Cross-Layer Resource Allocation for Downlink Access Using Instantaneous Fading and Queue Length Information

Luis M. Lopez-Ramos, Antonio G. Marques, Javier Ramos, and Antonio J. Caamaño

Dept. of Signal Theory and Communications, Rey Juan Carlos University, Camino del Molino s/n, Fuenlabrada, Madrid 28943, Spain.

{luismiguel.lopez, antonio.garcia.marques, javier.ramos, antonio.caamano}@urjc.es

Abstract—**Cross-layer algorithms that jointly allocate resources at different layers are known to foster the communication performance in wireless networks. Recent works have shown that fading and queue information are among the most critical parameters to consider in cross-layer designs. Motivated by those findings, this work develops optimal algorithms that use instantaneous fading and queue length information to allocate resources at transport, link and physical layers. Our focus is on the downlink channel of a cellular system where an access point sends different information to several users through a set flat-fading orthogonal channels. The queue stability and the average queuing delay of the developed algorithms are also investigated. Moreover, a simple mechanism to effect delay priorities among users is presented. Finally, methods that reduce the signalling overhead by allocating groups of channels to the same user are discussed at the end of the paper.**

Keywords: cross-layer design, network optimization, congestion control, dynamic resource management.

I. Introduction

In dynamic wireless networks where fading is time variant and the queues length vary instantaneously, nodes must be capable of adapting their transmission and reception parameters to the intended variations while adhering to power constraints and quality of service (QoS) requirements. For this reason, during the last years the design of optimal cross-layer resource allocation schemes has attracted considerable attention. Many works have relied on convex optimization and dual decomposition techniques to design the way resources are allocated; see, e.g., [1], [2], [3], [4]. Following a different approach, other works have relied in the dynamic backpressure policies introduced in the original work [5] to derive resource allocation algorithms that optimize the system performance while guarantee the queues of the wireless network to be stable; see, e.g, [6], [7], [8].

The goal of this paper is to optimally design an algorithm to allocate resources of different layers in the downlink channel of a fading wireless cellular network. The operating conditions of the system are as follows. At the transport layer, the access point (AP) receives packets from higher layers. The packets of each user entail different utility levels and the AP implements flow control mechanisms to keep the network stable. At the link layer, users share orthogonally a set of parallel flat-fading channels. At the physical layer, nodes can adapt their power and rate loadings in every channel.

The optimal cross-layer resource allocation (flows, powers, rates, and channels) is obtained as the solution of a constrained optimization problem, which naturally takes into account flow-specific utility functions, individual QoS requirements, and the operating conditions of the network. The resultant optimum dynamic resource allocation is found in closed form and it is shown to depend only on the current channel realization, and user-specific weights (Lagrange multipliers). An adaptive stochastic algorithm is developed to estimate the value of the Lagrange multipliers. The developed algorithm does not need knowledge of the channel distribution, and is asymptotically optimal.

Furthermore, it is shown that under very mild conditions, the stochastic estimations of some of the Lagrange multipliers are a scaled version of the length of queues. In other words, the developed schemes reveal the way in which the length of the queues can be used to optimize the performance of the network. Furthermore, by looking at the queues as a scaled version of the Lagrange multipliers, both the stability and the average queueing delay of our cross-layer algorithms are characterized. Joint consideration of wireless fading and instantaneous queue length in the present work represents a difference from most state-of-the-art works in the literature. In fact, not many works have addressed the design of cross-layer networking algorithms that take into account fading, see [1], [2], [3], [8], [9], [10], [11] for some of the exceptions.

The operating conditions of the system are described in Section II. The constrained optimization problem is formulated in Section III. The latter is solved in Section IV, where a stochastic method to estimate the optimum Lagrange multipliers is also developed. The relationship between those multipliers and the queues length is established at the end of that section. The changes in the formulation to reduce the signalling overhead via channel grouping are addressed in Section V. Numerical results and conclusions in Sections VI and VII wrap-up this paper. Due to space limitation no proofs have been included in the paper. [1]

II. Operating conditions

In this section, the system set-up and the model for the channel are first introduced. Then, the operation of the link and physical layers is presented. Finally, the flow-control mechanism and the dynamics of the queues are described.

Consider an access point (AP) transmitting over K flat-fading orthogonal channels to M wireless users (nodes).

Work in this paper was supported by the Spanish Government Grant No. TEC2009-12098.

