Joint Optimization of User Association and Dynamic TDD for Ultra-Dense Networks

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

[1] EURECOM, Sophia Antipolis, France, firstname.lastname@eurecom.fr

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

***Abstract*—Ultra-dense small cell networks will require sophisticated user association algorithms that consider (i) channel characteristics, (ii) base station load, and (iii) uplink/downlink (UL/DL) traffic profiles. They will also be characterized by high spatio-temporal variability in UL/DL traffic demand, due to the fewer users per BS. In this direction, *Dynamic TDD* is a promising new technique to match BS resources to actual demand. While plenty of literature exists on the problem of user association, and some recent on dynamic TDD, most works consider these separately. In this paper, we argue that user association policies are strongly coupled with the allocation of resources between UL and DL. We propose an algorithm that decomposes the problem into separate subproblems that can each be solved efficiently and in a distributed manner, and prove convergence to the global optimum. Simulation results suggest that our approach can improve UL and DL performance *at the same time*, with an aggregate improvement of more than $2\times$, compared to user association under static TDD allocation.**

I. Introduction

The trend towards base station (BS) densification will continue in 5G systems, towards *Ultra-Dense Networks (UDN)* where: (i) many different small cells (SCs) are in range of most users; (ii) a small number of users will be active at each SC [1], [2]. The resulting traffic variability makes optimal user association a challenging problem in UDNs [1].

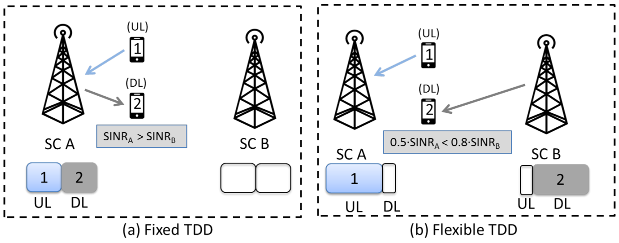
While optimization of current networks revolves around the downlink (DL) performance, social networks, Machine Type Communication (MTC), and other upload-intensive applications make uplink (UL) performance just as important. Some SCs might see their UL resources congested, while others their DL resources, depending on the type of user(s) associated with that SC. What is more, the same SC might experience higher UL or DL traffic demand over time. *Even sophisticated association policies might suffer under UL/DL traffic assymetry, if BS resources are not appropriately dimensioned.*

Conventional networks usually operate with the same amount of resources for UL and DL (FDD or static TDD) [3]. However, recently proposed *Dynamic or Flexible TDD* systems can better accommodate UL/DL traffic asymmetry, by varying the percentage of LTE subframes used for DL and UL transmission [4]. Hence, more UL resources can be allocated to SCs with UL intensive users, and vice versa.

Nevertheless, dynamic TDD introduces new challenges. First, *there is a strong interplay between user association and dynamic TDD policies*. Consider the simple example of Fig. 1, where a UL intensive user (1) and a DL intensive user (2) are both in range of a SC A (which is close) and a SC B (which is further away). Assume that each SC has initially the same amount of DL and UL resouces (50%). Both the DL and the UL user will connect to A, as it offers the best SINR. Now, notice that any change in the TDD schedule of SC A will hurt one of the two users. However, assume that A increases its UL resources to 80%. There are now $8/5$ (i.e. 60%) more resources for the UL user, which could lead to 60% higher rate. Furthermore, assume that SC B increases its DL resources to 80%. Connecting the DL user to B can increase the available resource blocks for her also by a factor of $8/5$. If the resulting SINR decrease has a smaller impact than this factor, then *both users can win by revisiting both the TDD schedules and the association decisions of BSs A and B*.

The previous example, while oversimplified, helps illustrate some of the dependencies at hand. One important omission in the above example, is that we ignored the DL-to-UL cross interference that might arise if nearby BSs have different schedules (see Fig. 1). E.g. a macro-cell transmitting on the DL, can really hurt a nearby SC transmitting on the UL [5]. Hence, excessive liberty in tuning UL and DL resources might hurt rather than help. Hence, the main goal in this paper is *to propose a generic framework for the joint α-fair optimization of user association and TDD allocation per BS*. Precisely:

- Associate users with BSs to optimize a chosen user- or network- centric performance metric (e.g. spectral efficiency, load-balancing, etc.).
- Choose the TDD UL/DL configuration for each SC to best match the UL/DL traffic demand for that metric.
- Consider the TDD UL/DL configuration of nearby SCs to avoid cross-interference.



(a) Fixed TDD

(b) Flexible TDD

Fig. 1: Interplay between user association and TDD configuration.

A. Related Work

User Association in HetNets and UDNs: Associating with the BS with the strongest SINR is common, but ignoring

This work was supported from the European Framework Program under H2020 grant agreement 671639 for the COHERENT project, as well as under the ICT theme of DG-CONNECT 317941 for the iJOIN project.

the BS load may fail in dense, heterogeneous networks. Some load-based modifications are considered in LTE SON (Self-Organized Networks), but are not widely deployed. Cell Range Expansion (CRE) techniques can somewhat improve the situation [6], but are more applicable in assymetric macro-cell/small-cell setups. As a result, a number of load-based association algorithms have been proposed to address the problem of optimal user association [7], [8], [9], [10]. In that context, in our previous work [11], [12], we tried to shed some light on the impact of traffic differentiation and backhaul limitations on user association. *Dual connectivity* aspects (e.g. splitting the UL and DL of the same user between different BSs) and *multi-connectivity* (where a user might download from two or more BSs in parallel for its DL) have also been considered recently [13]. We will explicitly consider the former in our framework, and defer the latter to future work.

Targetting UDN user association specifically, [14] considers the tradeoff between "spreading" the users among many BSs (reducing the load, and improving performance) vs. "concentrating" users in fewer BSs (to turn off empty BSs, saving energy or avoiding extra interference). Finally, a recent work attempts to jointly optimize user association and power control in UDNs [15]. Nevertheless, *all aforementioned works on user association either focus on DL traffic only [14], [7], assume fixed UL/DL resources at each base station [15], [11], [8], or focus on a single metric (e.g. spectral efficiency) [9].*

Dynamic TDD: Seven TDD uplink-downlink configurations are already supported in TD-LTE, but real TD-LTE systems are typically deployed with all cells using the same UL/DL configuration [5]. However, dynamic TDD is needed to adapt the configuration per cell, in dense networks. To deal with potential cross-interference, Almost Blank Subframes (ABS) (part of the enhanced Inter-Cell Interference Coordination (eICIC) specifications [16], [10]) could be transmitted for example in the DL direction for BS A, if a nearby BS B is currently in the UL. Yet, excessive use of ABS could waste a lot of capacity. "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA), standardized in LTE-A Release 13 [4], proposes a number of new techniques to deal with cross-interference [5] such as (a) different power control mechanisms, or (b) different pilot signal quality mechanisms for *flexible* subframes, and (c) *clustering* where a (small) number of nearby BSs must be synchronized, but different TDD configurations could be used per cluster. While useful, these techniques can only partially combat the problem.

