A DECOMPOSITION METHOD FOR OPTIMAL USER ASSIGNMENT IN CELLULAR NETWORKS WITH ORTHOGONAL TRANSMISSIONS

Antonio G. Marques, Luis Cadarso, Eduardo Morgado and Carlos Figuera

King Juan Carlos University (Madrid, Spain), Dept. of Signal Theory and Communications

ABSTRACT

Effective operation of next-generation communication networks requires the deployment of a high number of base stations (BSs) capable of adapting dynamically their available resources to the changing environment. The resources include link layer variables (user-channel allocation and user-BS assignment) that, due to their binary nature, render the design challenging. This work proposes algorithms for *user-BS* allocation in cellular networks where users access orthogonally and close-by BSs use non-interfering channels. The *user-BS* allocation algorithms are designed jointly with the power, rate, and user-channel allocation, and take into account the dynamic environment. Three different algorithms are designed, each of them updates (adapts) the *user-BS* allocation at a different speed. We show that although the linear relaxation of all the binary variables is not optimal, a Benders' decomposition approach can be used to find the optimal solution. To accomplish this, we split the original problem so that the user-BS variables are isolated, relax the remaining binary variables, and solve the (sub-)problems iteratively.

***Index Terms*—** User association, Base station selection, Benders' decomposition, Stochastic optimization.

1. INTRODUCTION

Association of users with base stations (BSs) is a classical problem in wireless communications. Traditional designs assign users to BSs trying to maximize their signal-to-noise ratio (SNR), while balancing load among cells. In contrast to traditional networks, future networks are expected to be very dense and their state (SNRs, user locations, traffic load and quality of service requirements) is expected to change quickly. In such scenarios, the design of user-to-BS assignment algorithms is more challenging [1, 2]. The algorithms must account for the operating conditions of *each* BS, adapt quickly to changes in the environment and be jointly designed with the resource allocation (RA) schemes that optimize the cells' performance.

Many recent works have investigated the problem of user-BS assignment both in the context of cellular and heterogeneous networks [3, 4]. The network operating conditions considered are diverse: uplink and downlink [5, 6, 7, 8] optimization, fast and slow-varying fading scenarios [9, 10], and co-channel and orthogonal deployments [10, 11, 12]. Preliminary works relied more on heuristic approaches, while more recent works typically cast the user-BS allocation problem as NP-hard and then propose (either heuristic [13] or optimization-based [14]) approximated schemes to solve it. Powers, rates and/or beamformers are designed to maximize the aggregated rate while keeping interference under control. Interest in developing energy-efficient schemes [15, 7] has been growing in the last years. Similarly, there is an effort to incorporate additional variables to the state information [16] and to develop distributed solutions [14, 9].

In this paper, we design fast RA and user-BS assignment schemes for cellular networks with *orthogonal transmissions* [12, 17]. The schemes are obtained as the solution to a non-convex optimization, which aims at maximizing the aggregated transmit rate while minimizing the aggregated transmit power. Rate and power prices are allowed to be different across users, BSs and time instants, providing a means to represent the differences among users and BSs. Constraints accounting for orthogonal transmissions within the same BS, *orthogonal transmissions within close-by BSs* and allowing users to be connected to only one BS, are also considered. We design three algorithms with different user-BS assignment updating frequency. First, we assume that the user-BS assignment can be instantaneously adapted at each time slot; then, that it can be adapted only at the beginning of each planning period (which is composed of several slots); and, finally, that it can change at each slot, but penalizing each handover. Our main contribution is to develop algorithms capable of finding the *optimal solution* with moderate complexity. Key to the design is to use decomposition methods that split the variables into different sets, isolating the ones hard to optimize (binary user-BS assignment variables). Due to the structure of our problem (and more specifically, the interference model considered [12, 17]), the *Benders' Decomposition Approach* (BDA), which is used in stochastic programming to handle large scale mixed-integer-linear programs [18], is well-suited to address such decomposition. To facilitate exposition, we use several simplifying assumptions: perfect channel state information, single-antenna communications, and fixed channel-to-BS allocation, to name a few. However, the approach in the paper can be extended to scenarios where those assumptions do not hold true. Since the expressions for the power and channel assignment are similar to those in other RA problems, the main novelty is on the design of the user-BS assignment algorithms.

2. SYSTEM MODEL

Consider a deployment with J BSs, I users, and K frequency-flat orthogonal channels (indexed by j, i and k, respectively). Binary variables $w_{i,j}$ are used to represent the user-BS assignment. Specifically, $w_{i,j}$=1 if user i is connected to BS j, and $w_{i,j}$=0 otherwise. Similarly, binary variable $w_{i,j}^k$ will be one if channel k in BS j is assigned to user i, and zero otherwise. Because we consider orthogonal access if $w_{i,j}^k$=1, then it must hold that $w_{i',j}^k$=0 for all $i' \neq i$. This is guaranteed if $\sum_{\forall i} w_{i,j}^k \leq 1, \ \forall(j,k)$.[1]

This work was supported by the EU contract Tucan3G FP7-ICT-2011-9-601102; and the Spanish Ministry of Economy grant No TEC2013- 41604-R.

