Joint Power Control and Access Point Scheduling in Fronthaul-Constrained Uplink Cell-Free Massive MIMO Systems

Mamoun Guenach, *Senior Member, IEEE*, Ali A. Gorji, and André Bourdoux, *Senior Member, IEEE*

***Abstract*— Cell-free (CF) massive Multiple-Input Multiple-Output (MIMO) with large number of distributed access points (APs) has emerged as a new paradigm allowing higher macro diversity for randomly distributed users. However, the fronthaul traffic bandwidth between central processing unit and the APs can explode in particular in the uplink, requiring expensive star-topology with point-to-point fronthaul links. To achieve a scalable CF massive MIMO architecture and a cost-effective fronthauling solution, we consider, in this paper, a point-to-multipoint fronthaul topology where (a subset of) the APs share a serial fronthaul link offering a per-user limited fronthaul bandwidth. We develop a novel unified optimization framework for iterative power control and AP scheduling that provides a systematic user-centric solution towards scalable uplink CF massive MIMO. Experimental results show that power control is not sufficient to guarantee the best objective and, therefore, the appropriate association of the users to the APs is required to improve the overall system signal-to-noise ratio. Under the stringent fronthaul bandwidth, the proposed joint optimization framework results in i) significant 5% outage data rate increase ii) near uniform distribution of the served users per APs and, hence, an increased diversity and iii) fast convergence of the algorithm within a few iterations.**

***Index Terms*— Cell-Free massive MIMO system, power control, access point scheduling, serial fronthaul.**

I. Introduction

THE massive wireless traffic growth goes hand-in-hand with the deployment of advanced radio interfaces as well as network densification. This, amongst others, has direct impact on the radio access architecture that is, nowadays, moving from centralized (e.g. macro-cells) to distributed (e.g. small-cells) deployments through the use of large number of collocated or distributed antennas or access points (APs), a concept referred to as massive Multiple-Input Multiple-Output (MIMO) [1].

Massive MIMO allows high beamforming gain and spatial multiplexing of the users [2]–[4], and has become a de facto key enabler of 5G systems. Following this trend, radio access networks are undergoing a deep transformation by distributing the APs close to the end-users and interconnecting them with a Central-Processing Unit (CPU) as shown in Fig. 1. In fact, wireless networks are moving away from the cell-centric paradigm, which is limited by the inter-cell-interference, to ubiquitous Cell-Free (CF) massive MIMO that is coherently user-centric with higher resilience to the interference and additional macro-diversity [5]. CF massive MIMO will most likely become an important component in the future wireless connectivity landscape of 6G (and beyond) featuring low-complexity, high-throughput, ultra-reliable and low-latency applications [6]. Benefiting from the channel hardening and favorable propagation properties [7], [8] resulting from the law of large numbers, higher Spectral Efficiency (SE) and reliability is achieved with CF massive MIMO. As a matter of a fact, the deterministic behavior of the effective channel simplifies both the power control and the AP scheduling problems with the final design becoming independent of the small-scale fading variations. As a result, the overall aforementioned problem only depends on the large scale fading that changes slowly with respect to the Time Division Duplxing (TDD) frame rate. Favorable propagation, on the other hand, means that multiple users can be spatially multiplexed with reduced mutual interference.

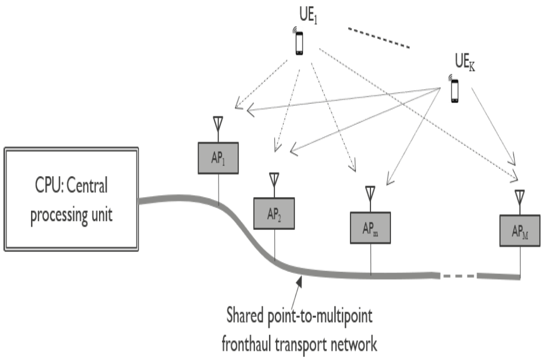
UE$_1$
UE$_K$
AP$_1$
AP$_2$
AP$_m$
AP$_M$
CPU: Central processing unit
Shared point-to-multipoint fronthaul transport network

Fig. 1. CF massive MIMO with a serial wired fronthaul.

