column vectors); $(\cdot)^T$ denotes transpose; $[\cdot]_{k,l}$ denotes the (k,l)th entry of a matrix, and $[\cdot]_k$ the kth entry of a vector. Calligraphic letters denote sets, with $|\mathcal{X}|$ denoting the cardinality of the set $\mathcal{X}$. If g is a continuous function, $\dot{g}$ denotes its derivative. Finally $\wedge$ denotes the logical "and" operator, and $\mathbf{I}_{\{\cdot\}}$ the indicator function ($\mathbf{I}_{\{x\}} = 1$ if x is true and zero otherwise).

$1, \forall k$. Let $\mathbf{P}$ and $\mathbf{R}$ represent the system power and rate matrices of size $M \times K$. When the mth user is the only one accessing the kth channel, the entries $[\mathbf{P}]_{m,k}$ and $[\mathbf{R}]_{m,k}$ represent *nominal* transmit-power and *nominal* transmit-rate, respectively. As the system is going to adapt the resources depending on $\mathbf{J}$, it follows that $\mathbf{W} = \mathbf{W}(\mathbf{J})$, $\mathbf{P} = \mathbf{P}(\mathbf{J})$, and $\mathbf{R} = \mathbf{R}(\mathbf{J})$ and consequently each of them can take, at most, $|\mathcal{J}|$ different values. Finally, let Υ (Υ^{-1}) be the power-rate (rate-power) function that through QoS requirements links $[\mathbf{P}]_{m,k}$ and $[\mathbf{R}]_{m,k}$ in the same region $\mathcal{R}$ (wherever appropriate, we will write $\Upsilon_{\mathcal{R}([\mathbf{J}]_{m,k})}$ to stress this fact). For illustration purposes, consider a system whose QoS requirements impose a maximum instantaneous BER of $\check{\epsilon}_{\max}$, and symbols are drawn from QAM constellations. The maximum BER for a given region can then be approximated by $\epsilon_{\max} = 0.2 \exp((-g^{\min}_{m,k}([\mathbf{J}]_{m,k})p_{m,k} / (2^{r_{m,k}} - 1))$, cf. [1, Eq. (9.8)]; where $g^{\min}_{m,k}([\mathbf{J}]_{m,k}) := \min_{g_{m,k}}\{g_{m,k} \in \mathcal{R}([\mathbf{J}]_{m,k})\}$. For this case, $\Upsilon_{\mathcal{R}([\mathbf{J}]_{m,k})}$ can be written as $\Upsilon_{\mathcal{R}([\mathbf{J}]_{m,k})}(x) = ((2^x - 1)\ln(0.2/\check{\epsilon}_{\max}))/g^{\min}_{m,k}([\mathbf{J}]_{m,k})$

2.1. Problem Formulation

We are interested in minimizing the average weighted transmit-power subject to individual average rate constraints. For ergodic channels, the average transmit-power and rate for the mth user are, $\bar{p}_m := \sum_{\forall \mathbf{J}} \sum_{k=1}^{K} ([\mathbf{P}(\mathbf{J})]_{m,k}[\mathbf{W}(\mathbf{J})]_{m,k}) \Pr\{\mathbf{J}\}$ and $\bar{r}_m := \sum_{\forall \mathbf{J}} \sum_{k=1}^{K} ([\mathbf{R}(\mathbf{J})]_{m,k}[\mathbf{W}(\mathbf{J})]_{m,k}) \Pr\{\mathbf{J}\}$, respectively. Note that evaluating $\Pr\{\mathbf{J}\}$ requires knowledge of the channel quantizer and the channel PDF.