In the modeling arena, [17] uses stochastic geometry to understand the amount of cross-interference in random network deployements. [18] also considers some simple heuristics for distributed TDD configuration. [19] models a random small cell network with stochastic traffic arrivals (rather than saturated users). Finally, [20] studies the impact of the time-scale of TDD adaptation to traffic variability, considering adaptation intervals from 10ms (the minimum) to 640ms. Despite progress in this area, *prior works assume users associations are given, based on some simple (e.g. SINR-based) association algorithm.*

B. Contributions

Summarizing, to our best knowledge this is the first work to address the problem of joint user association and dynamic TDD in UDNs. Our contributions are as follows:

TABLE I: Notation

	Downlink	Uplink
Key control variables		
Association probability of location x with BS i	$p_i^D(x)$	$p_i^U(x)$
TDD percentage of resources for BS i	ζ_i	$1-\zeta_i$
Auxiliary control variables		
Normalized BS load ($\zeta_i \to 1$ for DL), ($\zeta_i \to 0$ for UL)	ρ_i^D	ρ_i^U
Input variables		
Traffic arrival rate (flows/sec) at location x	$\lambda^D(x)$	$\lambda^U(x)$
Mean flow size at location x	$\overline{s}^D(x)$	$\overline{s}^U(x)$
Max. rate of BS i at location x	$c_i^D(x)$	$c_i^U(x)$
Load estimate of BS i (used for broadcast)	$\hat{\rho}_i^D$	$\hat{\rho}_i^U$
Load density of BS i at location x	$\rho_i^D(x)$	$\rho_i^U(x)$
$\alpha-$fair function parameter (for the objective)	α^D	α^U
α-fair total cost function	$\phi_\alpha^D(\cdot)$	$\phi_\alpha^U(\cdot)$
Cross interf.: neighbor set and penalty indicator (A.9)	C_i , $\mathcal{I}_{ij}$	

(1) We propose an analytical framework to study the joint optimization problem (Section II and Section III).

(2) We show that the joint problem is non-convex in general, and propose a primal decomposition algorithm that reduces complexity and can be implemented in a distributed manner. We then prove that this algorithm converges to the optimal solution of the joint problem (Section IV).

(3) Using simulations, we show our approach can *concurrently* improve UL and DL performance compared to the state of the art, showing more than 2× aggregate improvement, in the scenarios considered (Section V).

II. System Model and Assumptions

To keep notation consistent, the superscript "D" and "U" refer to downlink and uplink traffic, respectively, for all variables considered. For brevity, we'll sometimes omit explicitly defining the respective UL notation, if it is similar to the DL. In Table I, we summarize some key notation.

(A.1 - Network topology) We assume an area $\mathcal{L} \subset \mathbb{R}^2$ served by a set of base stations $\mathcal{B}$.

(A.2 - Traffic model) Downlink traffic at location $x \in \mathcal{L}$ consists of DL and UL file (or "flow") requests arriving as an inhomogeneous Poisson point processes with arrival rate per unit area $\lambda^D(x)$.[1] Similarly for the UL with rate $\lambda^U(x)$. Flow sizes are drawn from a *generic* distribution with means $\overline{s}^D(x)$ and $\overline{s}^U(x)$, respectively.

(A.3 - Resource allocation between UL and DL) Each BS $i \in \mathcal{B}$ has a total bandwidth w_i. A percentage $\zeta_i \cdot w_i$ is used to serve DL flows and $(1-\zeta_i) \cdot w_i$ for UL flows, where $\zeta_i \in (0,1)$ are *control variables* for our problem.[2]

(A.4 - Association variables) $p_i^D(x) \in [0,1]$ is the probability that DL (respectively UL) flows generated from users at location x get served by BS i. Respectively for $p_i^U(x)$. These are also control variables in our problem. Note that *per flow association* rather than per user, applies well to traffic steering

[1] If flow arrivals at each location are not Poisson, using the Palm-Khintchine theorem [21] we can show that the Poisson assumption is a good approximation for the aggregate input traffic to a BS, which comes from many locations.

[2] The model applies more generally, beyond TDD subframes, namely to allocate carriers, antennas, etc., between UL and DL.

in 5G New Radio (NR).[3]

(A.5 - Physical data rate) BS $i \in \mathcal{B}$ has 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 (i.e. if *all* DL resources are allocated to that user) given by the Shannon capacity:

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

where $\text{SINR}_i(x)$ is the SINR at location x when receiving from BS i.[4]

(A.6 - Offered load at x**)** The contribution of traffic arising at location x to the total DL load of a BS i, when location x is associated with BS i, is equal to $\rho_i^D(x,\zeta_i) \cdot dx$, where

$$\rho_i^D(x,\zeta_i) = \frac{\lambda^D(x) \cdot E[s^D](x)}{c_i^D(x,\zeta_i)}, \tag{2}$$

defines a load density.

(A.7 - BS utilization/load) $\rho_i^D(\zeta_i)$, defined as

$$\rho_i^D(\zeta_i) = \int_{\mathcal{L}} p_i^D(x)\rho_i^D(x,\zeta_i)dx, \tag{3}$$

represents the percentage of time the DL resources of BS i are busy [21], [24]. For convenience we also define the *normalized* load variables

$$\rho_i^D = \rho_i^D(\zeta_i = 1) \quad \rho_i^U = \rho_i^U(\zeta_i = 0). \tag{4}$$

Note that these are independent of ζ_i, and only depend on the association variables $p_i(x)$ (see also Eq.(1)-(3)). We will use them as *auxiliary control variables* in the optimization.

(A.8 - Scheduling assumptions) We assume each BS operates two separate queueing systems, one for DL and one for UL flows, which can be modelled each as an M/G/1 multi-class processor sharing (PS) system (similarly to [24], [7]). For such a system, it is easy to show that the expected number of active DL flows in BS i is given by $E[N_i] = \frac{\rho_i^D/\zeta_i}{1-\rho_i^D/\zeta_i}$ [21]. Furthermore, the expected troughput per flow depends both on the user's physical data rate $c_i^D(x,\zeta_i)$ (related to users at location x only) and the total BS load $\rho_i^D(\zeta_i)$ (related to *all* users associated with BS i) and is given by

$$c_i^D(x,\zeta_i) \cdot (1 - \rho_i^D/\zeta_i). \tag{5}$$

Hence, the throuhgput of that user can improve by: (i) associating to a BS with higher SINR, or (ii) reducing ρ_i^D in the serving BS i by associating fewer locations to i, or (iii) increasing the amount of DL resources ζ_i.

(A.9 - UL/DL cross interference avoidance) Without loss of generality, we assume that each BS i cross interferes with a subset of other BSs $\mathcal{C}_i \subseteq \mathcal{B} \setminus \{i\}$. If i is *active* on the DL and a BS $j \in \mathcal{C}_i$ *active* on the UL (or vice versa) at the same time, then these BSs might cause severe interference to each other. We introduce the constraints to prohibit cross interference:

$$\rho_i^D + \rho_j^U \leq q, \forall i \in \mathcal{B}, j \in \mathcal{C}_i, \tag{6}$$

where parameter $0 < q \leq 1$ controls the amount of cross-interference allowed on average, as follows. Consider two nearby BSs i and j. If $\zeta_i = \zeta_j$ then each BS has the same amount of DL and UL slots, and they could synchronize their slots (e.g. over the LTE X2 interface). If $\zeta_i \neq \zeta_j$, cross-interference might occur, but *it also depends on the utilization of each BS*. Observe that the sum of expected busy DL slots of i and UL slots of j is given by:

$$\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i + \frac{\rho_j^U}{1-\zeta_j} \cdot (1-\zeta_j).$$

Out of the ζ_i slots used for DL, only $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i = \rho_i$ will be busy on average. The rest could be blanked with ABS frames (or other interference management techniques) *without any capacity loss* (see Fig. 2). Similarly, for the percentage of slots that j will be *active* on the UL.Consequently, if $\rho_i^D + \rho_j^U \leq 1$, there are enough different slots on average in a frame to fit both i's DL and j's UL. However, choosing $\rho_i^D + \rho_j^U \leq q$, for $q \leq 1$ further reduces the probability of cross-interference.

