Hierarchical Optimization of User Association and Flexible TDD Allocation for Access and Backhaul Networks

Nikolaos Sapountzis[1], Thrasyvoulos Spyropoulos[1], Navid Nikaein[1], and Umer Salim[2]

[1]Mobile Communications Department, EURECOM, 06410, Biot, France, firstname.lastname@eurecom.fr

[2]TCL Communications Limited, Sophia Antipolis, 06410, France, umer.salim@tcl.com

***Abstract*—The success of future heterogeneous networks (HetNets) heavily depends on the interplay between user association and resource allocation on both the access and backhaul network. While user association is key to improve both the user and network performance, it is becoming a multi-objective optimization problem that should consider the number and type of base station in range. Furthermore, the increasing spatio-temporal heterogeneity in downlink(DL) and uplink(UL) traffic suggests that DL/UL resources can be tuned to optimally serve the respective workload. Split DL/UL association and flexible TDD offer such an opportunity. While much literature exists on these problems, the majority consider them separately. In this work, we develop a framework that tackles the optimal interplay of (i) user-association, (ii) radio resource allocation, and (iii) backhaul resource allocation of TDD resources, for a family of objective functions. We propose an algorithm that reduces the complexity of this problem by decomposing it into three optimization subproblems, each potentially solved by a different network element and at different timescales. We prove convergence to the global optimum, and provide simulation results that demonstrate the performance benefits of our approach.**

I. Introduction

LATELY, heterogeneous network (HetNet) deployments have been widely considered in 4G and beyond wireless networks. They are composed of conventional macro cells (MC) overlaid with a set of low-power small cells (SC). Due to the increasing number and type of base stations (BS) within the range of each user, the problem of user association becomes increasingly important. More advanced schemes beyond simple SINR-based ones are thus needed [1], [2] to balance user- and network-related performance goals.

While optimization of most current networks revolves around the downlink (DL) performance, social networks, augmented reality and other Machine Type Communication (MTC) applications suggest that uplink (UL) performance becomes as important. Recent approaches that aim to improve both DL and UL throughput suggest that UL/DL association should be in fact decoupled for optimal performance. As one example, a user equipment (UE) could be connected to a macro BS in the DL (from which it receives the highest signal level), and to an SC in the UL (where the pathloss is lower) [3], [4]. However, if the DL resources of the macro BS, or the UL resources of the SC are not sufficient, this approach can lead to *unnecessary* congestion or under-utilization in either direction.

Typically, in today's systems, each BS is given an amount of bandwidth resources to utilize for both DL and UL traffic by duplexing on the frequency (Frequency Division Duplex-FDD) or the time (Time Division Duplex-TDD) domain. While conventional networks are mainly designed for FDD or pre-configured TDD schemes, heterogeneous traffic demand, desired architectural flexibility, and scarcity of spectrum has increased interest in *flexible TDD* schemes, that can *match the UL and DL resources to the actual demand* [5].

Nevertheless, dynamic/flexible TDD schemes require additional considerations, in particular in asymmetric interference scenarios (see e.g. Figure 1). As a typical example, if an SC is doing UL while a nearby MC is transmitting on the DL (with much higher power), the performance of the SC might be significantly degraded from this *cross-interference*. Enhanced Inter-Cell Interference Coordination (eICIC) schemes such as Almost Blank Subframes (ABS) could alleviate this but only to some extent [6], [7]. Large amounts of mismatch might lead to excessive usage of resources for eICIC, instead of user traffic, leading instead to considerable performance degradation. Many additional allocation schemes have further been proposed to tackle this problem(s) [8] [9] [10], most of them revolving around a key-enabler for 5G networks, namely "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA), standardized in LTE-A Release 13 [11]. However, it is not clear which scheme is the best option and how it should interact with user association.

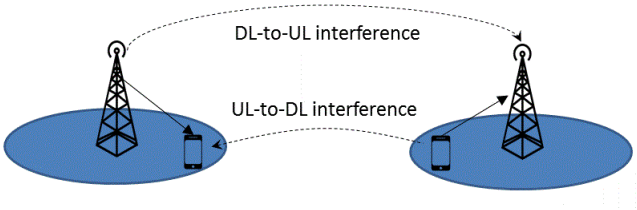


Fig. 1: Cross interference scenarios.

Additionally, a common limitation of most of the above works, and of others in that area, is that they focus solely on the radio access part, ignoring the backhaul (BH) network. This might be reasonable for legacy cellular networks, given that the macro-cell backhaul is often over-provisioned (e.g., fiber). However, expected backhaul limitations for small

cells [12] and the additional backhaul load for coordinated transmission (CoMP) and eICIC put a heavy toll on backhaul links, that might become the new bottleneck. This calls for a joint optimization of radio and backhaul e.g. [13], [14], [15], [16], [17]. Nevertheless, these works mostly focus on the DL. In our previous work [18] we have analytically derived optimal UL and DL user association rules for various backhaul-limited scenarios. However, we assumed a fixed and pre-allocated amount of resources for UL and DL, for both the radio access as well as backhaul, and we concluded that such a pre-configured resource allocation significantly penalizes performance. Undoubtedly, backhaul resource allocation policies should interact with the *user association* and *flexible TDD* radio access policies, in order to satisfy the UL and DL traffic demands that the latter generate. However, the analysis of such an interaction is not well understood yet and it is neither clear nor trivial how to properly address it, even though it is a key for the success of future HetNets.

Finally, most of the existing works rely on a centralized controller, e.g., for user association or TDD schemes etc., to improve load balancing or throughput [19], [20], [21], [8]. Such a *centralized* entity shall govern the BSs, the UEs, and potentially the backhaul links with access to all the necessary information. However, such an implementation may not be applicable. Additionally, even when it is applicable, it may (a) allow only for slow adaptation on the queuing statistics at relatively long timescales by failing to adapt to sharp fluctuation, since such a controller is usually implemented in a server deep in the core network, as well as (b) require excessive message overhead and computational complexity that increase exponentially in the network size (e.g. in the number of BSs, backhaul links, or users). Future MTC and Internet-of-Things (IoT) applications are expected to bring online more than 50 billion devices soon, by emerging the dramatic increase of the network sizes. Thus, to avoid relying on such a controller that is prone to failures, current systems shall aim on more distributed implementations.

In this paper, we propose an optimization framework that jointly considers all these problem dimensions. To our best knowledge, this is the first work to attempt it. Our main contributions can be summarized as follows:

(1) We propose an analytical model and then algorithm to study the interplay between (i) user association, (i) radio access resource allocation with cross-interference management, and (iii) backhaul resource allocation, significantly extending the popular framework of [1]. (Section II and Section III)

(2) We show that the joint problem is non-convex, unlike variants studies in the past [1], [18], [3], [17], but possesses some "hidden" convexity properties that allows its decomposition into three subproblems. These subproblems can be solved through convex optimizers, at possibly different elements (e.g. UE, BS, backhaul link), and at different timescales, facilitating an hierarchical implementation converging to the global optimum under some certain circumstances. (Section IV)

(3) Using extensive simulations, we highlight the complex trade-offs involved between the different subproblems, and show that significant performance improvements could be achieved compared to current standards. (Section V)

(4) We show that our framework allows for totally distributed implementations, is of low computational complexity, highly scalable, offers flexible performance optimization, and that is extendable to different future work directions (Section VI).

TABLE I: Notation

	Downlink	**Uplink**
Key control variables		
Access Resource Allocation Policy for BS i	ζ_i	$1-\zeta_i$
Backhaul Resource Allocation Policy for link k	$Z(k)$	$1-Z(k)$
Normalized load of BS i ($\zeta_i \to 1$ and $\zeta_i \to 0$)	ρ_i^D	ρ_i^U
Other variables		
Traffic arrival rate (flows/sec) at location x	$\lambda^D(x)$	$\lambda^U(x)$
Max. rate of BS i BS at location x	$c_i^D(x)$	$c_i^U(x)$
Load estimate of BS i (used for the broadcast)	$\hat{\rho}_i^D$	$\hat{\rho}_i^U$
Load density of BS i at location x	$\rho_i^D(x)$	$\rho_i^U(x)$
BS i max rate requirement for backhaul	$\tilde{c}_i^D$	$\tilde{c}_i^U$
(Effective) load of BS i	ρ_i^D/ζ_i	$\rho_i^U/(1-\zeta_i)$
Load-balancing degree	α^D	α^U
Association chance of location x with BS i	$p_i^D(x)$	$p_i^U(x)$
Penalty indicator for congestion at BH link k	$\mathcal{J}^D(k)$	$\mathcal{J}^U(k)$
Penalty indicator for cross interf. between BS i,j	$\mathcal{I}_{ij}$	

II. System Model and Assumptions

We use a similar problem setup as the one used in a number of related works [3], [1], [18], [22], and extend it accordingly. To keep notation consistent, for all variables considered, the superscript "D" and "U" refer to downlink and uplink traffic, respectively. For brevity, in the following *we present most notation and assumptions in terms of downlink traffic only, assuming that the uplink case and notation is symmetric*. Specific differences will be elaborated, where necessary. In Table I, we summarize some useful notation.

A. Traffic Model

(A.1 - Traffic arrival rates) Traffic at location $x \in \mathcal{L}$ consists of file (or more generally *flow*) requests arriving according to an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$[1]. Each new arriving request is for a *downlink (DL)* flow, with probability z^D, or *uplink* (UL) flow with probability $z^U = 1 - z^D$. Using a Poisson splitting argument [23], it follows that the above gives rise to 2 independent, Poisson flow arrival processes with rates

$$\lambda^D(x) = z^D \cdot \lambda(x), \;\; \lambda^U(x) = z^U \cdot \lambda(x). \tag{1}$$

(A.2 - Flow characteristics) *Flow-sizes* (in bits) are drawn from a generic distribution with mean $1/\mu^D(x)$.

B. Access Network

(B.1 - Access network topology) We assume an area $\mathcal{L} \subset \mathbb{R}^2$ served by a set of base stations $\mathcal{B}$, that are either macro BSs (eNBs) or small cells (SCs).