Positive power weights $\boldsymbol{\mu} := [\mu_1, \ldots, \mu_M]^T$ and individual rate constraints $\check{\mathbf{r}} := [\check{r}_1, \ldots, \check{r}_M]^T$ will be used to effect different priority levels among users. Taking into account these observations, the optimal resource allocation can be obtained as the solution of the following constrained optimization problem

$$\begin{cases} \min_{\mathbf{R}(\mathbf{J})\geq\mathbf{0}, \mathbf{W}(\mathbf{J})\geq\mathbf{0}} \sum_{m=1}^{M} [\boldsymbol{\mu}]_m \bar{p}_m \\ \text{s. to}: \ \bar{r}_m \geq [\check{\mathbf{r}}]_m, \ \ \forall m, \\ \qquad \sum_{m=1}^{M} [\mathbf{W}(\mathbf{J})]_{m,k} \leq 1, \ \ \forall k, \forall \mathbf{J} \end{cases} \tag{1}$$

Since $\Upsilon_{\mathcal{R}([\mathbf{J}]_{m,k})}$ links $\mathbf{R}$ with $\mathbf{P}$, only optimization over one of them is required. Assuming that Υ is strictly convex, (1) can be optimally solved using Lagrangian relaxation [3].

2.2. Optimal Resource Allocation

Let $\boldsymbol{\lambda}^R$ be an $M \times 1$ vector whose mth entry is the Lagrange multiplier associated with the mth rate constraint. Consider the rate allocation (recall $\dot{x}$ denotes derivative of x)

$$[\mathbf{R}(\mathbf{J})]_{m,k} = \dot{\Upsilon}^{-1}_{\mathcal{R}([\mathbf{J}]_{m,k})} \left(\frac{[\boldsymbol{\lambda}^R]_m}{[\boldsymbol{\mu}]_m} \right) \mathbf{I}_{\{\dot{\Upsilon}^{-1} \ \frac{[\boldsymbol{\lambda}^R]_m}{[\boldsymbol{\mu}]_m} > 0\}} \tag{2}$$

where $\dot{\Upsilon}^{-1}_{\mathcal{R}([\mathbf{J}]_{m,k})}$ is the inverse function of $\dot{\Upsilon}_{\mathcal{R}([\mathbf{J}]_{m,k})}$. We use (2) to define the cost of allocating user m to channel k as $[\mathbf{C}_{\mathbf{W}}(\mathbf{J})]_{m,k} := [\boldsymbol{\mu}]_m \Upsilon_{\mathcal{R}([\mathbf{J}]_{m,k})}([\mathbf{R}(\mathbf{J})]_{m,k}) - [\boldsymbol{\lambda}^R]_m [\mathbf{R}(\mathbf{J})]_{m,k}$. With ε representing a small positive number, we also define the vector $[\mathbf{c}_{\mathbf{W}^*}(\mathbf{J}, \boldsymbol{\lambda}^R))]_k := \min_m\{[\mathbf{C}_{\mathbf{W}}(\mathbf{J}, \boldsymbol{\lambda}^R))]_{m,k}\}_{m=1}^{M}$, and the set $\mathcal{M}(\mathbf{J}, k) := \{m : ([\mathbf{C}_{\mathbf{W}}(\mathbf{J}, \boldsymbol{\lambda}^R)]_{m,k} - [\mathbf{c}_{\mathbf{W}^*}(\mathbf{J}, \boldsymbol{\lambda}^R))]_k < \varepsilon) \ \wedge \ ([\mathbf{c}_{\mathbf{W}^*}(\mathbf{J}, \boldsymbol{\lambda}^R))]_k < 0)\}$. Based on the previous definitions, the following channel allocation (scheduling) is considered

$$[\mathbf{W}(\mathbf{J}, \boldsymbol{\lambda}^R)]_{m,k} := \mathbf{I}_{\{m \in \mathcal{M}(\mathbf{J},k)\}} \times \frac{\left(1 - \frac{[\mathbf{C}_{\mathbf{W}}(\mathbf{J},\boldsymbol{\lambda}^R)]_{m,k} - [\mathbf{c}_{\mathbf{W}^*}(\mathbf{J},\boldsymbol{\lambda}^R)]_k}{\varepsilon}\right)^2}{\sum_{m \in \mathcal{M}(\mathbf{J},k)} \left(1 - \frac{[\mathbf{C}_{\mathbf{W}}(\mathbf{J},\boldsymbol{\lambda}^R)]_{m,k} - [\mathbf{c}_{\mathbf{W}^*}(\mathbf{J},\boldsymbol{\lambda}^R)]_k}{\varepsilon}\right)^2}. \tag{3}$$