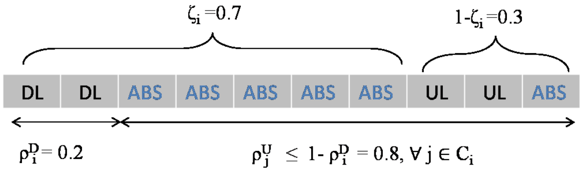


Fig. 2: Subframe allocation to avoid cross-interference.

III. Optimization Problem

Based on the discussed problem setup, we now formulate the joint optimizaton problem at hand.

Problem 1 (Joint Problem)**.** *The jointly optimal user association and dynamic TDD resource allocation is the solution of the following problem:*

$$\min_{\mathbf{p}^D,\mathbf{p}^U,\zeta}. \ \tau \cdot \textstyle\sum_{i \in \mathcal{B}} \frac{\left(1-\frac{\rho_i^D}{\zeta_i}\right)^{1-\alpha^D}}{\alpha^D - 1} + (1-\tau) \cdot \textstyle\sum_{i \in \mathcal{B}} \frac{\left(1-\frac{\rho_i^U}{1-\zeta_i}\right)^{1-\alpha^U}}{\alpha^U - 1} \tag{7}$$

$$\rho_i^y = \textstyle\int_{\mathcal{L}} p_i^y(x)\rho_i^y(x)dx, \ \ y \in \{U,D\} \tag{8}$$

$$\textstyle\sum_{i \in \mathcal{B}} p_i^D(x) = 1, \ \ \textstyle\sum_{i \in \mathcal{B}} p_i^U(x) = 1 \tag{9}$$

$$0 \leq p_i^y(x) \leq 1, \ \ \forall x \in \mathcal{L}, \ \ y \in \{U,D\}, \tag{10}$$

$$\zeta_{min} \leq \zeta_i \leq \zeta_{max}, \tag{11}$$

$$0 \leq \rho_i^D \leq (1-\epsilon) \cdot \zeta_i, \ \ \forall i \in \mathcal{B}, \tag{12}$$

$$0 \leq \rho_i^U \leq (1-\epsilon) \cdot (1-\zeta_i), \ \ \forall i \in \mathcal{B}, \tag{13}$$

$$\rho_i^D + \rho_j^U \leq q, \ \ \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \tag{14}$$

In case of $\alpha^D = 1$ *(or* $\alpha^U = 1$*) the respective term inside the objective sum is replaced with* $\log(1 - \rho_i^D/\zeta_i)^{-1}$.

Control Variables: There are two main sets of *control variables*: (i) the user association variables in each direction (A.4), namely $p_i^D(x)$ and $p_i^U(x)$; we will often use the vector

[3] W.l.o.g. we assume UEs have *dual-connectivity* capabilities and can associate to a different BS for UL traffic, i.e. $p_i^D(x) \neq p_i^u(x)$, as proposed in LTE Rel. 12 [22]. However, our framework is backward compatible when joint UL/DL association is required (see Eq.(15)).

[4] We assume $\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j \neq i} G_j(x)P_j + N_0}$, where 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.). Effects of fast fading are filtered out, and we assume the total intercell interference at location x is treated as another noise source, as in most aforementioned works [7], [23], [11]. Symmetrically for the UL case.

shorthand $\mathbf{p}^D = \{p_1^D(x), p_2^D(x), \ldots, p_{|\mathcal{B}|}^D(x)\}$ for all locations $x \in \mathcal{L}$, and $\mathbf{p}^U$. And, (ii) the dynamic TDD parameters ζ_i that split the resources of BS i between DL and UL (A.3).

Objective: Consider the left sum, which corresponds to DL performance. It has one term for each BS i corresponding to the popular α-fair function with $(1 - \rho_i^D/\zeta_i)$ as its argument [25]. This family of objectives can capture different tradeoffs between spectral efficiency and load-balancing by varying parameter α^D.[5] Note also that the control variables ζ_i appear explicitly in the objective, but the association variables $p_i^D(x)$ are "hidden" inside ρ_i^D (see Eq.(8)), which serve as auxiliary control variables.

Optimizing now *both* DL and UL performance is a typical *multi-criterion* optimization problem [26], the DL and UL performance being coupled through the TDD variables ζ_i. We are thus interested in finding *Pareto efficient* operating points. Let us denote the left sum as $\phi_\alpha^D(\mathbf{p}^D, \zeta)$ and the right sum as $\phi_\alpha^U(\mathbf{p}^U, \zeta)$, for convenience. A solution $\{\mathbf{p}^{D*}, \mathbf{p}^{U*}, \zeta^*\}$ is Pareto efficient if for any other feasible $\{\mathbf{p}^D, \mathbf{p}^U, \zeta\}$

$$\left(\phi_\alpha^D(\mathbf{p}^D, \zeta), \phi_\alpha^U(\mathbf{p}^U, \zeta)\right) \leq \left(\phi_\alpha^D(\mathbf{p}^{D*}, \zeta^*), \phi_\alpha^U(\mathbf{p}^{U*}, \zeta^*)\right)$$
$$\Rightarrow \{\mathbf{p}^D, \mathbf{p}^U, \zeta\} = \{\mathbf{p}^{D*}, \mathbf{p}^{U*}, \zeta^*\}$$

The above relation suggests that any other solution could perhaps improve the DL or the UL but not both. All Pareto efficient points can be found by scalarization [26], e.g. minimizing $\tau \cdot \phi_\alpha^D(\mathbf{p}^D, \zeta) + (1-\tau) \cdot \phi_\alpha^U(\mathbf{p}^U, \zeta)$ for different values of $\tau \in [0, 1]$.

Constraints: Constraint (8) defines the set of *auxiliary variables* ρ_i^D and ρ_i^U, where $\rho_i^D(x) = \rho_i^D(x, \zeta_i = 1)$ and $\rho_i^U(x) = \rho_i^U(x, \zeta_i = 0)$ (see Eq.(2)-(4)). Constraint (9) states that every location must be associated with one BS on the DL and one on the UL. Constraints (10) are box constraints for the association variables. Constraint (11) relates to the minimum and maximum TDD resources that can be allocated to the DL (and UL).[6] Constraints (12) and (13) ensures the *stability* of each BS on the DL and UL direction. Finally, constraint (14) handles cross-interference, with $0 < q < 1$ being an input parameter (see A.9).