(B.2 - Access Resource Allocation Policy) Each BS $i \in \mathcal{B}$ is associated with a total bandwidth w_i, and a resource allocation parameter $0 < \zeta_i < 1$ which reflects the amount of radio resources (e.g., time, frequency, space) available for DL

[1]As we are interested in the aggregation of all flows from all locations x assosicated to BS i, even if flow arrivals at each location are not Poisson the Palm-Khintchine theorem [23] suggests that Poisson assumption could be a good approximation for the input traffic to a BS.

transmissions. Without loss of generality, we focus on time resources, as e.g. in the context of the envisioned flexible TDD standard. [2] Hence, the (long-run) resources of BS i allocated to DL are $\zeta_i \cdot w_i$, whereas the UL ones are $(1-\zeta_i)\cdot w_i$, where ζ_i is a key *control variable* of our problem.

(B.3 - DL physical data rate) Each BS $i \in \mathcal{B}$ is associated with a transmit power P_i. It can deliver a *maximum* physical data transmission rate of $c_i^D(x,\zeta_i)$ to a user at location x in absence of any other flows served, given by Shannon capacity[3]

$$c_i^D(x,\zeta_i) = \zeta_i \cdot w_i \cdot \log_2(1+\text{SINR}_i(x)), \tag{2}$$

where $\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j\neq i} G_j(x)P_j+N_0}$. N_0 is the noise power, and $G_i(x)$ represents the path loss and shadowing effects between the i-th BS and the UE located at x (as well as antenna and coding gains, etc.)[4]. We assume that effects of fast fading are filtered out, and that the total intercell interference at location x is static, and considered as another noise source, as in most aforementioned works [1], [22], [18].

(B.4 - Load density) We introduce the *load density* at x

$$\rho_i^D(x,\zeta_i) = \frac{\lambda^D(x)}{\mu^D(x)c_i^D(x,\zeta_i)}, \tag{3}$$

which is the contribution of location x to the total load of a BS i, when location x is associated with BS i. Assume a simple example that there are only two locations x_1 and x_2 associated with a certain BS (we skip the subscripts i). Then, the total arrival rate of this BS is $\lambda = \lambda(x_1) + \lambda(x_2)$. The mean service time per flow $E[S]$, depends on the chance that a flow comes from x_1 or x_2, which is $\frac{\lambda(x_1)}{\lambda(x_1)+\lambda(x_2)}$ and $\frac{\lambda(x_2)}{\lambda(x_1)+\lambda(x_2)}$, respectively. Hence, $E[S] = \frac{\lambda(x_1)}{\lambda(x_1)+\lambda(x_2)} \cdot \frac{1}{\mu(x_1)c(x_1)} + \frac{\lambda(x_2)}{\lambda(x_1)+\lambda(x_2)} \cdot \frac{1}{\mu(x_2)c(x_2)} = \frac{\rho(x_1)+\rho(x_2)}{\lambda(x_1)+\lambda(x_2)}$. The BS load is $\rho = (\lambda(x_1)+\lambda(x_2))\cdot E[S] = \rho(x_1)+\rho(x_2)$.

(B.5 - BS load) Each location x is associated with routing probabilities $p_i^D(x) \in [0,1]$, which are the probabilities that DL flows generated for users at location x get associated with (i.e., are served by) BS i. Generalizing the simple example of B.4 to multiple locations x with infinitesimal load, we can define the total *load*, or *utilization*, for BS i as

$$\rho_i^D(\zeta_i) = \int_{\mathcal{L}} p_i^D(x)\rho_i^D(x,\zeta_i)dx. \tag{4}$$

Clearly, this load depends on (and is *coupled by*) the *new* control variables ζ_i, related to the UL/DL allocation problem. To make this relation explicit, in the following we will use the *normalized* load variables $\rho_i^D = \rho_i^D(\zeta_i = 1)$, i.e. the load when all resources are used for DL (similarly for UL). Note also that Eq. (4) is a generalization of a well known queueing result for servers with multiple traffic types (each location x corresponding to a different traffic type) [23], [24]. We are interested in the flow-level dynamics of this system, and model the service of DL flows at each BS as a queueing system with effective load $\frac{\rho_i^D}{\zeta_i}$. Also, since we are interested in the aggregation of all flows at BS level (i.e. all flows from all locations x associated with BS i), even if flow arrivals at each x is not Poisson (as in A.1), the Palm-Khintchine theorem [23] suggests that Poisson assumption is a good approximation for the BS input traffic. Note that ρ_i, directly associated with $p_i(x)$ as Eq. (4) shows, is the second set of our considered key *control variables*.

(B.6 - Scheduling) Proportionally fair scheduling is often implemented in LTE networks due to its good fairness and spectral efficiency properties [25]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system [23]. It is multi-class because each flow might get different rates for similarly allocated resources, due to different channel quality and modulation and coding scheme (MCS) observed at x.

(B.7 - Performance impact of BS load) The stationary number of flows in BS i is equal to $E[N_i] = \frac{\rho_i^D/\zeta_i}{1-\rho_i^D/\zeta_i}$ [23]. Hence, minimizing ρ_i^D/ζ_i minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that BS [23]. Also, the throughput for a flow at location x is $\zeta_i \cdot c_i^D(x)\cdot(1-\rho_i^D/\zeta_i)$. This observation is important to understand how the user's physical data rate $\zeta_i \cdot c_i^D(x)$ (related to users at location x only) and the BS load ρ_i^D/ζ_i (related to *all* users associated with BS i) affect the optimal association rule (e.g., in Eq. (16)).

(B.8 - UL/DL association split) In the following, we will assume that a UE is able to associate with up to two BSs, one for its DL and one for UL traffic, as proposed in LTE Rel. 12 [26]. However, our framework is backward compatible when joint UL/DL association is required (see Section VI).

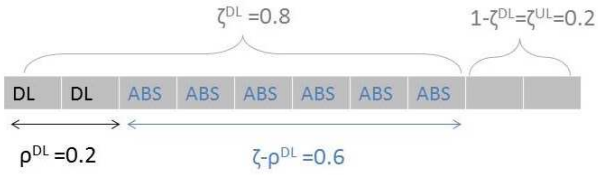


Fig. 2: A frame example for a certain BS.

(B.9 - UL/DL cross interference avoidance) Without loss of generality, we assume that each BS i cross interferes with a subset of other BSs $\mathcal{C}_i \subseteq \mathcal{B} \setminus \{i\}$. In practice, a distance based rule, or alternatively the cell cluster concept, can be used to determine these sets. If i is on the DL and a BS $j \in \mathcal{C}_i$ on the UL (or vice versa) then these BSs might cause severe interference to each other (that invalidates assumption B.3). We refer to this as *cross interference*. A sufficient condition to avoid cross-interference is

$$\rho_i^D + \rho_j^U \leq 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \tag{5}$$

We explain the above condition here. Consider two such BSs i and j. If $\zeta_i = \zeta_j$ then there is no cross-interference, because i and j can synchronize their DL (and UL) slots to avoid it. If $\zeta_i \neq \zeta_j$, cross-interference might occur, but *it also depends on the effective loads*. ζ_i slots are *at most* used for DL. But out of these only $\frac{\rho_i^D}{\zeta_i}\cdot\zeta_i = \rho_i^D$ will be busy (since $\frac{\rho_i^D}{\zeta_i}$ is the utilization of the downlink resources, according to B.5-B.7). The rest of the DL slots $(1-\frac{\rho_i^D}{\zeta_i})\cdot\zeta_i = \zeta_i - \rho_i^D$ could be blanked with ABS

[2] Although traditional LTE systems only allow some fixed and predefined values for ζ_i (depending on the TDD configuration), we relax them to be more generally applicable.

[3] We use Shannon capacity for clarity of presentation. However, our approach could be easily adapted to include modulation and coding schemes.

[4] In the UL, we assume that the Tx power of each user is P^{UE}, and slightly abuse notation for SINR, G, etc., as these don't play a major role later.

frames (see also Fig. 2). Similarly, the percentage of slots that j will be *active* on the UL is $\frac{\rho_j^U}{1-\zeta_j} \cdot (1-\zeta_j) = \rho_j^U$ slots. Hence, if $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i + \frac{\rho_j^U}{1-\zeta_j} \cdot (1-\zeta_j) \leq 1$, there are enough different slots in a frame to schedule all DL and UL of i and j without any overlap. Taking care for all such links on the interference graph, gives us Eq.(5). Finally, we stress that this constraint applies to the long-term allocation policy of resources. The actual MAC scheduling may still allocate resources in those time slots to transmissions that are non-interfering.

C. Backhaul Network

(C.1 - Backhaul network topology) Each access network node (either eNB or SC) is connected to the core network through an eNB aggregation gateway via a certain number of backhaul links that constitute the backhaul network. This connection can be either direct ("star" topology) or through one or more SC aggregation gateways ("mesh" topology).

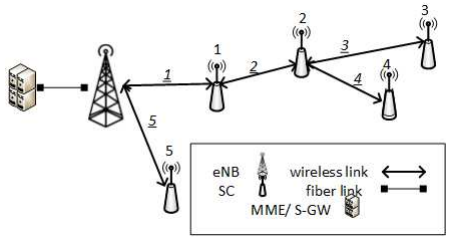


Fig. 3: Future Backhaul topology of a HetNet.

Without loss of generality, we assume that there is a fiber link from the eNB to the core network, and focus on the set of capacity-limited backhaul links (wired or wireless) connecting SCs to the eNB, denoted as $\mathcal{B}_h$. We denote as routing path $\mathcal{B}_h(i)$ the set of all backhaul links $j \in \mathcal{B}_h$ along which traffic is routed from BS i to an eNB aggregation point, and we assume that it is *given* (e.g., calculated in practice as a Layer 2 (L2) spanning tree). For example, in Fig. 3, $\mathcal{B}_h(1) = \{1\}$, and $\mathcal{B}_h(3) = \{1,2,3\}$. We further denote as $\mathcal{B}(j)$ the set of all BS $i \in \mathcal{B}$ whose traffic is routed over backhaul link j. E.g., $\mathcal{B}(1) = \{1,2,3,4\}$ and $\mathcal{B}(2) = \{2,3,4\}$ in Fig. 3.