[1] *N*otation: x^* denotes the optimal value of variable x; $\mathbb{1}_{\{\cdot\}}$ the indicator function ($\mathbb{1}_{\{x\}} = 1$ if x is true and zero otherwise); $[x]_a^b$ the projection of x onto the $[a,b]$ interval, i.e., $[x]_a^b = \min\{\max\{a,x\},b\}$; and $\mathcal{O}(\cdot)$ the Landau's big "O" order. Finally, for a function $f(\cdot)$, $(f)^{-1}(\cdot)$ denotes its inverse and $\dot{f}(\cdot)$ its derivative.

Channels and users will be indexed by k and m, respectively. The wireless link between the AP and user m at subcarrier k is characterized by its random square magnitude h_m^k, which is assumed normalized by the receiver noise variance. The overall $MK \times 1$ gain vector is denoted by $\mathbf{h} := \{h_m^k,\ m = 1, \ldots, M,\ \ k = 1, \ldots, K\}$. For simplicity, the channel is assumed to be ergodic and stationary. Moreover, it is assumed that the fading channel vector $\mathbf{h}$ remains invariant over a block of symbols but can vary from block-to-block (block fading channel model). In other words, if n denotes the current block index (whose duration is dictated by the channel coherence interval), we consider that $\mathbf{h}[n]$ remains constant and $\mathbf{h}[n] \neq \mathbf{h}[n+1]$. Notation $\mathbf{h}[n]$ will be used whenever the time variant nature of the channel needs to be emphasized.

Regarding the link layer operation, links at the outset can be scheduled to access simultaneously but orthogonally in time (or frequency) any of the channels; see, e.g., [9], [12]. To this end, let $w_m^k(\mathbf{h}) \in [0, 1]$ denote the nonnegative fraction of time (or channel bandwidth) the AP will use channel k to transmit information to user m during channel realization $\mathbf{h}$. Since every user's transmission interferes with all other transmissions, it must hold that

$$\sum\nolimits_m w_m^k(\mathbf{h}) \leq 1, \quad \forall k. \tag{1}$$

This way, if $w_m^k(\mathbf{h}) = 0.9$ and $w_{m'}^k(\mathbf{h}) = 0.1$, the traffic of user m is assigned to channel k during the 90% of the duration of realization $\mathbf{h}$, the traffic of user m' is assigned to channel k during the remaining 10%, and traffic of no other user is scheduled. The theoretical results in Section IV will show that in practice $w_m^k(\mathbf{h}) \in \{0, 1\}$.

The resources adapted at the physical layer will be power and rate per user and subcarrier. Specifically, $p_m^k(\mathbf{h})$ and $r_m^k(\mathbf{h})$ will denote respectively, the instantaneous power and rate transmitted over channel k to user m during the channel realization $\mathbf{h}$ if $w_m^k(\mathbf{h}) = 1$. Spectrum mask constraints are imposed by enforcing that the *instantaneous* $p_m^k(\mathbf{h})$ can never exceed a maximum pre-specified level $\check{p}^k$; i.e,

$$p_m^k(\mathbf{h}) \leq \check{p}^k, \quad \forall k. \tag{2}$$

On the other hand, the maximum *average* power the AP can transmit (denoted by $\bar{p}$) is also bounded by $\check{p}$; hence,

$$\begin{aligned} \bar{p} \ &\leq\ \check{p}, \text{ where} && (3)\\ \bar{p} \ &=\ \mathbb{E}_{\mathbf{h}}\left[\sum\nolimits_{m,k} p_m^k(\mathbf{h}) w_m^k(\mathbf{h})\right]. && (4) \end{aligned}$$

Moreover, under bit error rate or capacity constraints, $p_m^k(\mathbf{h})$ and $r_m^k(\mathbf{h})$ are coupled. This rate-power coupling will be represented by the function $C_m^k(\mathbf{h}, p_m^k(\mathbf{h}))$, which is assumed to be increasing and strictly concave. For instance, if strong coding schemes are used, $C_m^k(\mathbf{h}, p_m^k(\mathbf{h}))$ is given by Shannon's capacity formula [13]: $\log(1 + h_m^k p_m^k(\mathbf{h}))$, which is certainly increasing and strictly concave. The one-to-one mapping between $p_m^k(\mathbf{h})$ and $r_m^k(\mathbf{h})$ implies that when the optimization problem is formulated, there is only need to optimize over one of them.