Lemma 3.1. *The feasible set of Problem 1 is convex.*

Proof: All problem constraints are linear in the control variables and define a feasible region that is contained in the domain of the objective function. ■

Lemma 3.2. *The objective of Problem 1 is non-convex. However, it is biconvex in the set of variables* $\{\mathbf{p}^D, \mathbf{p}^U\}$ *and* $\{\zeta\}$*.*

Proof: Even considering two BSs only, the Hessian matrix of the problem is not positive semidefinitive for any values of α^D, α^U. Hence, the objective is non-convex. However, consider each term in the first sum of Eq.(7) and assume $\alpha^D > 1$. For fixed ζ_i, the term $(1 - \rho_i^D/\zeta_i)^{1-\alpha^D}$ is convex in the variables p_i^D, because its a composition of convex function $f(x) = (x)^{1-\alpha^D}$ with an affine one in ρ_i^D: $1 - \rho_i^D/\zeta_i$. This in turn is affine on the control variables $p_i^D(x)$ (Eq.(8)). Finally, a sum of convex functions is also convex. (The cases of $\alpha^D = 0, 1$ follow easily). If $\{\mathbf{p}^D, \mathbf{p}^U\}$ is fixed, then ρ_i^D is constant. The term $(1-\rho_i^D/\zeta_i)^{1-\alpha^D}$ is convex in ζ_i, as a composition of a convex and *non-increasing* function ($f(x) = (x)^{1-\alpha^D}$) with a concave one in ζ_i, namely $1 - \rho_i^D/\zeta_i$ [26]. ■

Summarizing, the joint optimization problem in hand is biconvex, in a convex feasible region. There are therefore two main obstacles to overcome, in order to achieve a jointly optimal user association and TDD allocation: *Non-convexity:* Biconvex problems cannot generally be solved efficiently with a guaranteed convergence to the global optimum [27]. *Complexity of centralized solution:* A centralized solution of this problem requires collecting a lot of information from a large number of users and BSs, and might not scale well.

Instead, *our proposal is to decompose the problem into subproblems of decreased complexity that are solved in an iterative manner and provably converge to the global minimum.* Furthermore, we show that these subproblems can be easily distributed among BSs and UEs.[7]

As a final note, Problem 1 applies to a dual-connectivity setup (see A.3). When each UE must be connected to the same BS for both DL and UL traffic, the following (linear) set of constraints must be added to Problem 1:

$$p_i^D(x) = p_i^U(x), \forall x \in \mathcal{L}. \tag{15}$$

This additional constraint does not change the convexity properties of the problem, nor our main algorithmic approach. We therefore focus on the UL/DL split case.

IV. Problem Decomposition

As shown in Lemma 3.2, the nonconvex objective in Eq. (7) becomes convex, if some of the variables are kept constant. To facilitate the solution of the joint problem, we therefore propose two things: (i) to decompose it into two *convex* subproblems, one corresponding to user association (given the TDD allocation), and another corresponding to TDD configuration (given the association decisions). We show that these problems can be solved efficiently and in a distributed manner; (ii) to solve them in an iterative manner, fixing a TDD allocation, optimizing the associations, then improving the TDD allocation a bit, then re-optimizing associations and so forth, until convergence. This is often referred to as a *primal decomposition*.

Let us denote the objective function of Eq.(7) compactly as

$$f(\mathbf{p}^D, \mathbf{p}^U, \zeta) = \tau \cdot \phi_\alpha^D(\mathbf{p}^D, \zeta) + (1-\tau) \cdot \phi_\alpha^U(\mathbf{p}^U, \zeta). \tag{16}$$

Our proposed decomposition is sketched in Algorithm 1. $k-1$ and k are indices simply denoting the respective quantities in the previous and current iteration of the algorithm.

We will first consider the detailed implementation of the *master* (Sec. IV-A) and *inner* (Sec. IV-B) problems. We will then prove the convergence of Algorithm 1 to the global optimum solution (Theorem 4.4) of the joint problem.

[5] E.g. setting $\alpha^D = 0$ leads to maximizing spectral efficiency. Increasing α^D leads to a tradeoff between spectral efficiency and balanced loads (e.g. $\alpha = 1$: maximizes per flow throughput; $\alpha = 2$ minimizing mean per flow delay; $\alpha \to \infty$: leads to max-min fairness in terms of loads); We refer the reader to [25] for an extensive discussion of α-fair functions.

[6] In current specifications of dynamic TDD for LTE $\zeta_{min} = 0.4$ and $\zeta_{max} = 0.9$ [4]. To generalize this, we'll assume that $\zeta_{min} = \epsilon$ and $\zeta_{max} = 1 - \epsilon$ (where $\epsilon > 0$ is a small constant)

[7] This problem decomposition can be applied in a centralized implementation as well, e.g. an SDN controller. However, such a centralized algorithm might not be able to update association decisions and TDD configuration for each cell fast enough (e.g. at each frame).

Algorithm 1 Primal Decomposition of Problem 1 into User Association and TDD Allocation Subproblems.

```
1: **Repeat** until $\|\zeta(k) - \zeta(k-1)\| < \epsilon$.
2: **Inner Problem** (User Association):
       $\{\mathbf{p}^D(k), \mathbf{p}^U(k)\} = \underset{\{\mathbf{p}^D, \mathbf{p}^U\}}{\text{argmin}}\, f\left(\mathbf{p}^D, \mathbf{p}^U, \zeta(k-1)\right),$   (17)
       subject to Eq.(8)-(10), Eq.(12)-(14).
3: **Master Problem** (TDD Allocation):
       $\zeta_i(k) = \zeta_i(k-1) + t_i(k)\Delta\zeta_i(k),$   (18)
   where $\Delta\zeta_i(k)$ is a descent direction (Eq.(19)) and $t_i(k)$
   an appropriate step size (see Section IV-A).
```

A. Master Problem (TDD Allocation)

There are plenty of methods to update the TDD allocation vector ζ. Convergence can generally be achieved with any subgradient of the objective in ζ [28]. Given that our objective function is differentiable in all feasible ζ, the *gradient descent direction* can be chosen:

$$\begin{aligned}\Delta\zeta_i(k) &= \tau\left(1 - \frac{\rho_i^D(k)}{\zeta_i(k-1)}\right)^{-\alpha^D} \frac{\rho_i^D(k)}{\zeta_i^2(k-1)} \\ &- (1-\tau)\left(1 - \frac{\rho_i^U(k)}{1-\zeta_i(k-1)}\right)^{-\alpha^U} \frac{\rho_i^U(k)}{(1-\zeta_i(k-1))^2}.\end{aligned} \tag{19}$$

Standard backtracking methods can be used to choose the *step size* $t_i(k)$ [26]. Finally, in practice we add a small noise vector with mean 0 to the gradient of Eq.(19). Such "noisy" gradient methods guarantee that the convergence point is not a saddle point [29], which is necessary as we'll show in Theorem 4.4. *The calculation of the next allocation vector ζ can thus be directly distributed among base stations, with each BS simply calculating Eq.(19) independently, using only locally available information (ρ_i and ζ_i).*

B. Inner Problem (User Association)

While the Master Problem updates of Eq.(18) can be performed in a distributed manner between BSs, the Inner Problem cannot yet be distributed between users (or even, between BSs). A key hurdle is the cross-interference constraint of Eq.(14) which *couples* the DL and UL loads of nearby base stations (and consequently the user association variables $\mathbf{p}^D$, $\mathbf{p}^U$ as well). To facilitate the distributed solution of the Inner Problem, we augment the objective of Eq.(17) with a penalty function for constraint (14).