[1] To make the notation less heavy, limits in the summations and variables specifying (indexing) the constraints will be dropped whenever clear from the context. For example, constraint $\sum_{\forall i} w_{i,j}^k \leq 1, \ \forall(j,k)$, will be written as $\sum_i w_{i,j}^k \leq 1$.

The system operates in time slots t, whose duration typically corresponds to the coherence time of the channel (e.g., 10 milliseconds). The state of the overall system at time t will be denoted as $\mathcal{S}[t]$. At the very least, $\mathcal{S}[t]$ will include the fading channel coefficients, but it may also account for other variables relevant for the RA (such as locations, queue lengths, or battery levels). For simplicity, we will assume that the value of the stochastic process $\mathcal{S}[t]$ at time t is perfectly known. However, the results in the paper can be easily extended to the case of imperfect state information.

At each time t, the transmission power and rate at link (i,j,k) are allowed to change (for the sake of exposition, let us assume that we focus on the downlink channel, so that the transmitter is the BS). Let $p^k_{i,j}[t]$ denote the "nominal" power allocated to that link. By "nominal" we mean that power $p^k_{i,j}[t]$ is consumed only if both $w_{i,j}[t]$ and $w^k_{i,j}[t]$ are one. Under bit error rate or capacity constraints, the instantaneous nominal rate $r^k_{i,j}[t]$ and the nominal power $p^k_{i,j}[t]$ are coupled. To be more specific, let $h^k_{i,j}[t]$ denote the power fading coefficient divided by the noise and interference at link (i,j,k), and assume that rate is given by the capacity formula. Then, the (increasing) power-to-rate function can be written as $C^k_{i,j}(\mathcal{S}[t],p)=C^k_{i,j}(h^k_{i,j}[t],p):=\log_2(1+h^k_{i,j}[t]p)$.

The optimal RA is obtained as the solution to a constrained optimization problem. The design variables are $\mathcal{X}:=\{p^k_{i,j}[t],\,w^k_{i,j}[t],\,w_{i,j}[t],\ \forall(i,j,k)$ and $t=1,...,T\}$, where T denotes the planning horizon (typically in the order of seconds). Powers are constrained to be non-negative and the assignment variables to be binary. Maximum peak power constraints require $p^k_{i,j}[t]\leq p^{\max}_{i,j}[t]$. As already explained, orthogonal access requires $\sum_i w^k_{i,j}[t]\leq 1$. Finally, we consider that at given time instant t, a user i can be served by at most one BS. Mathematically, this is guaranteed if $\sum_j w_{i,j}[t]\leq 1$. The constraint could be modified to allow users to be connected to more than one BS without changing the basic structure of the problem.

The optimization aims at maximizing the aggregate rate while minimizing the aggregate power. To be specific, let $\rho_{i,j}[t]$ represent the price of the rate transmitted at time t and let $\pi_{i,j}[t]$ represent the cost of the power transmitted at time t. Using such prices (their physical interpretation will be discussed later on), the nominal reward if tuple (i,j,k) is activated at time t is defined as $\varphi^k_{i,j}[t]:=\rho_{i,j}[t]C^k_{i,j}(h^k_{i,j}[t],p^k_{i,j}[t])-\pi_{i,j}[t]p^k_{i,j}[t]$. The objective to be optimized is then $f_0(\mathcal{X}):=\sum_{t=1}^T\sum_{j,i}w_{i,j}[t]\sum_k w^k_{i,j}[t]\varphi^k_{i,j}[t]$. The reason to consider generic prices $\rho_{i,j}[t]$ and $\pi_{i,j}[t]$ is to keep our formulation general. In practice, they can represent (fixed or real-time) prices set by operators; static multipliers associated with average rate and power constraints; marginal prices associated with utility/cost functions that effect fairness [19]; or state variables (congestion, battery levels) accounting for the backhaul network [20]; to name a few. Allowing these prices to be different for each i and j is instrumental to cope with the diverse and quick-changing conditions of future networks. For example, if j is a small BS powered by a battery, costs $\pi_{i,j}[t]$ will be set to high values. Differently, BSs connected to the power grid and capable of handling many connections are expected to use values of $\rho_{i,j}[t]$ and $\pi_{i,j}[t]$ that yield high rewards (high $\rho_{i,j}[t]$ and small $\pi_{i,j}[t]$), so that assignments to them are promoted. Another example is a BS whose main purpose is to give service to a specific (predetermined) group of users. In that case, the values of $\rho_{i,j}[t]$ and $\pi_{i,j}[t]$ will vary sharply with i, so that the reward for internal users is much higher than that for external users. Any linear combination of the previous is also possible.