Regarding the transport and network layers, the operation is as follows. Packets are generated exogenously. Packet streams will be referred to as flows and there will be as many flows as users; i.e., one flow per user is considered. The amount of information of flow m that arrives at the AP at a given instant n is a random variable denoted by $a_m[n]$. The average arrival rate of exogenous information of flow m is denoted by $\bar{a}_m$. We will assume that the AP is equipped with M queues (one per flow) where incoming packets are stored before transmission. Let $q_m[n]$ denote the queue length for flow m at time slot n. Then, the queues obeys the recursion

$$q_m[n+1] = \left[q_m[n] + a_m[n] - \sum\nolimits_k r_m^k(\mathbf{h}[n]) w_m^k(\mathbf{h}[n])\right]_0^\infty, \tag{5}$$

for all m. In practice, arrivals and departures are magnitudes that vary with time scale smaller than n. This implies that definitions slightly different than the one in (5) are also possible. Such differences are not relevant for the subsequent analysis, and (5) has been chosen for mathematical simplicity.

For such queues to be stable (in the sense that $\lim_{n\to\infty} \frac{1}{n}\sum_{l=1}^{n} q_m[l] \leq \infty$), the following necessary condition needs to be satisfied

$$\bar{a}_m \leq \sum\nolimits_k \mathbb{E}_{\mathbf{h}}\left[r_m^k(\mathbf{h}) w_m^k(\mathbf{h})\right], \ \forall m. \tag{6}$$

That is typically known as necessary average flow conservation condition. Condition in (6), together with (1), (2), (3) and (4) need to be considered in the optimization problem presented in the next section.

III. Problem formulation

The optimal resource allocation will be obtained in this section as the solution of a constrained optimization problem. The *objective* of this problem will be based on *utility functions* $U_m(\bar{a}_m)$, that are commonly used in resource allocation tasks. Utility functions $U_m(\cdot)$ are chosen to be increasing (so that solutions which allow for higher arrival rates are promoted) and concave (so that fairness among different users is enforced). Moreover, power consumption is penalized using a cost function $J(\bar{p})$. Function $J(\cdot)$ is chosen to be increasing and convex. Finally, to effect QoS, a *minimum* average transmitted rate $\check{a}_m$ is guaranteed for certain users (elastic traffic).

Once the optimality criteria and the system operation conditions of the system have been established, the optimal cross-layer resource allocation is obtained as the solution of the following problem:

$$\min_{\substack{\bar{a}_m, \bar{p} \\ w_m^k(\mathbf{h}), p_m^k(\mathbf{h})}} \quad J(\bar{p}) - \sum_m U_m(\bar{a}_m) \tag{7a}$$

$$\text{subject to:}\ \ \bar{a}_m \geq \check{a}_m, \ \forall m \tag{7b}$$

$$(1),\ (2),\ (3),\ (4),\ (6). \tag{7c}$$

Users that do not need a minimum average rate requirement, will set their corresponding $\check{a}_m$ in (7b) to zero. Both the cross-layer nature and channel-adaptive nature of the resource allocation problem are apparent since: variables of different layers are jointly optimized in (7) and several of the optimization variables and constraints are functions of $\mathbf{h}$. It is important to stress that the dynamics of the queues were not explicitly taken into account in the formulation in (7). Only the necessary condition for stability in (6) has been explicitly considered. However, it will be shown in upcoming sections that the queues can in fact be used as a stochastic estimation of the Lagrange multipliers associated with constraint (6). This finding implies that the solution of (7) will also depend on the state of the queues. The solution of (7) is presented next.

IV. Optimal Cross-Layer Allocation

The problem in (7) can be trivially transformed into a convex problem (see [12] for details), which can be solved using a dual approach. First, we present the optimal solution as a function of the Lagrange multipliers (dual variables). Next, a stochastic scheme that for every instant n estimates the value of the multipliers is developed. Lastly, the relationship between stochastic multipliers and queues is established and the stability and average of the algorithm is proved.