Problem 2 (Augmented Inner Problem)**.** *The following optimization problem is equivalent to the Inner Problem (Step 2) of Algorithm 1, when $\gamma \to \infty$.*

$$\min_{\{\mathbf{p}^D, \mathbf{p}^U\}} f\left(\mathbf{p}^D, \mathbf{p}^U, \zeta(k-1)\right) + \gamma \textstyle\sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^D + \rho_j^U - q)^2, \tag{20}$$

subject to Eq.(8)-(10), Eq.(12)-(13),

where $\mathcal{I}_{ij}$ is the indicator variable capturing whether BS i cross interferes with BS j, and γ is a penalty factor.

$$\mathcal{I}_{ij} = \begin{cases} 1, & \textit{when } \rho_i^D + \rho_j^U \geq q \\ 0, & \textit{otherwise.} \end{cases} \tag{21}$$

Quadratic penalty functions like the above are common and preserve convexity [30]. Note that Eq.(14) is no longer in the list of explicit constraints. Also, observe that for any finite value of γ the cross-interference constraint can now be slightly violated (if it helps improve the objective). We first show how to optimally solve Problem 2 for a fixed γ, then show that we can obtain the optimal solution of the original (non-augmented) by iteratively increasing γ, and re-solving Problem 2 (Theorem 4.3).

All explicit constraints are now independent for each BS, with respect to auxiliary variables ρ_i^D and ρ_i^U. To proceed further, we use a *first-order optimality condition for the augmented objective with respect to auxiliary variables ρ_i^D, ρ_i^U to derive optimal distributed policies for the association rules $\{\mathbf{p}^D, \mathbf{p}^U\}$*. We present our results and algorithm for the DL case, as the UL case is symmetric.

Lemma 4.1 (Inner Optimal Association Rules)**.** *If $\rho^{\mathbf{D}*} = (\rho_1^{D*}, \rho_2^{D*}, \cdots, \rho_{|\mathcal{B}|}^{D*})$ denotes the optimal load vector for the Augmented Inner Problem 2, the respective optimal DL association rule for a user at location x is given by $p_{i^*}^D(x) = 1$, $p_j^D(x) = 0, \forall j \neq i^*$, where $i^* =$*

$$\arg\max_{i\in\mathcal{B}} \left(c_i^D(x) \frac{\zeta_i(k-1)\cdot\left(1 - \frac{\rho_i^{D*}}{\zeta_i(k-1)}\right)^{\alpha^D}}{1 + 2\gamma\zeta_i(k-1)\left(1 - \frac{\rho_i^{D*}}{\zeta_i(k-1)}\right)^{\alpha^D} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{D*} + \rho_j^{U*} - q)} \right). \tag{22}$$

The optimal association rule for UL is given by replacing D with U, ζ_i with $1 - \zeta_i$, and exchanging indices i with j inside the term $(\rho_i^{D*} + \rho_j^{U*} - q)$.

Proof: The objective function of Problem (2) is only a function of the auxiliary variables ρ_i^D, ρ_i^U. As Problem (2) is convex, a sufficient (first-order) condition for the auxiliary variable vector ρ to be optimal is

$$\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle \geq 0, \text{ where } \Delta\rho^* = \rho - \rho^*, \tag{23}$$

where ρ is in the feasible region of Problem (2), and $\nabla\Phi(\rho^*)$ is the gradient of the objective of Problem 2.

If we consider the (partial) inner product, only along the DL variables ρ_i^D, we can write it explicitly as

$$\begin{aligned}&\sum_{i\in\mathcal{B}} \left(\frac{1}{\zeta_i\left(1 - \rho_i^{D*}/\zeta_i\right)^{\alpha^D}} + 2\gamma \sum_{j\in\mathcal{C}_i} I_{ij}(\rho_i^{D*} + \rho_j^{U*} - q) \right) (\rho_i^D - \rho_i^{D*}) = \\ &\sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma\cdot\zeta_i\left(1 - \rho_i^{D*}/\zeta_i\right)^{\alpha^D} \sum_{j\in\mathcal{C}_i} I_{ij}(\rho_i^{D*} + \rho_j^{U*} - q)}{\zeta_i\left(1 - \rho_i^{D*}/\zeta_i\right)^{\alpha^D}} \right) \\ &\qquad \cdot \int_{\mathcal{L}} \rho_i^D(x)\left(p_i^D(x) - p_i^{D*}(x)\right) dx = \\ &\int_{\mathcal{L}} \lambda^D(x)\overline{s}^D(x) \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma\zeta_i\left(1 - \frac{\rho_i^{D*}}{\zeta_i}\right)^{\alpha^D} \sum_{j\in\mathcal{C}_i} I_{ij}(\rho_i^{D*} + \rho_j^{U*} - q)}{c_i^D(x)\cdot\zeta_i\cdot\left(1 - \rho_i^{D*}/\zeta_i\right)^{\alpha^D}} \right). \\ &\qquad \cdot \left(p_i^D(x) - p_i^{D*}(x)\right) dx.\end{aligned}$$

However, the following holds

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

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

because $p_i^{D*}(x)$ is an indicator for the minimizer of the term in the parenthesis (see Eq.(22)). A similar argument holds for the partial inner product along the uplink variables ρ_i^U. This proves that $\langle\nabla\Phi(\rho^*),\Delta\rho^*\rangle\geq 0$ and the association rules of Eq.(22) define an optimal point for the inner problem. ■

The above association rules say what should hold at the optimal operating point, but not how to get there (i.e., when the base station loads ρ_i are not optimal). However, replacing ρ_i^{D*} with ρ_i^D (respectively for the UL) in the above association rules, suffices to define a *descent direction*. Based on this, we define Algorithm 2 which is a distributed, iterative algorithm that can converge to the optimal user associations (and loads) from an arbitrary initial association assignment.

Algorithm 2 Distributed, Iterative Optimization Algorithm for Inner Problem of Algorithm 1

```
1: Repeat until convergence.
2: Base Station (BS): (starting with a feasible BS load vector $\hat{\rho}(0)$) at iteration $n$, BS $i$ measures its current DL load $\rho_i^D(n)$ based on the current user associations $p_i^D(n)$ (see Eq. (3)) and then updates its load estimate as
       $$\hat{\rho}_i^D(n)=(1-\beta)\cdot\hat{\rho}_i^D(n-1)+\beta\cdot\rho_i^D(n), \qquad (24)$$
   where $\beta\in(0,1)$ is a parameter of an exponential moving average. A similar estimate $\hat{\rho}_i^U(n)$ is maintained for UL.
3: BS $i$ broadcasts to all UEs the complex fraction of Eq.(22).
4: User Equipment (UE): at iteration $n$, any user on location $x$ associates with BS $i^*=$
       $$\arg\max_{i\in\mathcal{B}}\left(c_i^D(x)\cdot\frac{\zeta_i(k-1)\left(1-\frac{\hat{\rho}_i^D(n)}{\zeta_i(k-1)}\right)^{\alpha^D}}{1+2\gamma\zeta_i(k-1)\cdot\left(1-\frac{\hat{\rho}_i^D(n)}{\zeta_i(k-1)}\right)^{\alpha^D}\sum\limits_{j\in\mathcal{C}_i}\mathcal{I}_{ij}(\hat{\rho}_i^D(n)+\hat{\rho}_j^U(n)-q)}\right). \qquad (25)$$
   Similarly for the UL association (see Lemma 4.1).
```