3. FAST USER-BS ASSIGNMENT

The first formulation considers that $w_{i,j}$ can be adapted at the same rate than $p^k_{i,j}$ and $w^k_{i,j}$ (i.e., every t). Hence, we aim at solving

$$\max_{\mathcal{X}}\ \sum_{t=1}^T\sum_{i,j,k}w^k_{i,j}[t]\varphi^k_{i,j}[t] \tag{1a}$$

$$\text{s. to}:\sum_i w^k_{i,j}[t]\leq 1,\ \ \sum_j w_{i,j}[t]\leq 1 \tag{1b}$$

$$w^k_{i,j}[t]=0 \text{ if } k\notin\mathcal{K}_j,\ w^k_{i,j}[t]\leq w_{i,j}[t] \tag{1c}$$

$$p^k_{i,j}[t]\in[0,p^{\max}_{i,j}],\ w^k_{i,j}[t]\in\{0,1\},\ w_{i,j}[t]\in\{0,1\}; \tag{1d}$$

where $\mathcal{K}_j$ is the set of channels that the operator allocated to BS j (if not all). Note that the objective in (1a) is not the same than the original one in $f_0(\mathcal{X})$. Specifically, variables $w_{i,j}[t]$ are not present in (1a). To guarantee that the nominal reward $\sum_k w^k_{i,j}[t]\varphi^k_{i,j}[t]$ is not considered if $w_{i,j}[t]=0$, constraint $w^k_{i,j}[t]\leq w_{i,j}[t]$ [cf. (1c)] was introduced. Clearly, if $w_{i,j}[t]\in\{0,1\}$ both formulations are equivalent (meaning that they yield the same objective value and the same effective power and rate values). The reason to write the problem as in (1) is twofold: i) the relaxed version of (1) (which deals with binary variables as if they were real variables in the $[0,1]$ interval) can be recast as a convex problem and ii) the structure of (1) is more amenable to be decomposed into smaller optimization problems (we will be more specific about this issue later on). Both issues will be critical to find efficient algorithms that solve the problem. Last but not least, additional constraints on $w_{i,j}[t]$ limiting the number of users per BS or allowing users to be connected to more than one BS can be incorporated to (1) without changing its basic structure.

Remark: *Interference among BSs.* The previous formulation (in particular, the definition of $C^k_{i,j}$) does not explicitly consider the effect that changing the power has on the interference at other BSs (the effect is considered implicitly, when acquiring the values of $h^k_{i,j}[t]$). This approach in not unusual in works dealing with user assignment. In [12] the interference is assumed to be negligible due to appropriate reuse factors or inter-cell interference cancellation techniques. In [17, 21, 22] rather than taking into account the dynamics of interference, it is assumed that transmissions are exposed to an average level of interference that depends only on long-term cell-load conditions. Hence, interference is assumed to be independent of the specific user assignment and scheduling. In an orthogonal-access configuration, interference can be considered negligible provided that if BS j and j' are close, then $\mathcal{K}_j$ and $\mathcal{K}_{j'}$ do not overlap. If the channel-BS allocation is not carried out beforehand (of if it does not guarantee orthogonality among close-by BSs), then constraints guaranteeing that conflicting BSs are not active simultaneously must be enforced: i.e., $w^k_{i,j}[t]+w^k_{i',j'}[t]\leq 1$ if (i,j) and (i',j') are in conflict.

3.1. Optimal Solution

Solving (1) requires optimizing jointly over the set of variables $\mathcal{X}$. To facilitate such a task, our approach is to split $\mathcal{X}$ into smaller sets and take advantage of the decomposable structure of (1). To be more specific, the idea is to leverage the fact the joint optimization over (x,y) of a generic function $f(x,y)$ can be performed as follows. Find first $x^*(y):=\arg\max_x f(x,y)$. Then, substitute $x^*(y)$ into the objective and solve $y^*=\arg\max_y f(x^*(y),y)$. Finally, find x^* as $x^*=x^*(y^*)$. For the problem in (1), $\mathcal{X}$ is split into three different sets of variables: $\mathcal{X}_1$, which accounts for the powers $p^k_{i,j}[t]$; $\mathcal{X}_2$, which accounts for all the channel allocation variables $w^k_{i,j}[t]$; and $\mathcal{X}_3$, which accounts for all the user-BS assignment variables $w_{i,j}[t]$. Next, we present several results dealing with the optimization of those subsets. Proofs are omitted due to space limitations.

We start with the optimization of $\mathcal{X}_1$ with $\mathcal{X}_2$ and $\mathcal{X}_3$ given, i.e., with finding $\mathcal{X}_1^*(\mathcal{X}_2, \mathcal{X}_3)$.