A. Optimal Allocation as a Function of the Multipliers

Let π, ρ_m, and α_m denote, respectively, the Lagrange multipliers associated with the *average* constraints in (4), (6), and (7b). All these multipliers will be collected into vector $\boldsymbol{\lambda}$. There is no need for dualizing constraints (1) and (2) because the solution presented in this section will automatically fulfill them. Furthermore, let $(\dot{U}_m)^{-1}(\cdot)$, $(\dot{J})^{-1}(\cdot)$ and $(\dot{C}_m^k)^{-1}(\mathbf{h},\cdot)$ denote, respectively, the inverse function of the derivative of $U_m(\cdot)$, $J(\cdot)$ and $C_m^k(\mathbf{h},\cdot)$. Remember also that x^* stands for the optimum value of a given variable x. Based on this notation and using the optimality Karush-Kuhn-Tucker (KKT) conditions [14] associated with (7), it can be shown that the optimal cross-layer resource allocation is

$$\bar{a}_m^*(\boldsymbol{\lambda}^*) = \Big[(\dot{U}_m)^{-1}(\rho_m^* - \alpha_m^*)\Big]_0^{\infty}, \qquad (8)$$

$$\bar{p}^*(\boldsymbol{\lambda}^*) = \Big[(\dot{J})^{-1}(\pi^*)\Big]_0^{\check{p}}, \qquad (9)$$

$$p_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*) = \Big[(\dot{C}_m^k)^{-1}(\mathbf{h},\pi^*/\rho_m^*)\Big]_0^{\check{p}^k}, \qquad (10)$$

$$r_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*) = C_m^k(\mathbf{h}, p_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*)). \qquad (11)$$

Moreover, upon defining $f(m,k,\mathbf{h},\boldsymbol{\lambda}^*) := \rho_m^* r_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*) - \pi_m^* p_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*)$, which represents the rate reward minus power cost for user m in channel k; the optimal scheduling is

$$w_m^{k*}(\mathbf{h},\boldsymbol{\lambda}^*) = \mathbb{1}_{\{m=\arg\max_{m'}\{f(m',k,\mathbf{h},\boldsymbol{\lambda}^*)\}\}}, \ \forall k. \qquad (12)$$

Interestingly, the optimal management of resources depends only on the current channel realization $\mathbf{h}$ and on the multipliers $\boldsymbol{\lambda}^*$. While (8)-(11) are easy to derive and interpret, the scheduling in (12) needs a more detailed explanation. Equation (12) establishes that *per channel realization* $\mathbf{h}$, each channel k is uniquely assigned to the user that maximizes the functional $f(m,k,\mathbf{h},\boldsymbol{\lambda}^*)$; i.e., the scheduling follows a winner-takes-all strategy. The winner-takes-all strategy has been shown to be optimal for other problems that also deal with orthogonal sharing of resources among users [7], [13], [15].

B. Stochastic Estimation of the Multipliers

In the previous section, the optimal resource allocation schemes were characterized as a function of $\mathbf{h}$ and $\boldsymbol{\lambda}^*$. Different alternatives can be used to find $\boldsymbol{\lambda}^*$. Here we pursue an online approach under which the exact value of $\boldsymbol{\lambda}^*$ is never found. Differently, using stochastic approximation algorithms, at every time index n we aim to find an estimate of $\boldsymbol{\lambda}^*$, call it $\hat{\boldsymbol{\lambda}}[n]$, which remains sufficiently close of $\boldsymbol{\lambda}^*$; see, e.g., [10], [16]. The motivation for this approach is threefold: (i) the computational complexity of the stochastic schemes is small, (ii) the stochastic schemes do not need to know the channel distribution and can cope with channel non-stationarities; and (iii) connections between the stochastic multipliers and the lengths of the queues can be established.

Let μ denote a *constant* stepsize and $a_m^*[\hat{\boldsymbol{\lambda}}[n]]$ the instantaneous arrival of flow m during block n, which is a random variable drawn from a distribution with mean $\bar{a}_m^*(\hat{\boldsymbol{\lambda}}[n])$ given by (8). Then, the following updates for the multipliers are proposed:

$$\hat{\pi}[n+1] = \Big[\hat{\pi}[n] - \mu\big(\bar{p}^*(\hat{\boldsymbol{\lambda}}[n]) \qquad (13)$$
$$- \textstyle\sum_{m,k} p_m^{k*}(\mathbf{h}[n],\hat{\boldsymbol{\lambda}}[n]) w_m^{k*}(\mathbf{h}[n],\hat{\boldsymbol{\lambda}}[n])\big)\Big]_0^{\infty},$$

$$\hat{\rho}_m[n+1] = \Big[\hat{\rho}_m[n] + \mu\big(a_m^*(\hat{\boldsymbol{\lambda}}[n]) \qquad (14)$$
$$- \textstyle\sum_k r_m^{k*}(\mathbf{h}[n],\hat{\boldsymbol{\lambda}}[n]) w_m^{k*}(\mathbf{h}[n],\hat{\boldsymbol{\lambda}}[n])\big)\Big]_0^{\infty},$$