The algorithm is fully distributed between BSs and UEs. At each iteration n: (step 2) each BS i measures its own DL load $\rho_i^D(n)$, and updates its estimate $\hat{\rho}_i^D(n)$. Also BS i receives (e.g. through the LTE X2 interface) a tuple $\{\hat{\rho}_j^D(n),\hat{\rho}_j^U(n)\}$ from its interfering BSs $j\in\mathcal{C}_i$ (this tuple is needed for the calculation of the sum in the denominator of the complex fraction in Eq. (22)). Then, (step 3) BS i is able to calculate and broadcast the complex fraction of Eq. (22) to all UEs in range[8]. Each user, (step 4) based on (i) the complex fraction received from all BSs in range, and (ii) the measurement of $c_i^D(x)$, is able to *locally* evaluate Eq. (22) and select the optimal BS for association. The steps are repeated until all BS load estimates converge (step 1). It is important to note that this algorithm is *scalable*, as it requires a constant amount of BS broadcast messages per round (only one value for DL and one for UL) irrespective of the number of users and interfering BSs, *of low complexity*, as each user receives information from the BSs in range only and then performs a max operation, and offers *flexible* performance (by tuning the α values).

Interpreting the association rule, if choosing a BS i does not lead to cross-interference with some neighboring BSs (i.e., $\mathcal{I}_{ij}=0,\forall j\in\mathcal{C}_i$), the UE associates to a base station only according to term $(c_i^D(x)\cdot\zeta_i(k-1)(1-\frac{\hat{\rho}_i^D(n)}{\zeta_i(k-1)})^{\alpha^D})$ (namely its congestion level, and the maximum rate at location x). However, when BS i cross interferes with another BS, an additional term in the denominator penalizes BS i.

The above algorithm implements a distributed gradient on the auxiliary variables ρ_i, and can be shown to converge to a unique fixed point, which is the global optimum of Problem 2. We omit the proof due to space limitations.

Lemma 4.2. *For a fixed value of γ, Algorithm 2 converges to the optimal solution of Problem 2.*

The following theorem further states that minimizing a sequence of cost functions of Problem 2 with increasing values for γ, converges to the optimal solution of the Inner Problem. We drop the "hat" notation and use ρ for simplicity.

Theorem 4.3. *Let $\{(\rho^{(0)},\gamma^{(0)}),(\rho^{(1)},\gamma^{(1)}),...,(\rho^{(k)},\gamma^{(k)})\}$ denote a sequence of optimal loads $\rho^{(k)}$, generated by Algorithm 2 for increasing values of $\gamma^{(k)}$. Then, any limiting point of this sequence is the minimum of the Inner Problem of Algorithm 1.*

Proof: Let us denote the cross-interference penalty term of Eq.(20), for a given load vector ρ, as $P(\rho)=\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{C}}(\rho_i^D+\rho_j^U-1)$. Let $\Phi(\rho,\gamma)$ further denote the augmented objective of Eq.(20), for a given load vector and penalty constant γ (ζ is a constant throughout the inner loop):

$$\Phi(\rho,\gamma)=f(\rho)+\gamma\cdot P(\rho). \qquad (26)$$

Let $\tilde{\rho}$ be the limiting point of the sequence $\rho^{(k)}$. From the continuity of $f(\cdot)$ we have

$$\lim_{k\to\infty} f(\rho^{(k)})=f(\tilde{\rho}). \qquad (27)$$

Furthermore, the sequence of values of $\Phi(\rho^{(k)},\gamma^{(k)})$ are non-decreasing:

$$\begin{aligned}\Phi(\rho^{(k)},\gamma^{(k)})&=f(\rho^{(k)})+\gamma^{(k)}\cdot P(\rho^k)\\&\leq f(\rho^{(k+1)})+\gamma^{(k)}\cdot P(\rho^{k+1})\\&\leq f(\rho^{(k+1)})+\gamma^{(k+1)}\cdot P(\rho^{k+1})=\Phi(\rho^{(k+1)},\gamma^{(k+1)}).\end{aligned}$$

The first inequality holds because $\rho^{(k)}$ is the optimal value of Problem 2 for penalty constant $\gamma^{(k)}$, and thus $\Phi(\rho^{(k)},\gamma^{(k)})\leq\Phi(r,\gamma^{(k)})$ for any other r, including $r=\rho^{(k+1)}$. The second inequality holds because $\gamma^{(k+1)}\geq\gamma^{(k)}$.

Let Φ^* be the optimal value of the inner user association problem (Inner Problem of Step 2). Then, the above sequence of values is bounded above by Φ^* for every k (since the augmented problems are relaxations of the original one). To prove this, let us denote with ρ^* the point corresponding to

[8] Note that ζ_i remains *constant* throughout the execution of Algorithm 2, and is only updated in the master problem of Algorithm 1.

the value Φ^*. Then, $P(\rho^*) = 0$ (since constraints are strictly satisfied in the non-augmented problem), and for any finite $\gamma^{(k)}$ it holds that:

$$\Phi^* = f(\rho^*)+\gamma^{(k)}\cdot P(\rho^*) \geq f(\rho^{(k)})+\gamma^{(k)}\cdot P(\rho^{(k)}) \geq f(\rho^{(k)}). \tag{28}$$

The relations imply that the following limit is a real number:

$$\lim_{k\to\infty} \Phi(\rho^{(k)}, \gamma^{(k)}) = q^* \leq \Phi^*. \tag{29}$$

Subtracting (29) from (27) yields

$$\lim_{k\to\infty} \gamma^{(k)} P(\rho^{(k)}) = q^* - f(\tilde{\rho}). \tag{30}$$

Since $P(\rho^{(k)}) \geq 0$ and $\gamma^{(k)} \to \infty$, Eq. (30) implies

$$\lim_{k\to\infty} P(\rho^{(k)}) = 0. \tag{31}$$

Using the continuity of $P(\cdot)$, this means that $P(\tilde{\rho}) = 0$ and thus $\tilde{\rho}$ is feasible. To show that $\tilde{\rho}$ is optimal we note from Eq. (28) that also $f(\rho^{(k)}) \leq \Phi^*$, and hence $f(\tilde{\rho}) = \lim_{k\to\infty} f(\rho^{(k)}) \leq \Phi^*$. ■

Given the optimality of the solution of the inner (user association) problem, shown so far, the following theorem suggests that the decomposition of the joint problem shown in Algorithm 1 leads to the optimal solution of the joint user association and dynamic TDD allocation.

Theorem 4.4. *Let $\{(\rho,\zeta)^k\}$ be the sequence generated by Algorithm 1. Then, the algorithm converges, and a limiting point of $\{(\rho,\zeta)^k\}$ is the desired global optimum of Problem 1.*

Proof: Our algorithm falls into the category of the popular Block Coordinate Descent (BCD) or nonlinear Gauss-Seidel method. As a special subcase, in primal decomposition algorithms, such methods usually employ a (sub-)gradient criterion for the master problem to control the inner problem blocks and monotonically improve the cost function. Our results on the individual subproblems (inner and master) of Algorithm 1 indeed prove that at the end of the k^{th} iteration

$$\phi_\alpha(\rho,\zeta)^{(k)} < \phi_\alpha(\rho,\zeta)^{(k-1)}.$$

It is well-known that such a generated sequence has a stationary point [28], [31]

For non-convex problems, this stationary point can generally be: (i) a saddle point, (ii) a local minimum, or (iii) the desired global minimum. We show below why we can exclude options (i) and (ii), in our case.