Proposition 1: *Optimal power allocation.* The optimal value of the nominal powers $\{p_{i,j}^{k*}[t]\ \forall(i,j,k,t)\}$ does not depend on the values of $\{w_{i,j}^{k*}[t], w_{i,j}^*[t]\}$ and can be found as $p_{i,j}^{k*}[t] = \arg\max_{0\leq p\leq p_{i,j}^{\max}} \rho_{i,j}[t]C_{i,j}(h_{i,j}^k[t], p) - \pi_{i,j}[t]p$.

The result in the proposition follows because: (1) can be decomposed (separated) across time; (1a) is non-decreasing with $p_{i,j}^k[t]$; and the value of $p_{i,j}^k[t]$ does not have an impact on $\mathcal{S}[t']$ for $t' > t$. Note also that the result deals with the *nominal* powers. Clearly, the *effective* transmit-powers *do* depend on $\{w_{i,j}^{k*}[t], w_{i,j}^*[t]\}$. [2]

Prop. 1 states that $\mathcal{X}_1^*(\mathcal{X}_2, \mathcal{X}_3) = \mathcal{X}_1^*$. The ensuing proposition deals with the optimization over $\mathcal{X}_2$ with $\mathcal{X}_3$ given and with $\mathcal{X}_1 = \mathcal{X}_1^*$. To simplify notation, the values of $p_{i,j}^{k*}[t]$ are used to define the *optimal* nominal reward per (j,k,t) tuple as: $\varphi_{i,j}^{k*}[t] = \rho_{i,j}[t]C_{i,j}(h_{i,j}^k[t], p_{i,j}^{k*}[t]) - \pi_{i,j}[t]p_{i,j}^{k*}[t]$.

Proposition 2: *Optimal channel-user assignment.* If $w_{i,j}[t] \in \{0,1\}$ are given and $p_{i,j}^k[t] = p_{i,j}^{k*}[t]$, then the optimal $w_{i,j}^{k*}[t] \in \{0,1\}$ can be found as follows. For each (j,k,t) tuple:
i) Define $\mathcal{I}_j^k[t] := \{i : w_{i,j}[t] = 1\ \&\ \varphi_{i,j}^{k*}[t] = \max_l \varphi_{l,j}^{k*}[t]w_{l,j}[t]\}$ as the set of winner users.
ii) Select one user from $\mathcal{I}_j^k[t]$ (the pick can be random), call it i', and then set $w_{i',j}^{k*}[t]$=1 and $w_{i,j}^{k*}[t]$=0 for all $i \neq i'$.

Prop. 2 states that the optimal channel assignment follows a greedy policy [23]. Specifically, for each (j,k,t) tuple, the policy first finds the users who are connected to the specific BS. Then, it selects the one with the highest optimal reward and assigns the channel to that user. If $\mathcal{I}_j^k[t]$ contains more than one user, any random pick is equally optimum. Clearly, this policy can be run in polynomial time.

To summarize the results so far, let $\tilde{f}_0(\mathcal{X})$ be the extended version of $f_0(\mathcal{X})$, so that $\tilde{f}_0(\mathcal{X}) = f_0(\mathcal{X})$ if $\mathcal{X}$ satisfies constraints (1b)-(1d), and $\tilde{f}_0(\mathcal{X})$=$-\infty$ otherwise. Our initial goal was to solve $\max_{\mathcal{X}} \tilde{f}_0(\mathcal{X})$. Upon splitting $\mathcal{X}$ into $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ we have that

$$\begin{aligned}\max_{\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3} \tilde{f}_0(\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3) &= \max_{\mathcal{X}_2,\mathcal{X}_3} \tilde{f}_0(\mathcal{X}_1^*(\mathcal{X}_2,\mathcal{X}_3),\mathcal{X}_2,\mathcal{X}_3)\\ = \max_{\mathcal{X}_2,\mathcal{X}_3} \tilde{f}_0(\mathcal{X}_1^*,\mathcal{X}_2,\mathcal{X}_3) &= \max_{\mathcal{X}_3} \tilde{f}_0(\mathcal{X}_1^*,\mathcal{X}_2^*(\mathcal{X}_1^*,\mathcal{X}_3),\mathcal{X}_3)\end{aligned} \quad (2)$$

The previous equations demonstrate that to solve the problem in (1), it suffices to execute the following steps: a) find $\mathcal{X}_1^*$ using Prop. 1; b) find $\mathcal{X}_3^*$ using the right hand side of (2), the output of step a) and Prop. 2; and c) find $\mathcal{X}_2^*$ using Prop. 2, $\mathcal{X}_1^*$ (step a) and $\mathcal{X}_3^*$ (step b).