When scheduling in (3) is implemented, only users in $\mathcal{M}(\mathbf{J}, k)$ (those with minimum cost) can access channel k. For most channel realizations, the set $\mathcal{M}(\mathbf{J}, k)$ will contain a single user, rendering the access opportunistic. However, there will also exist realizations for which $|\mathcal{M}(\mathbf{J}, k)| > 1$, and then a small group of users will access the channel. Using the results in [3], it can be proved that: *the scheduling in* (3) *and the rate (thus power) allocation in* (2) *are asymptotically optimal*.

It is important to underscore that CSI affects the optimal allocation in two different ways. On one hand, it depends on the current channel realization $\mathbf{J}$. On the other hand, it depends on the channel PDF via the Lagrange multiplier. Details of the computation of $\boldsymbol{\lambda}^R$ will be provided in the next section.

3. STOCHASTIC LAGRANGE MULTIPLIERS

Upon defining the $M \times 1$ vector $\partial^s D(\boldsymbol{\lambda}^R)$ with entries $[\partial^s D(\boldsymbol{\lambda}^R)]_m := [\check{\mathbf{r}}]_m - \bar{r}_m(\boldsymbol{\lambda}^R)$, the optimum value of the multiplier vector can be obtained through the iterations

$$\boldsymbol{\lambda}^{R(i)} = \left[\boldsymbol{\lambda}^{R(i-1)} + \beta \partial^s D(\boldsymbol{\lambda}^{R(i-1)})\right]^+ \tag{4}$$

where β is a small stepsize and i the iteration index (cf. [3, Prop. 5]). The computation of $\boldsymbol{\lambda}^R$ using (4) is performed off-line and requires knowledge of the channel statistics. Specifically, to find $\bar{r}_m(\boldsymbol{\lambda}^R)$, the probabilities $\Pr\{\mathbf{J}\}$ have to be known $\forall \mathbf{J}$. However, there are cases where this computation cannot be efficiently carried out or is not even feasible. Such cases include scenarios where the set-up (number of users, channel statistics, QoS requirements) changes so frequently that $\boldsymbol{\lambda}^R$ has to be continuously re-computed. Other examples include limited-complexity systems that cannot afford the off-line burden, or, when the channel statistics are unknown. For those cases, stochastic approximation arises as an alternative approach to estimating $\boldsymbol{\lambda}^R$ [2].

Let n denote the current block index (whose duration will correspond to the coherence time), $\mathbf{J}[n]$ the fading state during block n, and $r_m(\mathbf{J}[n], \boldsymbol{\lambda}^R[n]) := \sum_{\forall k} [\mathbf{R}(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])]_{m,k}$

$[\mathbf{W}(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])]_{m,k}$ the instantaneous transmit-rate for user m. Our proposal is to replace $\partial^s D(\boldsymbol{\lambda}^R)$ by its stochastic version $\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R)$, with $[\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R)]_m := [\check{\mathbf{r}}]_m - r_m(\mathbf{J}[n], \boldsymbol{\lambda}^R)$. Using this definition, the original iterations over $\boldsymbol{\lambda}^R$ in (4) can be replaced by the stochastic estimates

$$\hat{\boldsymbol{\lambda}}^R[n] = \left[\hat{\boldsymbol{\lambda}}^R[n-1] + \beta \partial^s D(\mathbf{J}[n], \hat{\boldsymbol{\lambda}}^R[n-1])\right]^+. \quad (5)$$

It must be emphasized that the stochastic iterations in (5) take into account the entire history of the channel. Specifically, $\mathbf{J}[n]$ is explicitly considered as an argument of $\partial^s D()$ while $\mathbf{J}[0], \ldots, \mathbf{J}[n-1]$ are implicitly considered via $\hat{\boldsymbol{\lambda}}^R[n-1]$.