$$\hat{\alpha}_m[n+1] = \Big[\hat{\alpha}_m[n] + \mu\big(\check{a}_m - a_m^*(\hat{\boldsymbol{\lambda}}[n])\big)\Big]_0^{\infty}; \qquad (15)$$

where the primal variables in (13)-(15) (namely, powers, rates, scheduling ratios, and flow rates) are found by substituting $\hat{\pi}[n]$, $\hat{\rho}_m[n]$, and $\hat{\alpha}_m[n]$ into (8)-(12). Those will be referred to as stochastic primal variables or stochastic allocation. Basically, (13)-(15) are stochastic versions of the subgradient of the dual function of (7); see [14, Ch. 6] for details.

Assuming that the updates in (13)-(15) are bounded, the optimality of the stochastic schemes can be proved. Specifically, let define the cumulative running averages $\hat{\bar{p}}[n] := \frac{1}{n}\sum_{l=1}^{n}\sum_{m,k} p_m^{k*}(\mathbf{h}[l],\hat{\boldsymbol{\lambda}}[l]) w_m^{k*}(\mathbf{h}[l],\hat{\boldsymbol{\lambda}}[l])]$ and $\hat{\bar{a}}_m[n] := \frac{1}{n}\sum_{l=1}^{n} a_m^*[\hat{\boldsymbol{\lambda}}[n]]$. Then,

Proposition 1: *If the stepsize μ is sufficiently small, then as $n \to \infty$: (i) $\hat{\bar{p}}[n]$ and $\hat{\bar{a}}_m[n]$ are strictly feasible and (ii) the objective in* (7a) *is at least, $O(\mu)$ higher than that of the optimum non-stochastic solution of* (7).

Proposition 1 guarantees the feasibility and (almost) optimality of the stochastic schemes. On top of being almost optimal, these schemes are also meaningful because they help us to unveil/reveil the cross-layer behavior of our algorithm. Comparing (14) to (5), it is clear that $\hat{\rho}_m[n]$ and $q_m[n]$ are related in a way that *the stochastic Lagrange multipliers can be interpreted as a scaled version of the queue sizes*. Specifically, if $\hat{\rho}_m[0] = \mu q_m[0]$, then it follows that $q_m[n] = \hat{\rho}_m[n]/\mu$. This finding is extremely useful because it reveals a way in which the queues can be used to allocate resources in the network. In fact, using Proposition 1 it follows that if ρ_m^* in (8)-(12) is replaced with $\mu q_m[n]$, the resultant stochastic resource allocation is optimum as long as μ is sufficiently small. Furthermore, the previous finding is also relevant from other points of view: (i) to analyze the queue stability of our algorithms; (ii) to estimate the queueing delay that packets will experience; and (iii) to establish connections with other well-known cross-layer resource allocation algorithms (e.g., with the celebrated dynamic backpressure algorithm [8]). In the next subsection we elaborate on the two first points.

C. Delay Analysis

Although the dynamics of the queues had not been explicitly considered into the formulation of (7), they emerged naturally as a scaled stochastic estimation of the Lagrange multipliers associated with (6). In this section we will characterize the queue stability and the queueing delay of the developed resource allocation algorithms.

To do so, let $\hat{\bar{q}}_m[n] := \frac{1}{n}\sum_{l=1}^{n} q_m[l]$ denote the cumulative running average of $q_m[n]$, and $\bar{d}_m$ the average delay that flow m experiences. Then, the following is true:

Proposition 2: *Let δ_q and δ_d be small positive numbers proportional to the maximum update in* (14)*, then :*

$$\text{(i)} \quad |\hat{\bar{q}}_m[n] - \rho_m^*/\mu| < \delta_q \text{ as } n \to \infty \qquad \text{w.p. } 1; \quad (16)$$

$$\text{(ii)} \quad |\bar{d}_m - \rho_m^*/(\mu \bar{a}_m^*)| < \delta_d. \quad (17)$$

The result in (i) readily implies stability, because we have that $\bar{q}_m[\infty] < \infty$ as far as $\rho_m^* < \infty$; i.e., as far as the original problem is feasible. The result (ii) builds on (i) and on the fact that the average delay is given by the average aggregate queue length divided by the average aggregate arrival rate (Little's result). Equation (17) reveals that the average delay of our stochastic algorithm can be estimated based on the optimal solution of (7), and the stepsize of the proposed iterations. Moreover, the KKT conditions can be used to show that $\rho_m^* = \dot{U}_m(\bar{a}_m^*) + \alpha_m^*$. Substituting this into (17), it follows that $\bar{d}_m \simeq (\dot{U}_m(\bar{a}_m^*) + \alpha_m^*)/(\mu \bar{a}_m^*)$. This means that based on the exogenous arrival rates of a given flow, either the value of the average delay for that flow (when either the user does not have a minimum rate constraint or when the constraint is not active) or a lower bound on that value (when the minimum rate constraint is active) can be found.