(1) *Saddle point escape:* As explained in Section IV-A the noisy gradient update we perform allows the algorithm to escape saddle points [29].

(2) *Unique optimum point:* Let us consider the problem in terms of the TDD variables ζ_i and auxiliary load variables ρ_i (finding optimal loads ρ_i also gives optimal association rules $p_i(x)$, through Lemma 4.1). While Problem 1 is non-convex (Lemma 3.2) it can be converted into a geometric program (GP). Instead of ζ_i, let us use the equivalent set of variables

$$\zeta_i^D = \zeta_i, \quad \zeta_i^U = 1 - \zeta_i$$

by adding one more constraint: $\zeta_i^D + \zeta_i^U \leq 1$. Observe now that all constraints involving ρ_i^D, ρ_i^U and ζ_i^D, ζ_i^U can be easily rewritten as standard *posynomial* inequalities [26]. E.g. Eq.(12) can be written as $(1-\epsilon)^{-1}\cdot\rho_i^D\cdot(\zeta_i^D)^{-1} \leq 1$ and Eq.(13) as $(1-\epsilon)^{-1}\cdot\rho_i^U\cdot(\zeta_i^U)^{-1} \leq 1$.

Regarding the objective, for $\alpha^D = 0$, $\alpha^U = 0$, it is easy to see that it's a posynomial in ρ_i and ζ_i. However, this is not the case for different values of α, and we need to perform some additional transformations. We introduce the new auxiliary variables $e_i^D \geq 0$, $e_i^U \geq 0$, where

$$e_i^D = 1 - \rho_i^D/\zeta_i^D \text{ and } e_i^U = 1 - \rho_i^U/\zeta_i^U. \tag{32}$$

The problem objective (Eq.(7)) can then be rewritten as

$$\tau \cdot \sum_{i\in\mathcal{B}} \frac{(e_i^D)^{1-\alpha^D}}{\alpha^D - 1} + (1-\tau)\cdot \sum_{i\in\mathcal{B}} \frac{(e_i^U)^{1-\alpha^U}}{\alpha^U - 1}, \tag{33}$$

which is also a posynomial in the variables e_i^D, e_i^U (trivially, as each term in the sum is a power of a single variable, and all coefficients are ≥ 0). Minimization takes place now in terms also of these auxiliary variables, subject to the additional constraints of Eq.(32). These constraints can be rewritten as

$$e_i^D + \rho_i^D \cdot (\zeta_i^D)^{-1} = 1 \text{ and } e_i^U + \rho_i^U \cdot (\zeta_i^U)^{-1} = 1.$$

Unfortunately, equality constraints require a monomial, in the standard GP form, while the above left handsides are posynomials. Nevertheless, we can relax these into inequalities, by replacing "= 1" with "≤ 1". Observe that minimizing the objective of Eq.(33) for $\alpha^{D/U} \geq 1$ requires picking as high values for e_i^D and e_i^U, as possible. Hence, at the optimal point, these constraints should be tight, and the new problem is equivalent with the original.

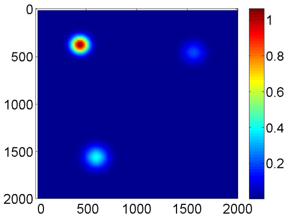
This concludes that the problem in hand has an equivalent GP form. Furthermore, GP problems can be converted to a convex equivalent with standard transformations [26], which has no local minima. Given that all transformations we did preserve a *bijection* between the equivalent problem and the original, this proves that our original non-convex problem also has no local minima and thus the point of convergence must be a global minimum. ■

V. Simulations

In this section, we evaluate the proposed algorithm in some representative scenarios. 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 qualitative (and often quantitative) benefits can be observed there.

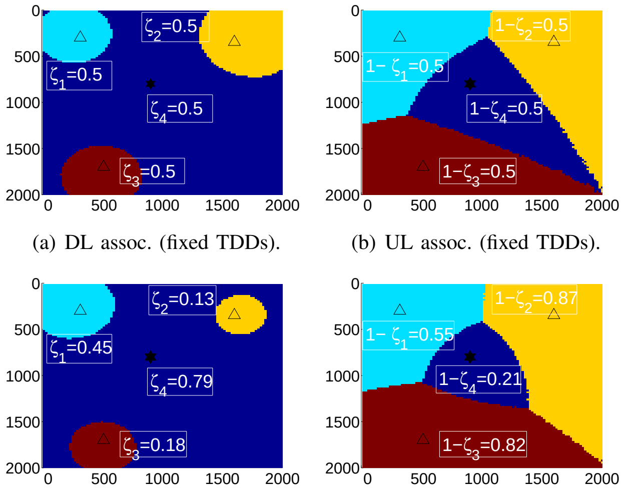
Scenario 1: We consider a 2×2 km^2 area. The figure below shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ (flows/hour per unit area). Out of this rate 70% is DL and 30% is UL, on average, with 3 hotspots (blue implying low traffic and red high). Without loss of generality, we assume that the macro BS cross interferes with all SCs (i.e., $C_4 = \{1,2,3\}$, $C_1 = C_2 = C_3 = \{4\}$, see A.9). We consider standard parameters as adopted in 3GPP [32], listed in Table II. We set $\alpha^D = \alpha^U = 1$ to optimize user throughput.

Coverage Snapshots: We first look at the coverage maps that different schemes create. We assume that the area is covered by one macro cell in the center (shown as a star with



Variable	Value
$P_{eNB,SC,UE}$	43,24,12 dBm
$w_i,\ \forall i \in \mathcal{B}$	10 MHz
N_0	-174 dBm/Hz
$\overline{s}^{D,U}(x)$	100,25KB $\forall x$
$\lambda^D(x)/\lambda^U(x)$	7/3
α^D, α^U	1,1

Fig. 3 & TABLE II: Traffic arrival rate and other simulation parameters.



(a) DL assoc. (fixed TDDs).

(b) UL assoc. (fixed TDDs).

(c) DL assoc. (flexible TDDs).

(d) UL assoc. (flexible TDDs).

Fig. 4: DL and UL user associations for different scenarios ($\tau = 0.5$).

BS number 4) and three equidistant SCs (shown as triangles using BS numbers 1-3). Fig. 4(a), 4(b) depict the optimal user associations for a static LTE-TDD configuration with UL/DL timeslot ratio $4:4$ i.e., fixed $\zeta_i = 0.5, \forall i \in \mathcal{B}$. 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). The DL coverage areas of the SCs also depend on 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).