It can be shown that for sufficiently small β the trajectories of the iterations in (4) and (5) are locked and the stochastic iterates in (5) converge to $\boldsymbol{\lambda}^R$. In a nutshell, we have established the following result.

Theorem 1 With similar initial conditions in (4) and (5) and given $T > 0$, there exist $b_T > 0$ and $\beta_T > 0$ so that

$$\max_{1 \leq n \leq T/\beta} \|\boldsymbol{\lambda}^{R(n)} - \hat{\boldsymbol{\lambda}}^R[n]\| \leq c_T(\beta) b_T \quad 0 \leq \beta \leq \beta_T$$

where $c_T(\beta) \to 0$ as $\beta \to 0$.

The result in Theorem 1 can be proved using the averaging approach in [5, Ch. 7]. Following the averaging method for the approximation of the trajectory of the difference (or differential) equations, updates in (5) and those in (4) can be seen as a pair of *primary* and averaged systems. Under general conditions, it is possible to show trajectory locking of these two systems via [5, Th. 7.2 and 7.3]. The full proof of the proposition is omitted due to space limitations, but the main idea is that the Lipschitz continuity of $\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R)$ with respect to (w.r.t.) $\boldsymbol{\lambda}^R$ can be used to prove that the most challenging conditions required in [5, Th. 7.2 and 7.3] hold.

4. QUEUEING AND DELAY ANALYSIS

Although not explicitly mentioned, the average rate constraint in (1) assumes the existence of a sufficient large input queue in every terminal m. This way, packets arriving at a rate $[\check{\mathbf{r}}]_m$ are stored in the queue and transmitted in a first-in-first-out (FIFO) fashion every time the mth user is scheduled to access the channel. In this context, it is of interest to characterize the dynamics of the queues as well as the delay performance of the resource allocation algorithm.

The first step is to recognize that the arrival rates of packets follow a random pattern as long as the *average* arrival rate is $[\check{\mathbf{r}}]_m$. In other words, if $a_m[n]$ denotes the number of arriving bits at slot n, then any arrival pattern satisfying $\bar{a}_m[n] = [\check{\mathbf{r}}]_m$ should be admissible. Random arrival rates can be incorporated into our stochastic algorithm by considering the following modified version of $\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])$

$$[\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R[n], \mathbf{a}[n])]_m := a_m[n] - r_m(\mathbf{J}[n], \boldsymbol{\lambda}^R[n]) \quad (6)$$

where the constant term $[\check{\mathbf{r}}]_m$ has been replaced by the stochastic $a_m[n]$. As far as the instantaneous arrivals are bounded (i.e., $a_m[n] < \infty$), it can be shown that the convergence claimed in Theorem 1 still holds when $\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])$ in (5) is replaced by $\partial^s D(\mathbf{J}[n], \boldsymbol{\lambda}^R[n], \mathbf{a}[n]))$.

To analyze the stability of the resource allocation algorithm, the queue dynamics need to be characterized. With $q_m[n]$ denoting the queue size of user m at the beginning of slot n, the mth queue with arrival rate of $a_m[n]$ and departure rate $r_m(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])$, obeys the recursion

$$q_m[n+1] = \left[q_m[n] - r_m(\mathbf{J}[n], \boldsymbol{\lambda}^R[n])\right]^+ + a_m[n]. \quad (7)$$

Substituting (6) into (5) and comparing the modified version of (5) with (7), we can conclude that: *the size of the queues can be interpreted as a scaled version of the stochastic Lagrange multipliers*. Specifically, if $q_m[0] = \hat{\boldsymbol{\lambda}}^R[0] = 0$ we have $q_m[n] \cong [\hat{\boldsymbol{\lambda}}^R[n]]_m/\beta$. In fact, the only difference between the recursions for $q_m[n]$ and those for $[\hat{\boldsymbol{\lambda}}^R[n]]_m/\beta$ is the way in which the projection operation in $[\cdot]^+$ is implemented. Nevertheless, since the rate constraints are always active (i.e., $\hat{\boldsymbol{\lambda}}^R$ are strictly positive), after an initialization period the projection operation is transparent and the approximation $q_m[n] = [\hat{\boldsymbol{\lambda}}^R[n]]_m/\beta$ is accurate.