As already discussed, the complexity to solve a) and c) is polynomial, so that the critical issue is how to solve the user-BS assignment in step b). The problem is not convex because variables $w_{i,j}[t]$ are binary. Moreover, it can be shown that, for the problem at hand, the relaxation $w_{i,j}[t] \in [0,1]$ is not tight. Several alternatives arise to solve the binary optimization. If the problem dimensionality is not very high, variants of the branch-and-bound algorithm for mixed integer programs can be used [24] (for each t we have to solve a problem with IJ binary variables and IJK real variables). If the dimensionality is too high, a reasonable choice is to rely on *decomposition* methods that solve for the binary variables iteratively and then use Props. 1 and 2 to solve for the remaining variables. One specially suited for the problem at hand is the BDA [18], which is extensively used in stochastic programming when dealing with very large-scale linear programs. Details on this method (and on how to use it for our problem) will be given in the next section.

4. SLOW USER-BS ASSIGNMENT

The formulation in (1) assumed that $w_{i,j}$ could be adapted at the same speed than $p_{i,j}^k$ and $w_{i,j}^k$. However, in many practical setups, changing the value of $w_{i,j}$ (which requires coordination among BSs) is more difficult than changing the value of $p_{i,j}^k$ and $w_{i,j}^k$ (which only requires coordination within the same BS). In this section, we solve the RA problem for $t = 1, ..., T$ assuming that $\{w_{i,j}\}$ remains constant during the planning horizon. To be specific, let $\mathcal{X}' := \{p_{i,j}^k[t], w_{i,j}^k[t], w_{i,j},\ \forall(i,j,k)$ and t=$1,...,T\}$ and solve [cf. (1)]

$$\max_{\mathcal{X}'} \ \sum_{t=1}^T \sum_{i,j,k} \mathbb{E}\Big[w_{i,j}^k[t]\varphi_{i,j}^k[t]\Big] \quad (3a)$$

$$\text{s. to}: \sum_i w_{i,j}^k[t]{\leq}1,\ \sum_i w_{i,j}{\leq}1,\ w_{i,j}^k[t] = 0 \text{ if } k \notin \mathcal{K}_j, \quad (3b)$$

$$w_{i,j}^k[t]{\leq}w_{i,j},\ p_{i,j}^k[t]{\in}[0, p_{i,j}^{\max}],\ w_{i,j}^k[t]{\in}\{0,1\},\ w_{i,j}{\in}\{0,1\}. \quad (3c)$$

Relative to (1), (3) not only replaces $w_{i,j}[t]$ with $w_{i,j}$, but also includes an expectation in (3a). As T grows large, the law of large numbers guarantees that the effect of including the expectation in (3a) vanishes. The reason to include the expectation is that $w_{i,j}$ must be found before the planning period starts (i.e., at $t = 0$) and, hence, the actual values of $\varphi_{i,j}^k[t]$ for $t = 1, ..., T$ (which are random) are not known[3]. The solution to (3) will be then found in two phases. The first phase (off-line phase) is executed at $t = 0$ and finds $w_{i,j}^*$. The second phase (online phase) is executed at each $t = 1, ..., T$ and finds $p_{i,j}^{k*}[t]$ and $w_{i,j}^{k*}[t]$ with $w_{i,j}^*$ given.

Strictly speaking, (3) is not separable across time. However, if the IJ values of $w_{i,j}$ are given, the problem can again be solved for each t separately. Equally important, if $w_{i,j}$ are given, $p_{i,j}^{k*}[t]$ and $w_{i,j}^{k*}[t]$ can be found in closed form (Props. 1 and 2), so that the computational complexity during the online phase is small. Moreover, we will see soon that this favorable structure can also be leveraged in the off-line phase. Although not investigated in this conference paper, convex algorithms can also take advantage of the aforementioned structure to design low-complexity *approximations* to $w_{i,j}^*$.

The remaining of the section is devoted to develop an algorithm to find $w_{i,j}^*$. Since the problem is not separable across time, the number of variables is much higher than that in Sec. 3 and branch-and-bound methods have to be discarded. Our approach to design an efficient algorithm builds on two observations. O1) After finding $p_{i,j}^{k*}[t]$, the remaining optimization is a mixed-integer-linear program (cf. Prop. 1). O2) The binary variables can be split into two sets: a smaller set $\mathcal{X}_3$ (containing $w_{i,j}$) with cardinality IJ that is difficult to optimize; and a larger set $\mathcal{X}_2$ (containing $w_{i,j}^k[t]$) with cardinality $IJKT$ that is easy to optimize (cf. Prop. 2). The BDA is specially suited for exploiting these properties. The main idea of the BDA is to decompose the problem into a master problem (dealing with the integer/binary optimization) and a subproblem (dealing with the real/linear optimization), which are solved iteratively (sequentially). Convergence to the optimal solution is guaranteed [18]. To describe the application of the BDA to the problem at hand more clearly, let l denote an iteration index and let $w_{i,j}^{(l)}$ and $w_{i,j}^{k,(l)}[t]$ denote, respectively, the solutions to the master problem and the subproblem *at iteration* l. The values of $w_{i,j}^{k,(l)}[t]$ are found using Prop. 2 with $w_{i,j} = w_{i,j}^{(l)}$. Solving the master problem is bit more intricate. For

[2] Due to space limits and to facilitate exposition, the paper assumes single-antenna communications. However, Prop. 1 also holds true for multiple-antenna scenarios. In such scenarios, beamformers/precoders for link (i,j,k) at time t should also be designed (adapted) to maximize $\varphi_{i,j}^k[t]$.