We then allow ζ_i (i.e. the TDD schedule) at each BS to vary, and apply the proposed joint association and TDD allocation algorithm. The resulting coverage maps and radio allocations (optimal ζ_i) are shown in Fig. 4(c), 4(d). We note that macro BS increases its $\zeta_4 = 0.79$ to serve more DL users, and SC increase their UL resources $1-\zeta_1 = 0.55, 1-\zeta_2 = 0.87, 1-\zeta_3 = 0.82$ to serve more UL. The joint algorithm has attempted to better match available resources to demand: e.g., SC 2 which has low traffic around it, has reduced its DL resources considerably (from 0.5 to 0.13), since this suffices to serve DL users, and has increased its UL resources signficantly to help the UL (who tends to suffer from the lower UE Tx power). Observe that this is not the case for SC 1 though, which has maintained its DL resources and coverage, as it lies on top of a hotspot. As we will see shortly, this new configuration is able to *simultaneously* improve both UL and DL performances.

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 between DL and UL performance). We compare the performance of the following three schemes:

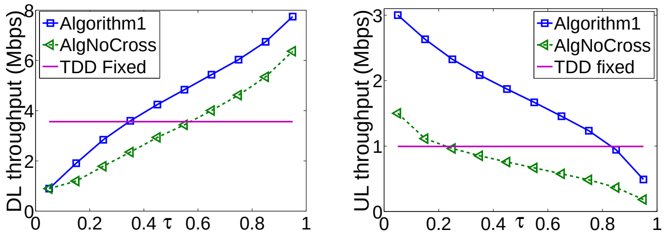
(**TDD Fixed**): Optimal user association algorithm with equal UL/DL resources for each BS ($\zeta_i = 0.5$).

(**Algorithm1**): Our proposed joint Algorithm 1.

(**AlgNoCross**): To better understand the importance of considering the cross-interference constraints, we implement a variation of Algorithm 1, where we do not take cross-interference into account. Any conflicts in the neighboring optimal UL/DL schedules leading to cross-interference are included in the SINR and the resuling rate decrease.

In Fig. 5 we depict the DL and UL user throughput as a function of τ in different scenarios. Observe that *Algorithm 1* can *significantly improve* DL or UL performance up to $2-3\times$, compared to *TDD fixed*, by an appropriate choice of τ. More importantly, for most intermediate τ values, it is able to *simultaneously improve DL and UL performance*. E.g., for $\tau = 0.5$ DL improves by approximately $15-20\%$ while UL imporves by almost 2×. Finally, observe that even when TDD fixed performs better in one direction (e.g. UL for $\tau \to 1$) the sum of UL and DL rates is better for Algorithm 1.

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. Finally, when $\tau \to 0$, all BSs operate almost exclusively on the UL, so DL performance suffers mostly due to limited resources rather than $UL \to DL$ cross-interference. While the two curves converge in the DL direction, *AlgNoCross* performs significantly worse in the UL. These observations underline the importance of directly considering cross interference constraints in our optimization framework through Eq.(6). In the remainder of this section we therefore only consider Algorithm 1 compared to TDD fixed.



(a) DL throughput.

(b) UL throughput.

Fig. 5: User-centric Performance.

Network-centric performance. Table III considers the performance improvement of *Algorithm 1* over *TDD Fixed*, in terms of network-related metrics, when $\tau = 0.5$ (equal importance between UL and DL). We consider two metrics: Spectral Efficiency (**SE**) in terms of bits/s/Hz, and Load

Balancing (**LB**) in terms of mean square difference between different BS loads. DL/UL spectral efficiency improves up to 44% since *flexible TDD better allocates the resources* with respect to the heterogeneous transmit powers that help physical data rates improve (see A.5). It also correctly considers related traffic statistics and asymmetries across users, diminishing BS under/over utilizations, and thus LB is improved. Note that simultaneous improvement of both these metrics implies improvement in overal user performance (see Eq.(5)).

TABLE III: Network (SE,LB) Performance (τ = 0.5)

	Downlink		Uplink	
Performance.	SE	LB	SE	LB
Percentage % of improvement.	44	17	44	55

Scenario 2: Having highlighted the sources of performance improvement in the basic scenario above, we now turn our attention to a larger network consisting of 4 macro BSs and 13 SCs with uniform traffic demand. Considerable improvements can be observed in this scenario too, as can be seen from Table IV (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 (*Algorithm1*) considerably improves the performance of the problematic BSs, average performance is not as affected; (b) the *additional cross interference* from a larger number of BSs reduces the range of permissible ζ_i values (i.e. TDD configurations) that can be considered.

TABLE IV: User (UE) and Network (SE, LB) Performance (τ = 0.5)

	Downlink			Uplink		
Scenario.	UE	SE	LB	UE	SE	LB
Percentage % of improvement.	30	41	6	86	43	52

VI. Conclusion

In this paper, we formulated a novel, distributed algorithm that jointly tackes the coupled problems of (i) user association, and (ii) TDD resource allocation, under *cross interference* constraints. Using optimization theory we proved that our algorithm converges to the global optimum. Simulation results corroborate the correctness of our framework and reveal promising qualitative and quantitative results, in terms of both user and network performance improvement.

References

[1] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Communications Surveys & Tutorials*, 2015.

[2] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Communication Surveys & Tutorials*, 2016.

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

[4] 3GPP, "TS 36.300, Release 13 (version 13.2.0)," 2016.

[5] J. Kerttula, A. Marttinen, K. Ruttik, R. Jäntti, and M. N. Alam, "Dynamic tdd in lte small cells," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 194, Aug 2016.

[6] S. Sesia, I. Toufik, and B. M., *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley, 2011.

[7] 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.

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

[9] E. L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proc. of IEEE INFOCOM*, 2008.

[10] 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*, 2016.

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

[12] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "User association in hetnets: Impact of traffic differentiation and backhaul limitations," *IEEE/ACM Transactions on Networking*, 2017.

[13] A. Ravanshid, P. Rost, D. S. Michalopoulos, V. V. Phan, H. Bakker, D. Aziz, S. Tayade, H. D. Schotten, S. Wong, and O. Holland, "Multi-connectivity functional architectures in 5G," in *Proc. of IEEE ICC*, 2016.

[14] A. G. Gotsis, S. Stefanatos, and A. Alexiou, "Optimal user association for massive MIMO empowered ultra-dense wireless networks," *IEEE ICC Workshops*, 2015.

[15] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter wave based ultra dense networks with energy harvesting base stations," *to appear in IEEE Journal on Selected Areas in Communication*, 2017.

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

[17] A. K. Gupta, M. N. Kulkarni, E. Visotsky, F. W. Vook, A. Ghosh, J. G. Andrews, and R. W. Heath, "Rate analysis and feasibility of dynamic TDD in 5g cellular systems," in *Proc. of IEEE ICC*, 2016.

[18] Y. Zhong, P. Cheng, N. Wang, and W. Zhang, "Dynamic tdd enhancement through distributed interference coordination," in *IEEE ICC*, 2015.

[19] H. Sun, M. Sheng, M. Wildemeersch, T. Q. S. Quek, and J. Li, "Traffic adaptation and energy efficiency for small cell networks with dynamic tdd," *IEEE Journal on Selected Areas in Comm.*, 2016.

[20] 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.

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

[22] 3GPP, "TR 36.842, Release 12 (version 12.0.0)," 2014.

[23] 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.

[24] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. of ACM Mobicom*, 2003.

[25] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, Oct. 2000.

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

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

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

[29] 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.

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

[31] A. M. Stephen Boyd, Lin Xiao and J. Mattingley, "Notes on decomposition methods," *Stanford University*, 2008.

[32] 3GPP, "TR 36.931 Release 13 (version 13.2.0)," 2016.