(C.2 - Backhaul Resource Allocation Policy) Each $j \in \mathcal{B}_h$ backhaul link is associated with a total capacity $C_h(j)$. While traditional backhaul links are multiplexed using FDD, nowadays TDD gains more ground due to the performance improvements it promises [27]. So, in the context of TDD, we introduce the backhaul resource allocation parameter $0 < Z(j) < 1$, that splits the backhaul capacity of the j link between DL ($Z(j) \to 1$) and UL ($Z(j) \to 0$). In other words, $Z(j) \cdot C_h(j)$ and $(1-Z(j)) \cdot C_h(j)$ correspond to the total resources that the j backhaul link allocates for the DL and UL traffic, accordingly, where $Z(j)$ is another key *control variable* of our problem. Note that, backhaul links usually don't implement any particular scheduling algorithm, so they can be seen as a data "pipe".

(C.3 - Backhaul capacity requirement) The DL capacity requirement of a backhaul link $j \in \mathcal{B}_h$ in terms of bits per second consists of the sum of DL loads of all BSs using that link ($i \in \mathcal{B}(j)$) [18]

$$\sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D}{\zeta_i} \cdot (\zeta_i \cdot \tilde{c}_i^D) = \sum_{i \in \mathcal{B}(j)} \rho_i^D \cdot \tilde{c}_i^D. \quad (6)$$

For example, if a single BS i only uses backhaul link j (e.g. a star topology), and i has a load of $\rho_i^D = 0.7$, i.e., is active 70% of the time on the downlink, then the average downlink *rate* on backhaul j will be $0.7 \cdot \tilde{c}_i^D$. As for $\tilde{c}_i^D$, this is a parameter tuned by the operator. It could be directly replaced with the average rate considering all possible locations (e.g. as in [24]). However, this is a rather optimistic value to use, and would lead to backhaul link capacities being violated often. Conversely, the use of peak rate (i.e. assuming the maximum MCS used for every active flow) corresponds to the most conservative choice for this parameter. However, it is well known that this is much higher than the average "busy" rate [12], and would lead to backhaul resources being wasted too often. Finally, the direct usage of $p_i(x)$ to derive $\tilde{c}_i$ would not only complicate significantly the problem at hand, but is also somewhat superfluous since in most "busy" scenarios the average rate mostly depends on the edge users [12] and does not change much. We therefore leave this to the operator as a design parameter, to set it depending on how conservative he wants to be and past statistics.Note that Eq. (6) is neither the BS load nor the backhaul link load but simply the *total rate requirement on the backhaul link* (which should not exceed capacity).

(C.4 - Backhaul provisioning) For each backhaul link $j \in \mathcal{B}_h$, we have formulated for the DL direction: the available resources given the allocation scheme ($Z(j) \cdot C_h(j)$, see C.2) and the capacity requirement ($\sum_{i \in \mathcal{B}(j)} \rho_i^D \cdot \tilde{c}_i^D$, see C.3). Each of these links shall introduce a backhaul *constraint* to avoid exceeding its maximum capacity and prohibit backhaul congestion $\forall j \in \mathcal{B}_h$

$$\begin{aligned} \sum_{i \in \mathcal{B}(j)} \rho_i^D \cdot \tilde{c}_i^D &< Z(j) \cdot C_h(j), \\ \sum_{i \in \mathcal{B}(j)} \rho_i^U \cdot \tilde{c}_i^U &< (1-Z(j)) \cdot C_h(j). \end{aligned} \quad (7)$$

Throughout this paper, we assume that the backhaul network is either *under-provisioned* if the capacity of at least one backhaul link is exceeded, or *provisioned* otherwise.

(C.5 - Interference-free Backhaul) Modern backhaul architectures are developed using (highly) directional P2P or P2MP static architectures [28]. These are planned topologies and thus cross interference between BH links with asymmetric UL/DL schedules can be considered negligible.

III. Optimization Problem

We are now ready to formulate our optimization framework, and jointly study the problems of (i) *user association*, (ii) *access*, and (iii) *backhaul TDD resource allocation*. We remind the reader that the variables associated with these problems are: ρ_i^D, ρ_i^U (see B.5), $\zeta_i, \forall i \in \mathcal{B}$ (see B.2) and $Z(j), j \in \mathcal{B}$ (see C.2), respectively.

We start by defining the feasible region of these optimization variables. This can be delimited by the requirement that the effective load of no BS (Eq. (8e)) as well as the TDD allocation policies of no BS (Eq. (8d)) and no BH link (Eq. (8f)) in either DL/ UL direction being exceeded.

Definition 1. ***(Feasible set)*** *If ϵ is an arbitrarily small positive constant, the feasible region of $(\rho^D;\rho^U;\zeta;Z) = ((\rho_1^D,\rho_2^D,\ldots,\rho_{\|\mathcal{B}\|}^D);(\rho_1^U,\rho_2^U,\ldots,\rho_{\|\mathcal{B}\|}^U);(\zeta_1,\zeta_2,\ldots,\zeta_{\|\mathcal{B}\|});(Z_1,Z_2,\ldots,Z_{\|\mathcal{B}_h\|}))$ is*

$$\mathcal{F} = \Big\{(\rho^D,\rho^U,\zeta,Z) \mid \rho_i^y = \int_{\mathcal{L}} p_i^y(x)\rho_i^y(x)dx, \tag{8a}$$

$$\sum_{i\in\mathcal{B}} p_i^y(x) = 1, \tag{8b}$$

$$0 \le p_i^y(x) \le 1, \ \ \forall x \in \mathcal{L}, \ \ y \in \{U,D\}, \tag{8c}$$

$$0+\epsilon \le \zeta_i \le 1-\epsilon, \tag{8d}$$

$$0 \le \frac{\rho_i^D}{\zeta_i}, \frac{\rho_i^U}{1-\zeta_i} \le 1-\epsilon, \ \ \forall i \in \mathcal{B}, \tag{8e}$$

$$0+\epsilon \le Z(j) \le 1-\epsilon, \ \ \forall j \in \mathcal{B}_h\Big\} \tag{8f}$$

Lemma 3.1. *The feasible set $\mathcal{F}$ is convex.*

Proof: The proof for the feasible set $\mathcal{F}$ without the last three constraints can be found in [1]. Constraints (8d), (8f) are linear, and constraint (8e) refers to the image of ρ under different perspectives. So they preserve convexity [29], and the complete feasible set remains convex. ■

We now proceed into our cost function. Following our previous work [18] we extend the proposed cost function that only considers the BS loads ρ_i^D,ρ_i^U, to also include the resource allocation variables $\zeta_i, \forall i \in \mathcal{B}$. The operator may weigh the importance of DL and UL traffic performance with a parameter $\tau \in [0,1]$. α^D controls the amount of load balancing desired in the DL resources, and α^U in the UL. Let $\alpha = [\alpha^D;\alpha^U]$, where α^D and α^U can have different values.

Definition 2. ***(Cost function)*** *Our α-fair cost function that considers the access network performance is*

$$\phi_\alpha(\rho,\zeta) = \sum_{i\in\mathcal{B}} \tau\frac{(1-\frac{\rho_i^D}{\zeta_i})^{1-\alpha^D}}{\alpha^D-1} + (1-\tau)\frac{(1-\frac{\rho_i^U}{1-\zeta_i})^{1-\alpha^U}}{\alpha^U-1}, \text{if } \alpha^D,\alpha^U \ne 1. \tag{9}$$

If α^D is equal to 1, the respective fraction must be replaced with $\log(1-\frac{\rho_i^D}{\zeta_i})^{-1}$. The respective α-fair functions can capture different objectives such as maximizing spectral efficiency ($\alpha = 0$) or throughput ($\alpha = 1$), minimizing mean per flow delay ($\alpha = 2$), and maxmin load-balancing ($\alpha \to \infty$); similarly for the UL.

Lemma 3.2. *The cost function $\phi_\alpha(\rho,\zeta)$ is a multi convex function, i.e., it is convex in ρ for fixed ζ, and versa.*

Proof: The objective function is the sum of the basic α function $\frac{(1-\frac{\rho}{\zeta})^{1-\alpha}}{\alpha-1}$ over different BSs, with $(\rho,\zeta) \in \mathcal{F}$. When ζ is fixed this is the simplest form of the well known α-fair function which is clearly convex in ρ. And so is the corresponding sum over all BSs (sum preserves convexity). For fixed ρ, the basic α function is also convex in ζ (it has non-negative second derivative, namely $2\rho\zeta^{-3}(1-\rho/\zeta)^{-\alpha} + \alpha\rho^2\zeta^{-4}(1-\rho/\zeta)^{-\alpha-1} \ge 0$), and so does its sum. ■

Summarizing, in Definition 1 we formulated the feasibility set of our control variables: ρ,ζ,Z, and in Definition 2 we defined our α-fair cost function considering the radio network performance using ρ,ζ. Two things remain to formulate our optimization problem. Firstly, to consider the *backhaul network constraints* defined in C.4 that (i) include the third dimension Z into the picture, and (ii) further delimit our solution space with respect to the available link capacities. Secondly, to also consider the *cross-interference constraints* with respect to the dynamic TDD slots, as explained in B.8. The direct consideration of these constraints is of utmost importance to avoid performance degradation in multiple scenarios (e.g. when backhaul capacities starts becoming under-provisioned or when a BS doing DL starts interfering with the UL of a neighboring one) as explained earlier. This will also be demonstrated later in the simulations with numerical examples.