A particular problem that hinders scalable CF massive MIMO deployment is the traffic bandwidth (BW) growth from/to CPU to/from APs, which can also explode with the massive number of distributed APs. This traffic needs to be fronthauled over wired links and, for demanding designs with high fronthaul (FH) throughput, a star-topology with

Manuscript received July 31, 2020; revised November 6, 2020; accepted December 18, 2020. Date of publication December 28, 2020; date of current version April 16, 2021. The associate editor coordinating the review of this article and approving it for publication was W. Chen. *(Corresponding author: Mamoun Guenach.)*

Mamoun Guenach is with imec, 3001 Leuven, Belgium, and also with the Department of Telecommunications and Information Processing, Ghent University, 9000 Gent, Belgium (e-mail: guenach@ieee.org).

Ali A. Gorji and André Bourdoux are with imec, 3001 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2020.3047801.

Digital Object Identifier 10.1109/TCOMM.2020.3047801

Point-to-Point (P2P) FH links is used. An alternative - and more cost effective - architecture is the Point-to-Multi-point (P2MP) consisting of a non-overlapping subset of APs sharing each a single FH serial link that offers a limited BW as depicted in Fig. 1. The radio-stripe architecture recently proposed in [5], [10] is an example of such a promising architecture with a serial shared FH that offers cheap CF Massive MIMO deployment. In the proposed architecture, distributed APs are confined in cables (stripes) enabling invisible installation in existing construction elements.

Another important design element in the distributed CF massive MIMO is the split processing between the CPU and different APs that could directly impact the required FH BW. There are essentially two classes of split processing architectures [9]. In the first class, referred to as analog-connected architecture, all the digital baseband processing up to the digital-to-analog and the analog-to-digital converters lies in the CPU. The second class is the digitally-connected architecture with the entire analog front-end processing and a fraction of digital baseband processing being both distributed in the APs. One particular subclass of interest is the digitally connected architecture with distributed beamforming [11] with a two-fold benefit when combined with the TDD mode. First, the channels can be locally estimated at each AP using orthogonal pilot sequences [11] (see subsection II-B for further details). In addition, the low-complexity conjugate digital beamformers [11] can be determined locally, which significantly reduces the amount of traffic to be fronthauled. This split processing architecture is preferred for practical deployments as mainly the payload is fronthauled while the channel state information is kept locally in the APs [11]. Although Downlink (DL) enjoys from a reasonable FH BW, the Uplink (UL) fronthauling explodes (it scales with the number of users times the number of APs) and this explosion requires a careful development of joint power control and AP scheduling.

In [11], the authors compared the performance of CF massive MIMO and traditional cellular systems. They proposed optimal max-min fairness power-control algorithms by means of respectively linear and second-order cone optimization for both UL and DL transmission using conjugate beamforming. The authors in [12] proposed a low-complexity DL power control algorithm with a performance close to the optimal algorithms developed in [11]. While the authors have further developed a near-optimal power allocation with zero-forcing precoding, none of the above-mentioned articles has proposed a mechanism to deal with the user-AP association.

Joint power allocation and AP scheduling is a crucial problem for practical deployments of CF massive MIMO for the reasons discussed earlier. In practice, not all APs will contribute equally to the SE per user due to the path-loss that rapidly decays with the propagation distance [5] on top of the shadowing that can also block the link between certain APs and users. In [11], it has been shown for instance that, in DL, the minimum number of effective APs contributing to at least 95% of the total power allocated to a certain user is rather limited e.g. 10-20% of the APs in the one-squared km area surrounding a user.

The authors in [13] addressed the above-mentioned problem of the main contributing APs by proposing two AP selection schemes for the DL transmission. For the power allocation problem, the energy efficiency (EE), which is defined as the ratio of the sum SE to the total power, is maximized subject to per user-SE and per-AP allocated power being upper-bounded. For this purpose, a power model of the AP accounting for the amplifier efficiencies, and a power model of the backhaul, which is linear in the sum of the per user-SE, are proposed. The first proposed AP-selection algorithm is a max-min fairness power control with received-power-based AP selection, and the algorithm is implemented in two major steps. First, assuming all the APs are active, the power is optimized. Then, the subset of effective APs serving a given user is created by considering those APs that contribute to, at least, a fraction (e.g. 95%) of the total power assigned to the reference user. The second algorithm is a max-min fairness power control with channel-quality-based AP selection wherein the APs with the best channel quality (largest large-scale fading coefficients) towards a certain user are selected. While both proposed algorithms offer a systematic optimization framework to obtain the allocated power, they both rely on heuristics to specify a set of APs that could ultimately improve the overall performance in DL. The authors in [15] proposed to jointly optimize the DL power control and the ON/OFF mode of the APs to minimize the total transceiver power consumption subject to a per-user minimum DL ergodic SE. The resulting optimization problem is a complex mixed-integer second order cone optimization problem for which the authors devised two low-complexity heuristics based on the transmit power and/or the sparsity. In the same way, several heuristics for ON/OFF AP activity have been proposed in [16] to improve the EE of the CF massive MIMO, and have been further adapted to mmwave scenarios in [17].