The previous result can help us characterize the stability and average delay of our stochastic resource allocation algorithm. On one hand, we know that: (i) a system is stable if $q_m[\infty] < \infty$, and (ii) the feasibility of (1) implies that $[\boldsymbol{\lambda}^R]_m = [\hat{\boldsymbol{\lambda}}^R[\infty]]_m < \infty$. From these considerations, we can readily conclude that our algorithm has a stable behavior. On the other hand, Little's result [6] asserts that with stable queues the average delay is given by the average queue length divided by the average arrival rate, i.e. $\bar{d}_m = \bar{q}_m/\bar{a}_m$. This in turn leads to an estimate of the average delay as

$$\bar{d}_m \cong [\boldsymbol{\lambda}^R]_m/(\beta[\check{\mathbf{r}}]_m). \quad (8)$$

In other words, the average delay of our stochastic algorithm can be estimated based on the optimal solution of (1), the rate requirements and the stepsize of the proposed iterations. It is worth mentioning that although in most cases the value of $[\boldsymbol{\lambda}^R]_m$ is not available in closed-form, convex optimization theory can be used to decipher properties of $[\boldsymbol{\lambda}^R]_m$. Specifically, analyzing the sensitivity of $[\boldsymbol{\lambda}^R]_m$ w.r.t. $[\check{\mathbf{r}}]_m$ will be critical to characterize how $\bar{d}_m$ varies w.r.t. $[\check{\mathbf{r}}]_m$.

Finally, it is important to stress the cross-layer characteristics of our stochastic algorithm. The scheduling in (3), prioritizes users with higher values of $[\boldsymbol{\lambda}^R]_m$ (higher $[\boldsymbol{\lambda}^R]_m$ will incur lower cost $[\mathbf{C}_{\mathbf{W}}(\mathbf{J})]_{m,k}$ $\forall k$ and thus a higher probability of accessing the channel). Since $[\hat{\boldsymbol{\lambda}}^R[n]]_m \cong \beta q_m[n]$, it is easy to deduce that users with larger queues have higher probability of being scheduled.

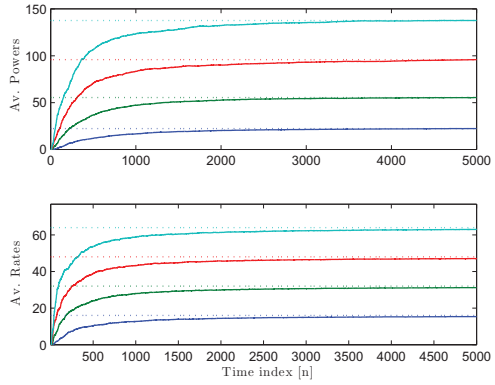

Fig. 1. Average power and rate.

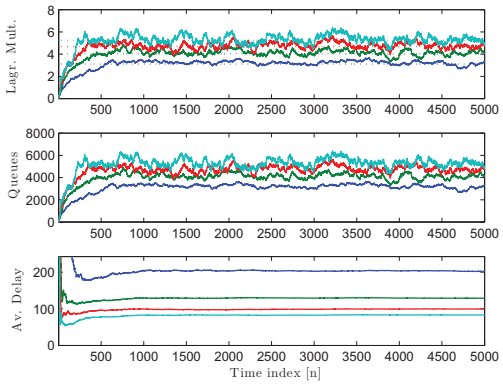

Fig. 2. Queueing and Delay.