Upon examining (17), it is also apparent that changes in the stepsize will induce changes in the average delay. Specifically, the higher the stepsize, the smaller the average queuing delay. The intuition is that high stepsizes will accelerate convergence and improve the ability of the schemes to react against events that otherwise, would increase the queuing delay. However, high stepsize values will also lead to more severe hovering in the dual domain and, when too big, will endanger convergence and stability. Equally interesting, (17) can also be used to effect different delay priorities. Key for this purpose is the fact that the iterations in (13)-(15) converge not only if the stepsize is the same for all the entries of $\boldsymbol{\lambda}$, but also if it is different for each entry. This way, flows/users that have stricter delay constraints can use a bigger stepsize. In other words, if we allow the stepsize to be dependent on m, i.e., μ_m; then different delay performances can be obtained.

V. Grouping channels

To implement the developed algorithms, the AP needs to know $(q_m[n], \hat{\alpha}_m[n])\ \forall m$, $\hat{\pi}[n]$, and $\mathbf{h}[n]$. In a downlink set-up $q_m[n]$, $\hat{\alpha}_m[n]$ and $\hat{\mu}[n]$ are available at the AP and the only problem is getting $\mathbf{h}[n]$. In Time Division Duplexing (TDD) systems, uplink and downlink channels are reciprocal. Hence, the AP can acquire $\mathbf{h}[n]$ by estimating the channel in the reverse link. However, in Frequency Division Duplexing (FDD) systems the forward and reverse channels are non-reciprocal, and therefore $\mathbf{h}[n]$ is not available at the AP. In that case, every user has to send the estimation of its own channels to the AP. If the channel varies sufficiently slow and the number of channels is not too high, the users could send an analog estimation of their channels. However, if it varies too fast or the number of channels is too high, the feedback rate needs to be reduced. In such scenario, it is convenient to exploit the correlation (if any) among channel gains. One of the simplest and most effective ways to exploit the correlation among channel gains is to group channels with highly correlated gains and assign the entire group to the same user. This is a common practice among wireless standards.

Our formulation can be slightly modified to account for this. The main required change is to modify the definition of $w_m^k(\mathbf{h})$. Specifically, let consider that channels are grouped into L groups (indexed by l) and that the channels in group l form the set $\mathcal{K}_l$. With $w_m^l(\mathbf{h})$ denoting the variable for user m being scheduled into group l, the following holds:

Proposition 3: *If $w_m^k(\mathbf{h})$ in* (7) *is replaced with $w_m^l(\mathbf{h})$, then:*
(i) *the optimal allocation in* (8)-(11) *holds;*
(ii) *the optimum scheduling is*

$$w_m^{l*}(\mathbf{h}, \boldsymbol{\lambda}^*) = \mathbb{1}_{\{m = \arg\max_{m'}\{\sum_{k \in \mathcal{K}_l} f(m',k,\mathbf{h},\boldsymbol{\lambda}^*)\}\}}.$$

There are different alternatives to implement this scheme. For example, the allocation could be done in two steps. First, the optimal group scheduling is found (this can be done based on a rough estimation of the channel gains; e.g., users only feed back the average gain for each group). Second, once the user who wins the full group is known, the AP can ask the winner to send a more accurate estimation of the channel. This will reduce the rate of signalling by a M factor. Another option is to implement only the first step. In this case the average gain is also used to find the final power and rate loadings. The latter alternative will reduce the signalling rate by a MK/L factor. This number can be significantly high. For example, it is known that in OFDM systems the loss of performance is negligible if L is $4-10$ times larger than the number of discrete taps of the multi-path channel [17]. Since the number of taps is typically in the order of 3-8 and the number of subcarriers is in the order of 64-2048, K/L can be in the order of 5-100.