Even though the optimal allocation problem has been extensively addressed in the literature, the optimal AP assignment has been tackled in an ad-hoc manner with no systematic or scalable framework being proposed for AP scheduling, at least, in the CF massive MIMO literature. More recently and in the context of distributed massive MIMO where each Remote Radio Head (RRH) is equipped with a high number of antennas (much higher than the number of active users), the joint optimization of user-RRH association and power antenna activation per RRH was addressed in [14]. In the user-RRH association, a user is constrained to communicate with only one RRH. Similar to [13], it turns out that the maximization of the EE that also adopts an elaborate power consumption model is an NP-hard problem. In [24], the authors investigated the joint power allocation and base station association for the UL of small-cell networks with non-orthogonal multiple access points. In the proposed scheme, each subset of users associated to a certain AP performs a superposition coding-based multiuser transmission followed by a serial interference cancellation at the reference AP that serves the aforementioned subset of users.

In this paper, we develop a novel unified optimization framework for iterative power control and AP scheduling in the UL of CF massive MIMO systems.

The main contributions of this paper are summarized next:

- A novel framework is proposed to deal with the joint power allocation and AP scheduling subject to a maximum power per user and to an upper bound on the shared FH BW. We address the non-convexity of the joint optimization problem by iterating between power allocation and AP assignment. Instead of the heuristics, this paper proposes two novel convex algorithms for AP scheduling, and tackles the overall joint problem as two separate convex optimization stages. We differ from [11] by focusing on joint power control and AP scheduling, while [11] was limited to power control only. We differ from [13] by investigating iterative power allocation and AP in the UL to address the limited shared FH BW, while [13] proposed two heuristics for DL based on the received power and large scale fading to improve the EE. We differ from [24] by i) using simple conjugate beamforming rather than serial interference cancellation, ii) considering CF massive MIMO where a user can be served by multiple APs rather than one AP at most, iii) maximizing the worst case Signal-to-Noise Ratio (SNR) rather than an increasing concave function of the total revenue per base station, and iv) by taking the limited shared FH BW into account to combine equalized data from different APs that serve a certain user.
- A thorough convergence analysis of the algorithms is conducted using the properties of the convex problems. The convergence of the algorithm is, especially, studied for large values of APs, and analytical terms are derived to justify the asymptotic convergence of the proposed framework.
- A detailed theoretical and empirical complexity analysis of the proposed algorithms is conducted.
- The optimality is assessed through a performance and complexity benchmarking with an exhaustive Nonlinear Mixed Integer Programming (NMIP) solver [26].
- A deep analysis of the proposed algorithms is done over a simulated scenario. The iterative algorithms are, especially, compared with a proposed heuristic for AP scheduling and the superior performance of our method for the joint optimization is proven against the state-of-art.

The rest of this paper is organized as follows. In section II, we describe the system model for both UL channel estimation and payload processing. In section III, a generic optimization framework is developed for the joint power allocation and AP scheduling where we propose two novel algorithms that each iterates between both convex power allocation and AP scheduling. A non-iterative heuristic for AP scheduling based on the large scale fading coefficients is described for benchmarking. Both convergence and complexity analysis of the proposed algorithms are discussed in detail, respectively, in sections IV and V. Numerical results are provided in section VI and conclusions are drawn in section VII.

II. System Model

A. General Assumptions

Consider a CF massive MIMO in Fig. 1 in which a CPU is connected through a serial wired FH link to a large number (M) of APs. We assume that each AP is equipped with N antennas receiving noisy data from the different users in the UL wherein $M \gg K$. A TDD of the UL and DL is considered in this paper. The main advantage here is to enable the estimation at the mth AP of the $N \times 1$ UL channel $\mathbf{g}_{m,k}^{(c)}$ between the mth AP and the kth user on the cth subcarrier of the Orthogonal Frequency Division Multiplexing (OFDM) modulation. In addition, the channel estimates can be further used for DL precoding by the virtue of the UL and DL channel reciprocity, assuming the AP front-ends non-reciprocities are estimated and compensated for [27].