[3] Some works assume that the state $\mathcal{S}[t]$ is known beforehand (non-causally). If that were the case, the expectation in (3a) would not be required.

each l, the master problem needs to incorporate information of the optimal solution of the subproblems for iterations $l' < l$. In particular, the value of the Lagrange multipliers associated with the constraints of the subproblem have to be used as input to the master problem. To be rigorous, at iteration l, the master problem is

$$\{w_{i,j}^{(l)}\} = \arg\max_{\{z, w_{i,j}\}} z \tag{4a}$$

$$\text{s. to} : z \leq \textstyle\sum_{t,j,k} \left[\beta_{t,j,k}^{(l')} + \sum_i \Omega_{t,i,j,k}^{(l')} w_{i,j}\right], \;\; l' < l \tag{4b}$$

$$\textstyle\sum_i w_{i,j} \leq 1, \quad w_{i,j} \in \{0, 1\}; \tag{4c}$$

where z is an auxiliary variable, and $\beta_{t,j,k}^{(l')}$ and $\Omega_{t,i,j,k}^{(l')}$ are, respectively, the values of the multipliers associated with constraints $\sum_i w_{i,j}^k[t] \leq 1$ and $w_{i,j}^k[t] \leq w_{i,j}^{(l')}$ of the subproblem at iteration l'. Unfortunately, due to space limitations, the details of the derivation of (4) cannot be included in the paper; see [18, 25] for details. A pseudo-code summarizing the steps of the algorithm is given next.

Algorithm 1: *Off-line phase.*

[*Step1*] Master problem: If $l = 0$, initialize $w_{i,j}^{(0)}$ so that each user is assigned to the closest BS. If $l \geq 1$, solve (4) using the outputs of [Step2] for $l' < l$.

[*Step2*] Subproblem: Use Props. 1 and 2 to solve (3) over $p_{i,j}^k[t]$ and $w_{i,j}^k[t]$, with $w_{i,j} = w_{i,j}^{(l)}$. Find and store the value of multipliers $\beta_{t,j,k}^{(l)}$ and $\Omega_{t,i,j,k}^{(l)}$.

[*Step3*] If $w_{i,j}^{(l-1)} = w_{i,j}^{(l)}$, stop. If not, set $l = l+1$ and go to [Step1].

Regarding the algorithm details, to deal with the expectations in (3), we use a MonteCarlo approach that draws samples of $\mathcal{S}[t]$. Moreover, in order to accelerate Benders convergence speed, we use Pareto-optimal cuts [26], [27]; we also find closed forms for the dual variables in [Step2], which reduce computational times for large-scale case studies. For the test cases considered in this paper, our algorithm converges in 3-10 iterations. As already explained, the algorithm to be run during the online phase consists in using Props. 1 and 2 for each t using as $w_{i,j}[t]$ the output generated by Alg. 1 (which is the same for all t within the planning horizon).

5. PENALIZING FAST USER-BS CHANGES

In Sec. 3, $w_{i,j}$ was allowed to change at every time t. In Sec. 4, our algorithm operated in two time scales and allowed $w_{i,j}$ to be updated only at the slow scale (every T slots). The second algorithm entails lower signalling costs, but it also achieves a lower aggregate reward (it has to satisfy constraints $w_{i,j}[t] = w_{i,j}[t-1]$, which are not imposed to the first algorithm). In this section, we follow a hybrid approach so that: i) $w_{i,j}$ is allowed to change at every time t; and ii) a new cost that penalizes user-BS updates is considered. By tuning the penalty cost, one can make the hybrid scheme as close as the schemes in Secs. 3 or 4 as desired. The basic idea is to modify the solution in Sec. 3. To that end, let $s_{i,j}[t] \in \{0, 1\}$ represent switching variables (which are one if $w_{i,j}[t] \neq w_{i,j}[t-1]$) and $\lambda_{i,j}[t]$ the non-negative cost of updating the value of $w_{i,j}[t]$. Using those conventions, now t the problem to be solved *at each time* is

$$\max_{\mathcal{X}_1[t], \mathcal{X}_2[t], \tilde{\mathcal{X}}_3[t]} \textstyle\sum_{i,j,k} w_{i,j}^k[t] \varphi_{i,j}^k[t] - \sum_{i,j} \lambda_{i,j}[t] s_{i,j}[t] \tag{5a}$$