VI. Numerical results

Numerical results are presented for OFDM downlink set-up with $M = 4$ users[2] and $K = 64$ parallel channels. The channel has 4 taps and the SNRs of each of the taps is exponentially distributed. The average power gain are 6dB, 4.5dB, 3dB and 1.5dB for users 1, 2, 3 and 4, respectively. The average power budget is $\check{p} = 30.0$; the power cost is $J(\bar{p}) = 0.001\bar{p}^2$; and the utilities to be maximized are $U_m(\bar{a}) = \log(1 + \bar{a})\ \forall m$. Table I lists the individual average rates, average power, sum-rate, and sum-utility for four different test-cases: (tc.1) an algorithm that maximizes the sum-rate; (tc.2) an algorithm that maximizes the sum-utility as in (7) but without minimum rate constraints, (tc.3) an algorithm that maximizes the sum-utility as in (7) and guarantees that $\bar{a}_1 \geq 35$, and (tc.4) a scheme that uses fixed subcarrier and (constant) power allocation but optimally adapts rate from a set of modes as the used in Wimax. The results in Table I corroborate that: (i) the proposed algorithm can provide fairness and guarantee minimum average rate requirements, and (ii) it performs better than existing alternatives.

The second simulation is devoted to analyze the dynamic behavior of the stochastic schemes developed in the paper for the test-case tc.3. Setting $\mu_m = 0.0002\ \forall m$, the plots in the first two rows of Figure 1 show the time-evolution of: the sample-average of the power transmitted by each user

[2] The relatively low number of users has been chosen because: (i) it facilitates the visualization of the plots, and (ii) the behavior and performance of our algorithm is similar than that for a large number of users.

TABLE I: Average rate, average power, sum-rate, and sum-utility for different test-cases (averages have been computed discarding the initialization period) . P^* denotes $P^* := J(\bar{p}) + \sum_m U_m(\bar{a}_m)$

	$[\bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4]$	$\bar{p}$	$\sum_m \bar{a}_m$	P^*
(tc.1)	$[28.3, 22.7, 16.2, 9.8]$	30.0	77.1	10.87
(tc.2)	$[24.3, 20.3, 16.7, 13.8]$	30.0	75.0	10.95
(tc.3)	$[35.0, 16.8, 13.8, 11.3]$	30.0	76.9	10.77
(tc.4)	$[11.1, 8.6, 6.4, 4.6]$	30.0	30.8	8.5

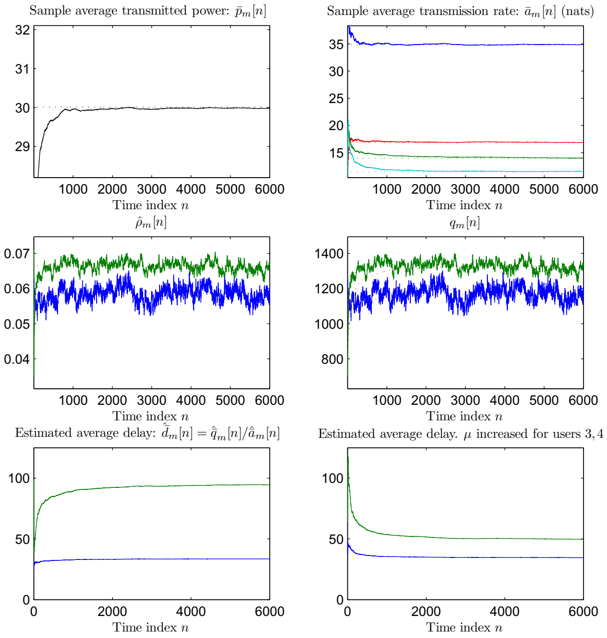

Fig. 1: Trajectories of different primal and dual variables.