Definition 3. ***(Optimization Problem 1)*** *The joint user association, radio access and backhaul resource allocation problem can be expressed as*

$$\begin{aligned} \min_{\rho,\zeta,Z}\{\phi_\alpha(\rho,\zeta) \mid \ & (\rho,\zeta,Z) \in \mathcal{F}\}, \\ \text{subject to} \ & \text{Eq. (5)}, \\ & \text{Eq. (7)}. \end{aligned} \tag{10}$$

Lemma 3.3. *Optimization Problem 1 is a multi convex minimization problem.*

Proof: This is a multi convex optimization problem since the objective function and the affine constraints are multi convex on the (multi) convex feasible set $\mathcal{F}$. For an analytical survey on bi/multi convex optimization problems, we refer the interested reader to [30]. ■

IV. Bottom-Up Optimization Approach

The first thing we notice is that this problem, unlike the original one and other variants, is *non convex* and thus the standard fixed point method or other convex solvers cannot be directly applied. Our solution direction is further delimited, baring in mind our additional constraints in terms of *complexity* and *implementation*, e.g. that our solution should not follow an exhausting searching procedure (e.g. as the multiple start branch for such non convex problems) that puts an unaffordable toll on the complexity based on the variable size. Or baring in mind our requirement that our method should be *distributed*, e.g. to be easily adoptable, to be scalable, and not to depend on centralized entities that are prone to failures.

Solution Roadmap. Following that direction, we suggest that the complex Optimization Problem 1 can be actually decomposed into three simpler optimization sub-problems, each potentially solved by a different network element, and at different timescale. This facilitates a multi-level hierarchical decomposition allowing, as it will turn out also later, for distributed implementations with multiple desirable properties under certain circumstances. We now sketch our approach to solve our three-level optimization problem using the bottom-up method. For the simplest scenario with solely one optimization level, where one tackles the convex problem of (i) user association, the problem is tractable by convex solvers and we refer the interested reader to past variants works including [1], [18]. For the joint consideration of

(i) user association and (ii) radio access resource allocation problems, we sketch a two level hierarchical decomposition algorithm, Algorithm 1, that globally converges as shown in Section IV-A; we analytically focus and tackle each decoupled problem in Sections IV-A1 and IV-A2, accordingly. Finally, for the complete problem that also considers the (iii) backhaul resource allocation problem we extend the latter algorithm and we build Algorithm 2 using an additional optimization level, as shown in Section IV-B.

A. Two-level Optimization Algorithm.

In this section we focus on the joint problem of (i) *user association*, and (ii) *access resource allocation*, i.e.

$$\begin{aligned} &\min_{\rho,\zeta}\{\phi_\alpha(\rho,\zeta) \mid \quad (\rho,\zeta) \in \mathcal{F}\}, \\ &\text{subject to} \qquad \text{Eq. (5)}. \end{aligned} \tag{11}$$

In the following we attempt to sketch an algorithm that solves this problem, and we assume that the backhaul network is over-provisioned and we thus forget about the backhaul network and the Z variables. (We come back to the complete problem later in Section IV-B). Note that this is a multi convex problem of two variables, also called biconvex.

The nonconvex objective in Eq. (11) is block separable in ρ^D, ρ^U. Indeed, if we fix ζ, the problem decomposes in two simpler problems with variables ρ^D and ρ^U, that are coupled from constraint (5), and so we call ζ the *complicating* variable. Therefore, it makes sense to decompose the objective into two levels of optimization, following the *primal decomposition method* [31]. Specifically, at the lower level there are *two sub-problems* that run in parallel, that aim to find the optimal values of ρ^{*D} and ρ^{*U}, namely $\rho^* = [\rho^{*D}; \rho^{*U}]$, upon a fixed ζ. At the higher level we encounter the *master problem*, where we attempt to update (and eventually optimize), the complicating variable ζ. Note that constraint (5) only depends on ρ and thus does not affect the master problem. Formally, the sub-problems (that we encounter at the lowest level) and the master problem (that we encounter at the higher level) are

$$\min_\rho\{\phi_\alpha(\rho,\zeta)\} \text{ subj. to Eq. (5) (sub-problems)} \tag{12}$$

$$\min_\zeta\{\phi_\alpha(\rho,\zeta)\} \quad \text{(master problem)} \tag{13}$$

The above decomposed problems are convex since the joint problem of Eq. (11) is biconvex. Thus, they can efficiently be tackled through convex optimizers.

Our proposed iterative algorithm is sketched in Algorithm 1.

Algorithm 1 Two-level Optimization Algorithm that solves the user association and access TDD allocation problem.

1: **Repeat** until $\|\zeta^{(m)} - \zeta^{(m-1)}\| < \epsilon$.
2: *(Update the master problem (Section (IV-A2)).)*
3: Resource allocation: $\zeta \rightarrow$ DL, $1-\zeta \rightarrow$ UL.
4: *(Solve the two subproblems (Section (IV-A1)).)*
5: Derive ρ^{*D} given the available resources (ζ).
6: Derive ρ^{*U} given the available resources ($1-\zeta$).

At the (m) iteration step the master problem allocates the available resources by directly giving each sub-problem the amount of resources that it can use ($\zeta^{(m)}$ for the DL and $(1-\zeta^{(m)})$ for the UL traffic) [Algorithm 1 line 3]. Then, we solve the two sub-problems (derive $\rho^* = [\rho^{*D}; \rho^{*U}]$) based on their given resources currently and the coupling constraint [Algorithm 1 line 5-6]. In the next iteration $(m+1)$, we update the complicating parameter (derive $\zeta^{(m+1)}$), and re-solve the two sub-problems. We repeat the process until $\zeta^{(m)}$ converges to a stationary point [Algorithm 1 line 1]. In Section IV-A1 we present an iterative algorithm running at the UE that efficiently derives the optimal BS load vector ρ^* given a fixed $\zeta^{(m)}$ or simpler ζ (see also Lemma 4.3). Similarly, in Section IV-A2 we present the rule to descently update the resource allocation $\zeta^{(m)}$ given a fixed ρ^* or simpler ρ (see also Lemma 4.4).

Now, we show convergence of this algorithm through the next Lemma. The proof can be found in Section VIII-A.

Lemma 4.1. *Let $\{(\rho^*,\zeta)^{(\mathbf{m})}\}$ be the sequence generated within the (m) iterations by Algorithm 1, when the two (lower level) sub-problems are solved on a faster timescale than the (higher level) master problem. Then, any limiting point of $\{(\rho^*,\zeta)^{(\mathbf{m})}\}$ is the global optimum of the problem of Eq. (11).*

1) Two-level Optimization Algorithm: sub-problem running at the UE.: In this section we focus on the two sub-problems of Algorithm 1, by keeping a fixed resource allocation policy ζ. As discussed, these sub-problems, also defined in Eq. (12), correspond to the DL and UL user association problem. Our objective is to provide distributed user association rules that should run at the UE level and push the control variable $\rho = [\rho^D; \rho^U]$ to converge to its optimum value $\rho^* = [\rho^{*D}; \rho^{*U}]$ given the current access resource allocation ζ.

An efficient way to tackle the coupling cross-interference constraints in a distributed implementation setup is to directly include the constraints in the objective as *penalty functions* that increase the objective when a cross-interference constraint is violated [29]. We can then solve the new *unconstrained* problem, coupled by the penalty constant $\gamma > 0$, as follows

$$\min_\rho \left\{ \Phi_\alpha(\rho,\zeta,\gamma) = \phi_\alpha(\rho,\zeta) + \gamma \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^D + \rho_j^U - 1)^2 \right\}, \tag{14}$$

where $\phi_\alpha(\rho,\zeta)$ is our α-fair cost function discussed earlier (see e.g. Definition 2). The desired penalty is introduced from the sum using the indicator variable $\mathcal{I}_{ij}$ that reveals whether BS i cross interferes with BS j, i.e.

$$\mathcal{I}_{ij} = \begin{cases} 1, & \text{when } \rho_i^D + \rho_j^U > 1 \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

This penalty function is quadratic on the amount of excess cross interference. (Quadratic penalty functions like the above are common [32] and preserve the convexity[5].) Parameter γ can be chosen as a small constant, introducing a "soft" constraint (i.e., cross-interference could be slightly exceeded, if this really improves our main objective), *or preferably be increased monotonically, so as to converge to a "hard" constraint* [32].

We now tackle the problem of minimizing $\Phi_\alpha(\rho,\zeta,\gamma)$ in ρ as seen in Eq. (14). We start our discussion by assuming that γ is fixed to a small positive constant referring to "soft"

[5]This can be easily seen, since the function $(x+y-1)^2$ has Hessian matrix the $[2,2;2,2]$, and so it is positive semidefinite and convex.

cross interference constraints. We remind the reader that in this section ζ is also assumed to be static. To deal with this minimization problem, we sketch a distributed iterative algorithm, where at each iteration step our cost function is monotonically improved, and two parts are involved: the *mobile device* and the *BS*. We assume that the starting point $\rho^{(0)}$ is feasible.

Mobile Device. At each iteration step (l), each user at location x receives a BS broadcast message and simply associates with the BS maximizing the respective quantity as shown in the following theorem, in a distributed manner. This updates the association variables $p_i(x)$.

Theorem 4.2. *If $\rho^{(l)} = (\rho_1^{(l)}, \rho_2^{(l)}, \cdots, \rho_{\|\mathcal{B}\|}^{(l)})$ denotes the current BS load vector, the DL association rule for a user at location x giving a descent direction is expressed by (similarly in UL)*

$$i^D(x) = \arg\max_{i\in\mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{user knowledge}} \cdot \overbrace{P_i^D}^{\text{BS broadcast message}} \right) \quad (16)$$

where $P_i^D = \dfrac{\zeta_i \cdot \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma\cdot\zeta_i \cdot \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{(l),D} + \rho_j^{(l),U} - 1)}$.

The proof of the above Theorem can be found in Section VIII-B. Regarding the derived rules, we note that when the interference constraints for the BS i are not violated (i.e., $\mathcal{I}_{ij} = 0, \forall j \in \mathcal{C}_i$), the above theorem states that the optimal downlink associations are the same as the one in [1]. However, when the BS i cross interferes with another BS, an additional term is added in the denominator that penalizes BS i making it less preferable to users at location x. Note that the amount of penalization depends on the amount of *total* cross interference (sum term) from nearby BSs. This penalization makes sense, since additional users to that BS would increase its effective load as well as its DL busy slots, and thus increase the cross interference. Finally, note that our derived rules are scalable, since each BS i needs to only broadcast one value per dimension (i.e., P_i^D for downlink) no matter the number of cross-interfering BSs $i \in \mathcal{C}_i$, they are of low-complexity since the user only needs to perform a simple max operation to find the best BS for association, and they offer flexible performance given the α value.