$$\text{s. to} : (1b), \; (1c), \; (1d) \tag{5b}$$

$$- s_{i,j}[t] \leq w_{i,j}[t] - w_{i,j}[t-1] \leq s_{i,j}[t], \; s_{i,j}[t] \in \{0, 1\}; \tag{5c}$$

where $\mathcal{X}_1[t] = \{p_{i,j}^k[t]\}_{\forall i,j,k}$, $\mathcal{X}_2[t] = \{w_{i,j}^k[t]\}_{\forall i,j,k}$, and $\tilde{\mathcal{X}}_3[t] = \{w_{i,j}[t], s_{i,j}[t]\}_{\forall i,j}$. The main difference relative to (1) is the incorporation of the penalty in (5a) and the constraints in (5c), which require $s_{i,j}[t]$ to be one if $|w_{i,j}[t] - w_{i,j}[t-1]| = 1$. The approach to solve (5) is similar to that to solve (1). If variables in $\tilde{\mathcal{X}}_3^*[t]$ (which now also include $s_{i,j}[t]$) are known, then $\mathcal{X}_1^*[t]$ and $\mathcal{X}_2^*[t]$ can be found using Props. 1 and 2. To solve for $\tilde{\mathcal{X}}_3^*[t]$ either branch-and-bound algorithms or a small modification of the BDA presented in Alg. 1 can be used. We stress that the formulation proposed in (5) considers $w_{i,j}[t-1]$ as given and does not need to account for the effect of the optimal RA on future time instants.

The value of $\lambda_{i,j}[t]$ must be set based on the operating conditions of the system. For example, if $\lambda_{i,j}[t] = 0$ the solution to (5) is the same than that in Sec. 3. Similarly, if the user-BS assignment is initialized using Algorithm 1 and one sets $\lambda_{i,j}[t] = \infty$, the algorithm reduces to that in Sec. 4. A more "sophisticated" alternative is to set a maximum rate of handovers (say η) and then rely on stochastic dual methods [28, 29] to update the cost as $\lambda_{i,j}[t+1] = \max\{0, \lambda_{i,j}[t] + \mu(s_{i,j}[t] - \eta)\}$, where μ is a small stepsize.

6. NUMERICAL RESULTS

Due to space limits, only results for two test cases are presented. Additional results will be available in the journal version of the paper. We consider a 2x2 kilometers (km) grid, where I=120 users are located uniformly at random and J=12 BSs are located at positions (in km): (0.5,0.5), (0.5,1.5), (1.5,1.5), (1.5,1.5), (0.65,0.85), (0.25,0.85), (1,0.9), (1.75,1), (1.25,0.75), (0.9,1.5), (0.3,1.75), (1.25,1.25). There are K=32 channels, each with a bandwidth of B_c. Channel gains follow a space-free model and a block fading Rayleigh channel is assumed. BSs $j = 1, 2, 3, 4$ transmit 10 watts and the remaining ones 2 watts. Noise level is set so that the average SNR in the grid (relative to the closest BS) is 15 dB. The planning horizon is T=100 and results are averaged over 30 planning horizons. Prices are set so that $\rho_{i,j}[t] = I/J + \rho_i$ and $\pi_{i,j}[t] = \pi_j$. The values of ρ_i and π_j are set so that, for the algorithm in Sec. 3, the BSs transmit with their maximum power and all users receive at least 10 bits per $1/B_c$. Users move with a speed of 5 meters per T.

Since we focus on designing algorithms for user-BS association, all tested schemes implement the optimal power and user-channel policies in Props. 1 and 2. We are interested in comparing the performance of: sc1) the scheme in Sec. 3; sc2) the scheme in Sec. 4; sc3) the scheme in Sec. 5 with $\lambda_{i,j} = 0.01$; sc4) the scheme in Sec. 5 with $\lambda_{i,j} = 5$; sc5) the scheme in Sec. 5 where $\lambda_{i,j}$ are tuned so that $\eta = 5\%$; sc6) a heuristic scheme that assigns the user to the BS with the highest *instantaneous* SNR; and sc7) a scheme that assigns the user to the BS with the highest *average* SNR.

The absolute gap relative to the optimal solution and the percentage of handovers are: (0.00%, 33.7%) sc1; (3.55%, 2.7%) sc2; (0.03%, 32.4%) sc3; (8.67%, 2.7%) sc4; (5.35%, 4.8%) sc5; (13.9%, 5.3%) sc6; and (19.7%, 1.7%) sc7). If only 4 of the small BSs and the 32 closest users are considered, and T is reduced to $T = 10$, the results are (0.00%, 11.8%) sc1; (16.2%, 0.6%) sc2; (0.01%, 10.9%) sc3; (12.1%, 4.0%) sc4; (2.90%, 5.1%) sc5; (4.14%, 6.40%) sc6; and (23.8%, 0.4%) sc7. If the schemes in Sec. 5, are not initialized with the solution in Sec. 4, the loss (gap) increases around 0.5-1%. Results confirm the validity or the theoretical claims and show moderate gains for the more advanced schemes. This seems to suggest that if the power and user-channel assignment are adapted at a fast rate, the rate of adaptation of the user-BS assignment is not critical. Another interesting point is that by tuning the value of $\lambda_{i,j}$ [cf. (5a)], the rate of handovers can be effectively kept under control. Simulations also reveal that the schemes in Sec. 5 are not very sensitive to initialization. Hence, they constitute an interesting alternative for practical deployments. Lastly, the close performance of the proposed algorithms, motivate the development of suboptimal approximations to the optimal solution.