5. SIMULATIONS

To numerically test our designs, we consider that each $g_{m,k}$ domain is divided into 5 scalar quantization regions. We will further assume that fading processes for different users are uncorrelated, channels are complex Gaussian distributed and symbols are drawn from QAM constellations so that the BER can be approximated by $0.2\exp\left(-g_{m,k}p_{m,k}/(2^{r_{m,k}}-1)\right)$. We test an OFDMA system with $M = 4$, $K = 64$, and 8 exponential taps that requires $\check{\mathbf{r}} = [16, 32, 48, 64]^T$ and BER not exceeding 10^{-3} per user. Arrival rates follow a Poisson distribution with $p = 0.5$; stochastic iterations in (6) with $\beta = 10^{-3}$ are implemented and the quantization regions simulated correspond to the low-complexity quantizer of [4, Sec. IV.B].

Figure 1 shows the time-evolution of the average power (top) and rate (bottom) for each user. Specifically, solid lines represent $\hat{\bar{p}}_m(n) = \frac{1}{n}\sum_{k=1}^{n} p_m[n]$ and $\hat{\bar{r}}_m(n) = \frac{1}{n}\sum_{k=1}^{n} r_m[n]$, while dotted lines depict the values obtained from the optimal off-line solution (assuming perfect knowledge of the channel PDF). The results indicate that the proposed algorithm converges to the optimal values (minimum power cost while satisfying the rate constraints) in a finite number of iterations. Results related with the queueing dynamics and delay are presented in Figure 2. In the first subplot solid lines represent $[\hat{\boldsymbol{\lambda}}^R[n]]_m$, while dotted lines depict the optimal values $[\boldsymbol{\lambda}^R]_m$ obtained from the off-line solution. The second subplot shows the queues size of each user. Clearly, users with higher rate requirements have larger queues. On the other hand, comparing the trajectories of $q_m[n]$ and $[\hat{\boldsymbol{\lambda}}^R[n]]_m$, we verify the validity of the approximation $q_m[n] \cong [\hat{\boldsymbol{\lambda}}^R[n]]_m/\beta$. The third subplot represents the expected delay at every time instant, and confirms the accuracy of the approximation in (8). Interestingly, simulations show that users with higher rate requirements, experience smaller delays; hence, the algorithm "prioritizes" information of users with high rate demand.

6. CONCLUSIONS

We developed a stochastic cross-layer algorithm that relies on a quantized version of the fading realization to specify power, rate and scheduling decisions so that the average weighted transmit-power is minimized. The resultant resource allocation is a function of the current channel realization and the Lagrange multipliers whose values depend on the history of the channel and the QoS requirements. The algorithm acquires the channel statistics on-the-fly and has provable convergence. Last but not least, upon identifying a relationship between the size of the input queues and the Lagrange multipliers, we were able to obtain the average queue delay of the novel scheme.[2]

7. REFERENCES

[1] A. Goldsmith, *Wireless Communications*, Cambridge University Press, 2005.

[2] X. Wang, G. B. Giannakis, and A. G. Marques, "A unified approach to qoS-guaranteed scheduling for channel-adaptive wireless networks," *Proceedings of the IEEE.*, vol. 95, no. 12, pp. 2410–24312, Dec. 2007.

[3] A. G. Marques, G. B. Giannakis, and F. J. Ramos, "Optimum scheduling for orthogonal multiple access in fading channels using quantized CSI," in *Proc. of Intl. Symp. on SPAWC*, Recife, Brasil, July 6-9 2008.

[4] A. G. Marques, G. B. Giannakis, F. Digham, and F. J. Ramos, "Power-efficient wireless OFDMA using limited-rate feedback," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 2, pp. 681–692, Feb. 2008.

[5] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*, Prentice Hall, 1995.

[6] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, New York: Willey, 1975.

[2]The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of the Army Research Laboratory or the U.S. Government.