Base Station. Each BS maintains an estimate $\hat{\rho}_i$ of its average utilization load. To deal with the utilization constraint ($\rho_i < 1 - \epsilon$), the parallel and potentially asynchronous updates of $p_i(x)$ variables, and non-stationarities in the traffic demand, the BS load estimate $\hat{\rho}_i$ is updated regularly as follows:

$$\hat{\rho}_i^{(l+1)} = (1 - \beta^{(l)}) \cdot \rho_i^{(l)} + \beta^{(k)} \cdot \hat{\rho}_i^{(l)}. \quad (17)$$

This is an exponential moving average with parameter $\beta^{(l)} \in [0, 1)$. $\rho_i^{(l)}$ is the current load measurement (derived e.g. using Eq. (4)) while $\hat{\rho}_i^{(l)}$ is the current load average estimate. $\hat{\rho}_i^{(l+1)}$ is used for the next iteration broadcast message.

The above algorithm essentially implements a distributed gradient on ρ_i, within the (l) iterations, for the minimization of the cost function of Eq. (14) by assuming that γ is a small positive constant. Lets denote the sequence generated within iterations as $\{\rho^{(1)}\}_{(\gamma)}$. Then, any limiting point of $\{\rho^{(1)}\}_{(\gamma)}$ is the global optimum for that problem with respect to this γ, by requiring a simple modification of the proof found of the original algorithm [1]. [6] In the following lemma we show that minimization of a sequence of these cost functions using increasing values for γ (chosen such that the solution of the next problem is close to the previous one; otherwise we risk getting stuck in steep valleys) transforms the "soft" constraints to "hard" ones, and it gradually leads to the global optimum of the original sub-problems of Eq. (12). The proof can be found in Section VIII-C.

Lemma 4.3. *Let $\{\rho^{(\mathbf{k})}\}$ be the sequence generated from the limiting points of $\{\rho^{(1)}\}_{(\gamma)}$ using increasing values for γ within the (k) iterations. Then, any limiting point of $\{\rho^{(\mathbf{k})}\}$ is the global optimum point ρ^* of the sub-problems of Algorithm 1.*

2) Two-level Optimization Algorithm: master problem running at the BS.: In this section we focus on the master problem of Algorithm 1, by assuming fixed user association policies and fixed BS load vector ρ. As discussed, this problem, also defined in Eq. (13), correspond to the access TDD allocation problem distributing the BS resources between UL and DL. Our objective is to provide distributed TDD allocation rules that should run at the BS level and push the control variable ζ to converge to its optimum value.

There are plenty of methods to update the access TDD allocation vector ζ at the next iteration $(m + 1)$ and thus the master problem. Following optimization theory [29], the general rule to improve the objective is through a *descent method*, as it follows in the next lemma.

Lemma 4.4. *If $\zeta_i^{(m)}$ is the current TDD allocation for BS i, at the next iteration it should be updated to*

$$\zeta_i^{(m+1)} = \zeta_i^{(m)} + t_i^{(m)} \Delta\zeta_i^{(m)}. \quad (18)$$

Then, $\zeta_i^{(m+1)}$ is a descent update of the master problem of Algorithm 1.

To calculate this update we need two parameters. Firstly, the *descent direction*, $\Delta\zeta_i^{(m)}$ that can be found from the first derivative criterion:

$$\Delta\zeta_i^{(m)} = \tau \left(1 - \frac{\rho_i^D}{\zeta_i^{(m)}}\right)^{-\alpha^D} \frac{\rho_i^D}{\zeta_i^{(m)}} + (1 - \tau) \left(1 - \frac{\rho_i^U}{1 - \zeta_i^{(m)}}\right)^{-\alpha^U} \frac{\rho_i^U}{1 - \zeta_i^{(m)}}. \quad (19)$$

Secondly, the *step size*. The backtracking method suggests that the this step size τ_i can be found by starting with $\tau_i = 1$ and repeat $\tau_i = \nu \cdot \tau_i$, until

$$\phi(\rho, \zeta_i \cdot t_i \cdot \Delta\zeta_i) < \phi(\rho, \zeta_i) + A \cdot t \cdot \nabla\phi(\rho, \zeta_i)^T \cdot \Delta\zeta_i, \quad (20)$$

where $A \in (0, 0.5), \nu \in (0, 1)$. [7]

[6] The descent direction at x improving the objective at the next iteration, through the corresponding inner product, is now provided from Eq. (16). This formula appropriately projects the direction *under* the cross interference constraint of Eq. (5) as shown in the proof of Theorem 4.2.

[7] Alternatively, there are other methods to calculate the descent direction and the step size. For instance, one could use the *Newton method* that provides the steepest descent direction in local Hessian norm, or the exact line search for the step size.

Note that, the TDD allocation update ζ_i of each BS i is performed in a distributed manner and independently from each other. The rules are scalable, the corresponding computational complexity is kept low since the first-order derivative criterion is only needed, and they offer flexible performance based on the α value. Finally, when stationarity is reached, we ensure that this is not a saddle point through a "stochastic" gradient: a noise vector with mean 0 is added to the gradient direction of stationary points that provably pushes them away from saddle points [33].

B. Three-level Optimization Algorithm.

We have successfully solved the joint problem of (i) user association and (ii) access resource allocation as shown in Algorithm 1, assuming that backhaul is always over-provisioned, e.g. assuming that $C_h \to \infty$. However, in real time implementations backhaul capacities are usually (quite) limited, leading to under-provisioned backhaul links. The latter emerges the need of including the backhaul constraints of Eq. (7), and thus also of using the third control variable Z namely (iii) backhaul resource allocation distributing the backhaul resources of a link between DL and UL. We remind the reader that, formally, considering these three dimensions together corresponds to Optimization Problem 1. In this section, we will provide an algorithm that efficiently tackles it, by building on Algorithm 1.

The backhaul constraints defined in Eq. (7) can be again tackled using penalty functions, by keeping the distributiveness of our framework. Thus, our new unconstrained cost function is similar in nature with Eq. (14), where now it should aggregate the penalties coming from both cross interference and backhaul constraints. As the analysis is an extension following similar logic to the previous section, we will not get into details. and we will immediately provide the algorithmic sketch in Algorithm 2.

Algorithm 2 (Complete) Three-level Optimization Algorithm that solves the user association and access and backhaul TDD allocation problem i.e. Optimization Problem 1.

1: **Repeat** until $\|Z^{(M)} - Z^{(M-1)}\| < \epsilon$.
2: *Update the master problem.*
3: BH resource allocation: $Z \to$ DL, $1 - Z \to$ UL.
4: *Update the secondary master and the two sub-problems.*
5: Run Algorithm 1.

We now show the convergence of this algorithm through the next lemma.

Lemma 4.5. *Let $\{(\rho^*, \zeta^*, Z)^{(\mathbf{M})}\}$ be the sequence generated within the (M) iterations by Algorithm 2, if at each iteration of a higher level master problems all the lower level problems have already converged. Then, any limiting point of $\{(\rho, \zeta, Z)^{(\mathbf{M})}\}$ is the global optimum of Optimization Problem 1.*

Regarding the master and secondary master problem update, the resource allocation parameters update (ζ, Z update) shall follow the same logic as described in section IV-A2. As for the two sub-problems (ρ optimization), the distributed algorithm and the optimal user association rules of section IV-A1 stay same in nature, i.e.

$$i^D(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{user knowledge}} \cdot \overbrace{P_i^D}^{\text{BS broadcast message}} \right) \quad (21)$$

but now the BS broadcast message part P_i^D should include all the penalties coming from both backhaul and cross interference constraints, i.e. $P_i^D = \frac{\zeta_i \cdot \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma \cdot \zeta_i \cdot \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D} \cdot \Pi_i^D}$, where Π_i^D is

$$\sum_{k \in \mathcal{B}_h(i)} \frac{\mathcal{J}^D(k) \tilde{c}_i^D}{Z(k) C_h(k)} \left(\frac{\sum_{l \in \mathcal{B}(k)} \rho_w^{(l),D} \tilde{c}_w^D}{Z(k) \cdot C_h(k)} - 1 \right) + \sum_{w \in \mathcal{C}_i} \mathcal{I}_{ij} (\rho_i^{(l),D} + \rho_w^{(l),U} - 1).$$

$\mathcal{J}^D(k)$ indicates whether the k backhaul link is congested in the DL ($\mathcal{J}^D(k) = 1$ when $\frac{\sum_{i \in \mathcal{B}(k)} \rho_i \tilde{c}_i}{Z(k) C_h(k)} > 1$). Note that, when the capacity constraints for the backhaul links k are not violated (e.g., $\mathcal{J}^D(k) = 0$), the above rules state that the optimal associations are the same as the one in Eq. (16). However, when a link becomes congested, an additional term is added that penalizes that BS making it less preferable to UEs at location x. Note that this penalty considers the whole backhaul path $\mathcal{B}_h(i)$ that traffic from BS i traverses, and adds a penalty for *every* link along that path that is congested (first sum of Π_i^D). Overall, our derived rules provide the optimal way to penalize the performance of a BS i depending on the total amount of congestion of all backhaul links this BS needs to traverse up to the core ($k \in \mathcal{B}_h(i)$), and amount of cross interference of other neighboring BSs ($w \in \mathcal{C}_i$).

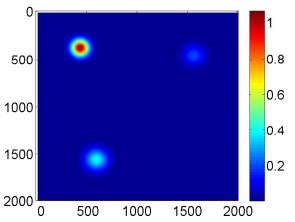
Finally, note that even in the under-provisioned backhaul scenario our association rules are scalable, since each BS i needs to only broadcast one value per dimension (i.e., P_i^D for downlink) no matter the number of cross-interfering BSs $i \in \mathcal{C}_i$ and the number of backhaul links it needs to traverse up to the core, they are of low-complexity and they offer flexible performance. Similarly for the TDD allocation updates.