7. REFERENCES

[1] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: Key challenges for the radio-access network," *IEEE Veh. Technology Mag.*, vol. 8, no. 3, pp. 47–53, Sep. 2013.

[2] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T.Q.S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, June 2011.

[3] P. Hande, S. Patil, and H.G. Myung, "Distributed load-balancing in a multi-carrier wireless system," in *IEEE Wireless Commun. and Networking Conf. (WCNC)*, Apr. 2009, pp. 1–6.

[4] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J.G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[5] M. Hong and A. Garcia, "Mechanism design for base station association and resource allocation in downlink OFDMA network," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 11, pp. 2238–2250, Dec. 2012.

[6] A. Alsawah and I. Fijalkow, "Base-station and subcarrier assignment in two-cell OFDMA downlink under QoS fairness," in *IEEE Intl. Sympos. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Sept. 2008, pp. 1–6.

[7] K. Son, S. Nagaraj, M. Sarkar, and S. Dey, "QoS-aware dynamic cell reconfiguration for energy conservation in cellular networks," in *IEEE Wireless Commun. and Networking Conf. (WCNC)*, Apr. 2013, pp. 2022–2027.

[8] J. Lin, Y. Li, and Q. Peng, "Joint power allocation, base station assignment and beamformer design for an uplink simo heterogeneous network," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, May 2014, pp. 434–438.

[9] V.N. Ha and L.B. Le, "Distributed base station association and power control for heterogeneous cellular networks," *IEEE Trans. Veh. Technology*, vol. 63, no. 1, pp. 282–296, Jan. 2014.

[10] S. Singh and J.G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[11] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

[12] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[13] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[14] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3214–3228, June 2013.

[15] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, Sept. 2011.

[16] H. Galeana and R. Ferrus, "Enhanced base station assignment approach for coping with backhaul constraints in OFDMA-based cellular networks," in *European Wireless Conf. (EW)*, Apr. 2010, pp. 254–260.

[17] A.J. Fehske, H. Klessig, J. Voigt, and G.P. Fettweis, "Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks," *IEEE Trans. Veh. Technology*, vol. 62, no. 5, pp. 1974–1988, June 2013.

[18] J.F. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numerische Mathematik*, vol. 4, no. 1, pp. 238–252, 1962.

[19] A.G. Marques, L.M. Lopez-Ramos, G.B. Giannakis, J. Ramos, and A.J. Caamaño, "Optimal cross-layer resource allocation in cellular networks using channel-and queue-state information," *IEEE Trans. Veh. Technology*, vol. 61, no. 6, pp. 2789–2807, July 2012.

[20] N. Gatsis and A.G. Marques, "A stochastic approximation approach to load shedding inpower networks," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, May 2014, pp. 6464–6468.

[21] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[22] A.J. Fehske and G.P. Fettweis, "Aggregation of variables in load models for interference-coupled cellular data networks," in *IEEE Intl. Conf. Commun. (ICC)*, June 2012, pp. 5102–5107.

[23] A.J. Goldsmith and K. Li, "Capacity and optimal resource allocation for fading broadcast channels I: Ergodic capacity," *IEEE Trans. Info. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[24] G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization*, Wiley-Interscience, New York, NY, USA, 1988.

[25] G. Codato and M. Fischetti, "Combinatorial benders' cuts for mixed-integer linear programming," *Operations Research*, vol. 54, no. 4, pp. 756–766, Aug. 2006.

[26] T.L. Magnanti and R.T. Wong, "Accelerating benders decomposition: Algorithmic enhancement and model selection criteria," *Operations Research*, vol. 29, no. 3, pp. 464–484, 1981.

[27] J. F. Cordeau, G. Stojković, F. Soumis, and J. Desrosiers, "Benders decomposition for simultaneous aircraft routing and crew scheduling," *Transportation Science*, vol. 35, no. 4, pp. 375–388, Nov. 2001.

[28] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6369–6386, Dec. 2010.

[29] A.G. Marques, L.M. Lopez-Ramos, G.B. Giannakis, and J. Ramos, "Resource allocation for interweave and underlay CRs under probability-of-interference constraints," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1922–1933, Oct. 2012.