$\hat{\bar{p}}_m[n] = n^{-1}\sum_{l=1}^{n} p_m[l]$; sample-average of the rate that arrives at each queue, $\hat{\bar{a}}_m[n] = n^{-1}\sum_{l=1}^{n} a_m[l]$, instantaneous estimation value of the stochastic multiplier, $\hat{\rho}_m[n]$; and both instantaneous and sample-average value of the queues, $q_m[n]$ and $\bar{q}_m[n] = n^{-1}\sum_{l=1}^{n} q_m[l]$. To facilitate visualization, only Lagrange multipliers and queues of users 1 and 3 have been plotted. In the first three subplots, dotted lines represent the values obtained from the optimal off-line solution. The results indicate that the proposed algorithm converges arbitrarily close to the optimal values [cf. Table I] in a finite number of iterations, that the queues are stable, and that the relationship between $\hat{\rho}_m[n]$ and $q_m[n]$ holds in practice. Moreover, to corroborate that changing the stepsize leads to different delay performance, we plot the estimated delay for two different stepsize values in the last row of Figure 1. Specifically, the subplots depict $\hat{\bar{d}}_m[n] = \hat{\bar{q}}_m[n]/\hat{\bar{a}}_m[n]$ vs. time for the case of: $\mu_m = 0.0002$ $\forall m$ (left plot) and $\mu_1 = \mu_2 = 0.0001$ and $\mu_3 = \mu_4 = 0.0004$ (right plot). Simulations confirm if the stepsize of each is modified separately, different delay performances can be obtained.

The last simulation tests the grouping scheme proposed in Section V. Four different values of L are considered: 8, 16, 32, and 64 (no feedback reduction). For each group the users feed back the smallest channel gain within the group (robust design). The utility loss is 4%, 1%, 0%, and 0%. Hence, results confirm that grouping techniques attain almost optimal performance even for small-medium values of L.

VII. Concluding Summary

This paper designed cross-layer algorithms to allocate resources (flows, channel access, power and rates) in a downlink system which transmits over a set of orthogonal parallel channels. The developed resource allocation is a function of the instantaneous fading, the length of the queues, and user-weights. The latter correspond to Lagrange multipliers and are estimated using stochastic approximation tools. Capitalizing on a relation between the estimates of the multipliers and the length of the queues, the stability and average delay were analyzed. Moreover, a mechanism to effect delay priorities among users by tuning the value of a stepsize parameter was also discussed. Finally, methods that reduced the signaling overhead by allocating groups of channels to the same user were briefly discussed at the end of the paper.

References

[1] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[2] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[3] Y.-H. Lin and R. L. Cruz, "Opportunistic link scheduling, power control, and routing for multi-hop wireless networks over time-varying channels," in *Proc. 43rd Annu. Allerton Conf. Commun., Control, and Computing*, Monticello, IL, Sep. 2005, pp. 976–985.

[4] P. Soldati, B. Johansson, and M. Johansson, "Proportionally fair allocation of end-to-end bandwidth in STDMA wireless networks," in *Proc. 7th ACM Int. Symp. Mobile Ad Hoc Networking and Computing*, Florence, Italy, May 2006, pp. 286–297.

[5] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[6] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and MAC for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2006.

[7] A. Stolyar, "Maximizing Queueing Network Utility Subject to Stability: Greedy Primal-Dual Algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, 2005.

[8] L. Georgiadis, M. Neely, and L. Tassiulas, "Resource Allocation and Cross-Layer Control in Wireless Networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-144, Apr. 2006.

[9] A. Ribeiro and G. B. Giannakis, "Optimal FDMA over wireless fading ad hoc networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, Mar.-Apr. 2008, pp. 2765–2768.

[10] A. G. Marques, G. B. Giannakis, and J. Ramos, "Stochastic Cross-Layer Resource Allocation for Wireless Networks Using Orthogonal Access: Optimality and Delay Analysis", in *Proc. IEEE Proc. of Intl. Conf. on Acoustics, Speech and Signal Process.*, Dallas, TX, Mar. 14-19, 2010.

[11] N. Gatsis, A. Ribeiro, G. B. Giannakis, "A class of convergent algorithms for resource allocation in wireless fading networks," *IEEE Trans. on Wireless Commun.*, vol.9, no.5, pp.1808-1823, May 2010.

[12] C.Y. Wong, R.S. Cheng, K.B. Lataief, and R.D. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

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

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

[15] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic Power Scheduling for Dynamic Multi-server Wireless Systems," *IEEE Trans. on Wireless Systems*, vol 5., no. 6, pp. 1506–1515, Jun. 2004.

[16] X. Wang and N. Gao, "Stochastic Resource Allocation in Fading Multiple Access and Broadcast Channels," *IEEE Trans. on Inform. Theory*, vol. 56, no. 5, pp. 2382-2391, May 2010.

[17] A. G. Marques, G. B. Giannakis, F. F. Digham, and F. J. Ramos, "Power Efficient Wireless OFDMA using Limited-Rate Feedback",*IEEE Trans. on Wireless Commun.*, vol. 7, no. 2, pp. 685-696, Feb. 2008.