V. Simulations

In this section, we evaluate our proposed algorithm on example scenarios, and discuss related insights. We first consider a simple scenario with one macro BS and three SCs, in order to better elucidate the qualitative behavior of our algorithm, compared to standard practices, as well as better trace its performance benefits and where these come from. We then consider a larger network scenario and demonstrate that similar benefits can be observed there as well. Note that our main focus is directly on Algorithm 2, referred as proposed algorithm hereafter, that considers the complete Optimization Problem 1. Specific elaborations on particular subproblems, constraints and tradeoffs will be stressed out where necessary.

Scenario 1: We consider a 2×2 km^2 area. Fig. 4 shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ (*flows/hour* per unit area) with 3 hotspots (blue implying low traffic and red high). We assume that this area is covered by three SCs (referred with BS numbers 1-3), and one macro cell (BS number 4). Without loss of generality, we assume

that each SC offloads its traffic through a dedicated backhaul link (corresponding BH link numbers 1-3) to the macro BS, and that the macro BS cross interferes with all SCs (i.e., $C_4 = \{1, 2, 3\}$, $C_1 = C_2 = C_3 = \{4\}$, see B.9). We consider standard parameters as adopted in 3GPP [34], listed in Table II[8]. We set $\alpha^D = \alpha^U = 1$ to optimize user throughput. (We have also considered other values, with similar conclusions.)



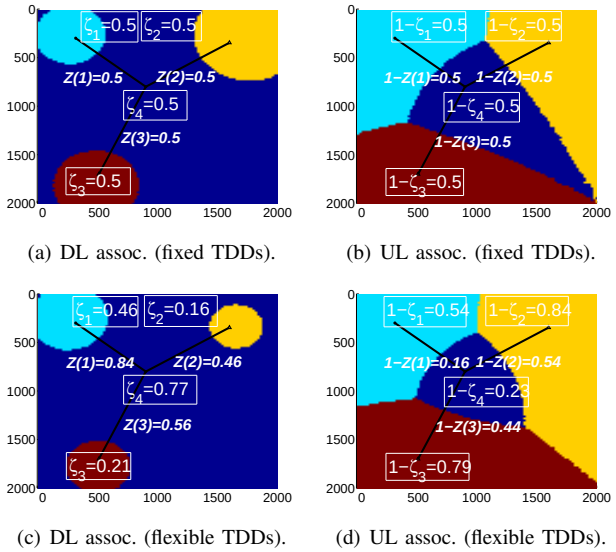
Variable	Value
$P_{eNB,SC,UE}$	43,24,12 dBm
$w_i,\ \forall i \in \mathcal{B}$	10 MHz
$C_h(j),\ \forall j \in \mathcal{B}_h$	100 MBps
N_0	-174 dBm/Hz
$1/\mu^{D,U}(x)$	100,25Kbytes
z^D, z^U	0.7,0.3
α^D, α^U	1,1

Fig. 4 & TABLE II: Traffic arrival rate and other simulation parameters.

Coverage Snapshots: We first look at the coverage maps that different schemes create. Figure 5(a), 5(b) depict the optimal user associations for fixed LTE-TDD configuration 1 that assumes static UL/DL timeslot ratio $4:4$ i.e., fixed $\zeta_i = 0.5, \forall i \in \mathcal{B}$. Similarly for the BH links $Z(j) = 0.5, \forall j \in \mathcal{B}_h$. As a first note, we see that in DL most users are associated with the macro BS, and a few to SCs (macro BS attracts more DL users due to the higher transmit power). In the UL, users tend to form Voronoi cells (to minimize path loss and improve UL SINR). Secondly, we note that the DL coverage areas of the various SCs are decreased according to the corresponding traffic arrival intensity: e.g. SC 1 that serves the most intense hotspot (see Fig.3) has the smallest coverage area, while SC 3 which sees lower traffic intensity has the largest). The main reason is that the SCs have limited DL backhaul capacities that force some users to the far away macro BS. This alleviates the backhaul link congestion but hurts overall performance. At the same time, a high amount of the pre-configured UL backhaul resources might remain wasted (due, to assymetry in DL/UL traffic intensity for example).

Summarizing, the observed coverage maps for this scenario demonstrate two possible shortcomings of pre-configured TDD: (a) asymmetry in the DL/UL coverage areas and corresponding transmit powers suggest that a TDD allocation other than 50-50% could improve performance; (b) some (usually DL) user associations could be suboptimal, dictated by backhaul capacity limitations arising from the preconfigured fixed allocation on the BH, even if the total BH resources would suffice for the sum of both UL and DL traffic.

To explore these possibilities, we now relax the allocation variables ζ and Z (see B.2 and C.2) and apply our proposed algorithm. Clearly, in this simple example, a single-step improvement in either direction described above ((a) or (b)) could improve performance. We remind the reader that our proposed algorithm goes beyond this single step, alternating between optimizing coverage maps and TDD resource allocation, until it finds the best possible combination. The resulting coverage maps (i.e. optimal ρ values) and radio/BH allocations (optimal ζ and Z values) are shown in Fig. 5(c), 5(d). We first note that macro BS increases its $\zeta_4 = 0.77$ to serve more DL users, and SC increase their UL resources $1-\zeta_1 = 0.54, 1-\zeta_2 = 0.84, 1-\zeta_3 = 0.79$ to serve more UL, bewaring to avoid *cross interference*. Interestingly, such an allocation simultaneously improves both UL and DL performances (we will explicitly show this later). Also, the DL BH allocated resources ($Z(j)$) are increased to accommodate more DL traffic, while ensuring not to exceed a maximum value that would congest the UL.



(a) DL assoc. (fixed TDDs). (b) UL assoc. (fixed TDDs).

(c) DL assoc. (flexible TDDs). (d) UL assoc. (flexible TDDs).

Fig. 5: DL and UL user associations for different scenarios ($\tau = 0.5$).

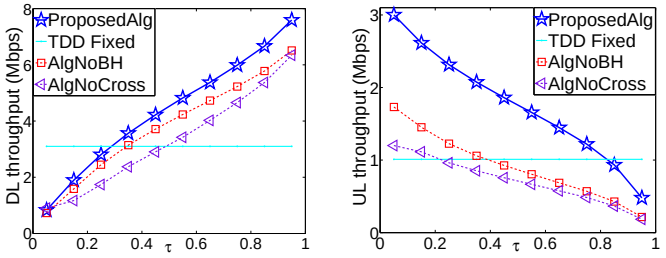
User-centric performance: We now go beyond the above qualitative behavior and evaluate the quantitative benefits. We first focus on user-centric performance and consider various τ values (we remind the reader that τ is a parameter that balances the importance of DL vs UL performance). We compare the performance of the following main schemes. (*ProposedAlg*): our proposed algorithm – see Algorithm 2; (*TDD Fixed*): the optimal allocation algorithm of [18] with equal, pre-confiigured UL/DL resources on both radio access and BH. To better understand the importance of considering the cross-interference and BH capacity constraints, we also include results for the following schemes. (*AlgNoCross*): jointly optimal allocation, but not taking cross-interference into account. If there is an eventual asymmetry in the optimal UL/DL schedules, potential cross-interference is included in the SINR to capture its impact. (*AlgNoBH*): jointly optimal allocation without considering the backhaul constraints. Here, we assume that all BSs associated with a BH link that is congested decrease their performance proportionally to the amount of congestion.

In Fig 6 we depict the DL and UL user throughput as a function of τ in different scenarios. It is easy to see that our *ProposedAlg significantly outperforms* the *TDD fixed* policy by up to $2.5-3\times$. What is more, for most intermediate τ values, it is able to simultaneously improve both DL and UL performance. As τ increases further, the emphasis of *ProposedAlg* moves exclusively to the DL (and vice versa) which is consistent with our expectations, unlike the fixed TDD scheme where DL and UL performances are optimized

[8]As for the sizes and ratios of different flows, as well as BH capacities, we can use different values in order to capture different simulation scenarios.

independently of τ (decoupled objective).

Regarding the impact of the cross interference constraint, *AlgNoCross* can still offer some improvement on the DL for $\tau > 0.5$, compared to the baseline (*TDD Fixed*). However, it does so with a significant penalty on UL performance (up to 3× worse), which is the most sensitive to cross-interference (this DL-to-UL interference is a key problem for future Flexible TDD [35]). This underlines the importance of directly considering cross interference constraints in our optimization framework through Eq.(5). Finally, the performance of *AlgNoBH* shows similar behavior, where it can sometimes provide better performance for the DL or the UL (compared to *TDD fixed*) but not both.



(a) DL throughput. (b) UL throughput.

Fig. 6: User-centric Performance.

One could notice that user throughput does not drops significantly when we neglect the cross interference constraint and we end up with cross-interfering BSs, as in the *(AlgNoCross)*, e.g. as $\tau \to 0$ UL throughput drops $3 \to 1.2$ Mbps (2.5 times). Or, when we neglect backhaul constraints and we end up with congested links, as in the *(AlgNoBH)*, e.g. as $\tau \to 1$ DL throughput drops $7.8 \to 6.2$ Mbps (1.3 times). This is due to the fact that: cross-interfering BSs and congested backhaul links do not affect the whole network, but specific groups of users associated with the cells that suffer from cross-interference or low backhaul capacity. To better illustrate this, in Table III we show the average throughput of such affected users, as a function of τ. Indeed, their performance is severely affected: e.g. as $\tau \to 0$ UL throughput now drops all the way to 0.3 Mbps (10 times) for *(AlgNoCross)*, or as $\tau \to 1$ DL throughput drops to 1.8 Mbps (4.5 times) for *(AlgNoBH)*.

TABLE III: Mean Throughput for negatively affected users (in Mbps)

Scenario.	$\tau \to 0$	$\tau \to 0.5$	$\tau \to 1$
DL and UL thr. for *(AlgNoBH)*	0.15 and 1.5	1 and 0.7	1.8 and 0.1
DL and UL thr. for *(AlgNoCross)*	0.1 and 0.3	3.2 and 0.15	6.1 and 0.05

Summarizing, the following important conclusions can be drawn from the above analysis: (a) jointly optimal allocation of user association and DL/UL radio resources can actually lead to considerable performance degradation, unless cross-interference is taken explicitly into account; (b) a jointly optimal allocation, even with cross-interference taken into account, might still be quite suboptimal, if the DL/UL resources on the BH are not also optimized to conform to the new load requirements imposed by the BSs; (c) joint optimization of all these dimensions is feasible, and can offer significant performance improvement for both DL/UL.

Network-centric performance. Table IV considers the performance improvements in the same comparison scenario (*ProposedAlg* and *TDD Fixed* [18]), but now from the network perspective when $\tau = 0.5$. We consider two metrics: Spectral Efficiency (SE) in terms of bits/s/Hz, and Load Balancing (LB) in terms of mean square error between different BS loads, similar to what is assumed in [18]. DL/UL spectral efficiency improve up to 44% since *flexible TDD better allocates the resources* with respect to the heterogeneous transmit powers that help physical data rates improve (see B.2-B.3). It also considers related traffic statistics and asymmetries across users (see A.1-A.2) by diminishing the BS load fluctuations (e.g., BS under/over utilizations) and thus LB is improved. It is interesting to note that simultaneous improvement of these metrics implies improvement in user performance, as showed previously and explained in B.7.

TABLE IV: Network (SE,LB) Performance ($\tau = 0.5$)

	Downlink		Uplink	
Performance.	SE	LB	SE	LB
Percentage % of improvement.	42	16	44	54

Scenario 2: Having highlighted the different tradeoffs and sources of performance improvement in the basic scenario above, we now turn our attention to a larger network topology consisting of 4 macro BSs and 13 SCs. Considerable performance improvements can be observed in this scenario as well if we compare our proposed algorithm with the *TDD Fixed*, as can be seen from Table V (e.g. 86% better UL user performance). Relative lower improvement values compared to the smaller Scenario 1 are mainly due to: (a) not all BSs experience bad performance now so even if *ProposedAlg* considerably improves the performance of the problematic BSs, average performance is not as affected; (b) the *additional cross interference constraints* posed from the neighboring clusters.

TABLE V: User (UE) and Network (SE, LB) Performance ($\tau = 0.5$)

	Downlink			Uplink		
Scenario.	UE	SE	LB	UE	SE	LB
Percentage % of improvement.	29	39	4	86	42	51

VI. Discussion and Future Work

Distributed algorithm. Our proposed framework proposes a novel convergent algorithm that solves the coupled problem of joint user association, access and backhaul TDD allocation, in a totally distributed manner. Decoupling it in three simpler optimization sub-problems, we showed how to solve them in three different network elements: user, base station, and backhaul links at potentially different timescales. There, each network element solves a certain problem at different rounds, by only requiring some simple message exchanges between rounds, facilitating a distributed implementation that globally converges without the need of any centralized entity. For example, the user is able to select where to associate based on own

measurements and BS broadcast information. Note that this is inline with user association in current LTE systems, where user association depends on device centric information (e.g. SINR measurements) but also BS-transmitted information (e.g. priority lists of BSs to monitor). In a similar manner, aligned with the eIMTA enabler, each BS distributes its resources between UL/DL by avoiding cross-interference based on the user demand in either direction and the previous allocation policy, or each link based on the corresponding demand by the BSs it offloads and the previous allocation policy.

Scalability, Complexity and Flexibility. Our derived user association rules are: scalable (constant amount of the BS broadcast messages irrespective of the number of users, backhaul topology, and cross-interference map), of low complexity (requiring a simple max operation) and offer flexible performance (defined from α values); see e.g. Eq. (21). Similarly, the rules for the access/backhaul resource allocation update satisfy similar characteristics by only requiring a first order derivative criterion; see e.g. Lemma 4.4.

Decomposition order. While our proposed decomposition is not the only possible decomposition, we believe this lends itself to a natural implementation between different network elements. As discussed, the hierarchical decomposition can be done by a number of different decomposition orders and all would converge to the global optimum under the mentioned certain circumstances. Specifically, upon n optimization problems there are $n!$ (factorial of n) possible decomposition orders; for us this would be $3! = 2 * 3 = 6$. However, our proposed decomposition order captures and comforts to the network dynamics in reality. User association is proposed to run in the fastest timescale to adapt to the high traffic fluctuations across different locations and users. The load of a single BS depends on the sum of its attached users and is subject to fewer fluctuations. It only has to react to (slower) traffic shifts of the aggregate loads, by updating its ζ parameter accordingly. Finally, a backhaul link further aggregates the traffic of multiple BS, and can update its optimal allocation Z at an even slower timescale.

Cross-layer and cross-network optimization: In this framework we perform cross-layer optimization since we jointly optimize different functionalities coming from different layers: e.g. the user association problem (coming from the network layer), the TDD allocation problem (coming from the MAC layer), as well as the cross-interference management (coming from the PHY layer). Also, we perform cross-network optimization since we jointly consider and optimize different functionalities and characteristics of both radio access and backhaul networks. In our future work plans we include the consideration of fronthaul network too, as explained in the next paragraph.

Fronthaul Network and C-RAN. In the proposed framework we have successfully considered the backhaul network and the constraints related to it along with the radio access. However, modern networks tend to increasingly focus on Centralized-Radio Access Network (C-RAN) architectures, fact that has lead fronthaul networks to be rather under-provisioned and their architecture to be revisited. Thus, the introduction of the fronthaul in our framework, along with the potentially influenced by, backhaul network, and their interaction with radio access is another promising extension.

Joint UL/ DL association: Our framework is also applicable when DL and UL traffic at a location x have to be offloaded to the same BSs (see B.8), by requiring $p_i^D(x) = p_i^U(x)$ in the association rule derivations. We defer to future work other similar splits, e.g., for control/data channels, or best effort/dedicated traffic [25].

VII. Conclusion

In this paper, we formulated a novel algorithm that carefully studies the coupled problems of (i) user association, TDD (ii) access, and (iii) backhaul resource allocation under the emerging *backhaul* and *cross interference* constraints. Using optimization theory we proved that under certain circumstances it converges to the global optimum. Simulation results corroborate the correctness of our framework and reveal promising qualitative and quantitative results.

VIII. Proofs Of Theoretical Results

A. Proof of Lemma 4.1

Our proposed decomposition algorithm falls into the category of Alternate Convex Search (ACS) [36], [30], that is a special case of the popular Block Coordinate Decent (BCD) or Gauss-Seidel method [37]. Thus, our proposed algorithm provably converges to a stationary point that could be either a (local or global) optimal, or even a saddle point. There, starting from an initial feasible point, one attempts to minimize the objective by cyclically iterating through the different optimization directions with respect to one coordinate direction at a time. Precisely, in our case at the end of the (m) iteration it is

$$\phi_\alpha(\rho, \zeta^{(m)}) \prec \phi_\alpha(\rho, \zeta^{(m-1)}).$$

This will continue until convergence to a stationary point, where the gradient vanishes and the above inequality approaches equality. ACS algorithms in its simplest form suggest that the stationary point could be a saddle point, a local or global optimal [36]. However, Algorithm 1 guarantees convergence to the global optimum due to the following two points.

(1) *Uniqueness of optimum point:* The considered problem of Eq. (11) can be converted to a geometric programming (GP) problem, since both its objective and constraints can be written as a sum of posynomials terms composed of positive monomials, according to the transformation in [38]. Such problems have a single optimum. (The GP equivalent form of our problem is not convenient for decomposition, so we use this argument only to prove uniqueness, but not to solve the joint problem.)

(2) *Saddle point escape:* Our proposed algorithm can escape from potential saddle points, as discussed in Section IV-A2, with the use of stochastic gradient.

B. Proof of Theorem 4.2

To write the proof compactly with respect to the coupling constraints, we denote (only within the proof) $\zeta^D = \zeta, \zeta^U = 1 - \zeta$, $I(D) = \mathcal{I}_{ij}, I(U) = \mathcal{I}_{ji}$ and assume that L is either D or U ($L \in \{D, U\}$) with complementary value $\tilde{L}$.

For $\rho \in \mathcal{F}$, we will show that the rules derived in Theorem 4.2 update the association variables and eventually the BS load vector at the (l) iteration to $\rho^{(l)}$ in a descent direction. A sufficient condition for this is to ensure if $\langle \nabla\Phi(\rho^{(l)}, \zeta, \gamma), \Delta\rho^{(l)} \rangle \geq 0$ for all $\rho \in \mathcal{F}$, where $\Delta\rho^{(l)} = \rho - \rho^{(l)}$. In addition, these rules maximize this inner product with gradient at ρ, ensuring the steepest descent direction. Let $p(x)$ and $p^{(l)}(x)$ be the associated routing probability vectors for ρ and $\rho^{(l)}$, respectively.

Using the deterministic DL and UL cell coverage generated by (16) the respective optimal association variables at the (l) iteration, denoted as $p_i^{(l),L}(x)$, are $p_i^{(l),L}(x) = \mathbf{1}\left\{ i = i^L(x) \right\}$. Then, the inner product $\langle \nabla\Phi\left(\rho^{(l)}, \zeta, \gamma\right), \Delta\rho^{(l)} \rangle$ is equal to

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1}{\zeta_i^L \left(1 - \frac{\rho_i^{(l),L}}{\zeta_i^L}\right)^{\alpha^L}} + 2\gamma \sum_{j\in\mathcal{C}_i} I(L)(\rho_i^{(l),L} + \rho_j^{(l),\tilde{L}} - 1) \right) (\rho_i^L - \rho_i^{(l),L}) =$$

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma \cdot \zeta_i^L \left(1 - \frac{\rho_i^{(l),L}}{\zeta_i^L}\right)^{\alpha^L} \sum_{j\in\mathcal{C}_i} I(L)(\rho_i^{(l),L} + \rho_j^{(l),\tilde{L}} - 1)}{\zeta_i^L \left(1 - \frac{\rho_i^{(l),L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \int_{\mathcal{L}} \rho_i^L(x) \left(p_i^L(x) - p_i^{(l),L}(x)\right) dx =$$

$$\sum_L \int_{\mathcal{L}} \frac{\lambda^L(x)}{\mu^L(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma\zeta_i^L \left(1 - \frac{\rho_i^{(l),L}}{\zeta_i^L}\right)^{\alpha^L} \sum_{j\in\mathcal{C}_i} I(L)(\rho_i^{(l),L} + \rho_j^{(l),\tilde{L}} - 1)}{\zeta_i^L c_i^L(x) \left(1 - \frac{\rho_i^{(l),L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \left(p_i^L(x) - p_i^{(l),L}(x)\right) dx.$$

Note that in the DL i.e. $L = D$ (similarly in UL)

$$\sum_{i\in\mathcal{B}} p_i^D(x) \left(\frac{1 + 2\gamma \cdot \zeta_i^L \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{(l),D} + \rho_j^{(l),U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D}} \right) \geq$$

$$\sum_{i\in\mathcal{B}} p_i^{(l),D}(x) \left(\frac{1 + 2\gamma \cdot \zeta_i^L \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{(l),D} + \rho_j^{(l),U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D}} \right)$$

holds because $p_i^{(l),D}(x)$ is an indicator for the minimizer of $\frac{1+2\gamma\zeta_i^L \left(1-\frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^L} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{(l),D}+\rho_j^{(l),U}-1)}{\zeta_i c_i^L(x)\left(1-\frac{\rho_i^{(l),D}}{\zeta_i}\right)^{\alpha^D}}$.

So, $\langle \nabla\Phi(\rho^{(l)}, \zeta, \gamma), \Delta\rho^{(l)} \rangle \geq 0$ holds.

C. Proof of Lemma 4.3

Let within this proof $\rho^{(k)}$ be the optimum of the problem of minimizing the cost function $\Phi_\alpha(\rho, \zeta, \gamma^{(k)})$ of Eq. (14) using the penalty factor $\gamma^{(k)}$ at the (k) iteration, i.e. of the problem $\min_\rho \Phi_\alpha(\rho, \zeta, \gamma^{(k)})$. Then, let us denote as $\rho^{(\mathbf{k})} = \{\rho^{(0)}, \rho^{(1)}, ..., \rho^{(k)}, ...\}$ the sequence of optimums of the cost function $\Phi_\alpha(\rho, \zeta, \gamma^{(k)})$ using increasing values of $\gamma^{(k)}$: $\gamma^{(0)} < \gamma^{(1)} < \gamma^{(2)} < ... < \gamma^{(k)} < ...$, at each iteration (k). We will show that any limiting point $\tilde{\rho}$ of $\rho^{(\mathbf{k})}$ is the global optimal point of the set of the original sub-problems defined in Eq. (12).

Given the cross-interference constraints defined in Eq. (5), let us define as $P(\rho) = \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{C}} (\rho_i^D + \rho_j^U - 1)^2$. It is

$$\Phi_\alpha(\rho, \zeta, \gamma) = \phi_\alpha(\rho, \zeta) + \gamma \cdot P(\rho). \tag{22}$$

From the continuity of ϕ_α we have

$$\lim_{k\to\infty} \phi_\alpha(\rho^{(k)}, \zeta) = \phi_\alpha(\tilde{\rho}, \zeta). \tag{23}$$

Let ρ^* be the solution of the problem of Eq. (12), and f^* be the value of the considered cost function at this point. It can be shown that the sequence of values of $\Phi_\alpha(\rho^{(k)}, \zeta, \gamma^{(k)})$ are non-decreasing:

$$\begin{aligned} \Phi_\alpha(\rho^{(k+1)}, \zeta, \gamma^{(k+1)}) &= \phi_\alpha(\rho^{(k+1)}, \zeta) + \gamma^{(k+1)} \cdot P(\rho^{(k+1)}) \geq \\ &\geq \phi_\alpha(\rho^{(k+1)}, \zeta) + \gamma^{(k)} \cdot P(\rho^{(k+1)}) \geq \\ &\geq \phi_\alpha(\rho^{(k)}, \zeta) + \gamma^{(k)} \cdot P(\rho^{(k)}) = \\ &= \Phi_\alpha(\rho^{(k)}, \zeta, \gamma^{(k)}), \end{aligned}$$

and bounded above by f^* for each k:

$$\begin{aligned} f^* = \phi_\alpha(\rho^*, \zeta) + \gamma^{(k)} \cdot P(\rho^*) \geq \phi_\alpha(\rho^{(k)}, \zeta) + \gamma^{(k)} \cdot P(\rho^{(k)}) \geq \\ \geq \phi_\alpha(\rho^{(k)}, \zeta). \end{aligned} \tag{24}$$

Thus, from the above-mentioned equations we can easily imply that the following limit is a real number:

$$\lim_{k\to\infty} \Phi_\alpha(\rho^{(k)}, \zeta, \gamma^{(k)}) = q^* \leq f^*. \tag{25}$$

Subtracting (25) from (23) yields

$$\lim_{k\to\infty} \gamma^{(k)} P(\rho^{(k)}) = q^* - \phi_\alpha(\tilde{\rho}, \zeta). \tag{26}$$

Since $P(\rho^{(k)}) \geq 0$ and $\gamma^{(k)} \to \infty$, Eq. (26) implies that surely

$$\lim_{k\to\infty} P(\rho^{(k)}) = 0. \tag{27}$$

Using the continuity of P, this implies that $P(\tilde{\rho}) = 0$ i.e. the constraints have converged to "hard" ones and thus $\tilde{\rho}$ is feasible. To show that $\tilde{\rho}$ is optimal we note from Eq. (24) that also $\phi_\alpha(\rho^{(k)}, \zeta) \leq f^*$, and hence

$$\phi_\alpha(\tilde{\rho}, \zeta) = \lim_{k\to\infty} \phi_\alpha(\rho^{(k)}, \zeta) \leq f^*. \tag{28}$$

REFERENCES

[1] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, 2012.

[2] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[3] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An analytical framework for optimal downlink-uplink user association in hetnets with traffic differentiation," in *Proc. IEEE Globecom*, 2015.

[4] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Globecom*, 2014.

[5] V. Pauli and E. Seide, *Dynamic TDD for LTE-A and 5G*, 2015.

[6] G. 36.133, "Evolved universal terrestrial radio access (E-UTRA) and radio access network (E-UTRAN); overall description," 2012.

[7] R. Sivaraj, I. Broustis, N. K. Shankaranarayanan, V. Aggarwal, R. Jana, and P. Mohapatra, "A QoS-enabled holistic optimization framework for LTE-advanced heterogeneous networks," in *Proc. IEEE Infocom*, 2015.

[8] M. Ding, D. L. Perez, A. V. Vasilakos, and W. Chen, "Dynamic TDD transmissions in homogeneous small cell networks," in *Proc. IEEE ICC Communications Workshops*, 2014.

[9] H. Ji, Y. Kim, S. Choi, J. Cho, and J. Lee, "Dynamic resource adaptation in beyond LTE-A TDD heterogeneous networks," in *Proc. IEEE ICC Communications Workshops*, 2013.

[10] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic uplink-downlink optimization in TDD-based small cell networks," in *International Symposium on Wireless Communications Systems*, 2014.

[11] 3GPP, "TS 36.300, Release 13 (version 13.2.0)," 2016.

[12] *Backhaul technologies for small cells*, Small Cell Forum, 2014.

[13] D. Chen, T. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, 2015.

[14] J. Ghimire and C. Rosenberg, "Revisiting scheduling in heterogeneous networks when the backhaul is limited," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2015.

[15] A. D. Domenico, V. Savin, and D. Ktenas, "A backhaul-aware cell selection algorithm for heterogeneous cellular networks," in *Proc. of IEEE PIMRC*, 2013.

[16] Z. Cui and R. Adve, "Joint user association and resource allocation in small cell networks with backhaul constraints," in *Proc. IEEE CISS*, 2014.

[17] M. Shariat, E. Pateromichelakis, A. Quddus, and R. Tafazolli, "Joint TDD backhaul and access optimization in dense small cell networks," *IEEE Transactions on Vehicular Technology*, 2013.

[18] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal downlink and uplink user association in Backhaul-limited HetNets," in *Proc. IEEE Infocom*, 2016.

[19] A. D. Domenico, V. Savin, and D. Ktenas, "A backhaul-aware cell selection algorithm for heterogeneous cellular networks," in *in Proc. IEEE PIMRC*, 2013.

[20] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. of IEEE Infocom*, 2006.

[21] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. of IEEE Infocom*, 2003.

[22] 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 Journal on Selected Areas in Comm.*, 2011.

[23] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Imperial college press, 2010.

[24] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. Mobile Computing and Networking (MobiCom)*, 2003.

[25] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communication Surveys and Tutorials*, 2013.

[26] 3GPP, "TR 36.842, Release 12 (version 12.0.0)," 2014.

[27] E. Metsala and J. Salmelin, *Mobile Backhaul*. Wiley, 2012.

[28] Cambridge Broadband Networks, "Solutions mobile backhaul," 2015.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[30] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, 2007.

[31] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, 2006.

[32] Z. G. Raphael T. Haftka, *Elements of Structural Optimization*. Springer Netherlands, 1992.

[33] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points-online stochastic gradient for tensor decomposition," in *Proc. of the 28th Conference on Learning Theory*, 2015.

[34] 3GPP, "TR 36.931 Release 13 (version 13.2.0)," 2016.

[35] Z. Shen, A. Khoryaev, E. Eriksson, and X. Pan, "Dynamic uplink-downlink configuration and interference management in TD-LTE," *IEEE Communications Magazine*, 2012.

[36] R. E. Wendell and A. P. Hurter, "Minimization of a non-separable objective function subject to disjoint constraints," *Journal on Operation Research*, 1976.

[37] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.

[38] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*. Springer-Verlag New York, Inc., 2000.