The flat fading channel on the cth subcarrier is modeled as $\mathbf{g}_{m,k}^{(c)} = \sqrt{\beta_{m,k}}\mathbf{h}_{m,k}^{(c)}$ [13]. This model is typically valid for micro-waves such as sub-6GHz[1] and consists of a subcarrier independent large scale fading $\beta_{m,k}$ and an $N \times 1$ complex small scale fading $\mathbf{h}_{m,k}^{(c)}$ that is essentially both subcarrier and receive antenna dependent. The small scale fading coefficients $h_{m,k}^{(c)}[n]$ are independent and identically distributed (iid) zero-mean Gaussian random variables $[h_{m,k}^{(c)}]_n \sim \mathcal{N}(0,1)$ where $\mathrm{E}\left\{[h_{m,k}^{(c)}]_n[h_{m',k'}^{(c')}]_{n'}^*\right\} = \delta_{m,m'}\delta_{k,k'}\delta_{n,n'}\delta_{c,c'}$ with $\delta_{k,k'}$ being the delta-Dirac function centered at $k = k'$. Next and for the sake of notational simplicity, we will omit the subcarrier index, hence $\mathbf{g}_{m,k} = \sqrt{\beta_{m,k}}\mathbf{h}_{m,k}$. It is worth to mention that the small scale fading determines the channel coherence time, while the large scale fading is slowly varying and can only change per a number of coherence time intervals [18].

Because of the FH BW limitations and seeking higher SE per user, we propose different joint power control and AP scheduling strategies. In the AP scheduling we determine, depending on e.g. the traffic FH BW requirements, a mapping between each active user and its serving APs. More formally, we will denote the set of APs serving the kth user as $\boldsymbol{A}_k$.

B. Channel Estimation

The TDD frame structure consists of UL training and payload, a guard time, and DL training and payload of durations respectively τ_{up}, τ_{ud}, τ_{gi}, τ_{dp} and τ_{dd} where the channel coherence time T_{c} and the channel coherence bandwidth B_{c} are such that $\tau_{\text{c}} = B_{\text{c}}T_{\text{c}} = \tau_{\text{up}} + \tau_{\text{ud}} + \tau_{\text{gi}} + \tau_{\text{dp}} + \tau_{\text{dd}}$. With channel hardening in CF massive MIMO, the DL channel estimation is not needed and the user equipment (UE) can only rely on the knowledge of the average channel gain for data detection [8] so that the DL training can be removed ($\tau_{\text{dp}} = 0$).

For the purpose of the channel estimation, we use the orthogonal pilot sequence [20] wherein, during the UL

[1] Note that for mmwave spectrum, the statistics of the small scale fading is different from the microwave and strongly depends on the geometry. For instance across the receive antennas per AP and for a given user and subcarrier, these coefficients are not iid and the dependency is tied with the array geometry of the N receive antennas allowing e.g. to perform analog beamforming [19].

training, the active users synchronously send orthogonal pilot sequences $\sqrt{\tau_{\text{up}}}\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau_{\text{up}}\times 1}$ for $k = 1,\ldots,K$ with $\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'} = \delta_{k,k'}$. The underlying assumption is that $\tau_{\text{up}} \geq K$ to ensure the orthogonality of the pairwise pilot sequences. In e.g. an industrial set-up with UEs moving at a limited-speed and with the sub-6GHz transmission spectrum, large coherence time can be achieved and, consequently, a large number of orthogonal pilot sequences can be obtained. It can be easily shown that the minimum mean-square error channel estimate (see (3) from [11]) is unbiased and that the variance of the nth channel estimate $[\hat{\mathbf{g}}_{m,k}]_n$ for a given maximum transmit power per UE denoted ρ and a noise variance σ_v^2 is

$$\alpha_{m,k} = \mathrm{E}\left\{\left|[\hat{\mathbf{g}}_{m,k}]_n\right|^2\right\} = \frac{\tau_{\text{up}}\rho\beta_{m,k}^2}{\tau_{\text{up}}\rho\beta_{m,k}+\sigma_v^2}. \tag{1}$$

C. UL Payload Transmission

In the UL, the $N \times 1$ received signal at the mth AP ($\mathbf{r}_m$) is corrupted by an $N \times 1$ additive white Gaussian noise vector ($\mathbf{w}_m$) that is $\sim \mathcal{N}(0\cdot\mathbf{1}_N, \sigma_w^2\mathbf{I}_N)$ distributed. It can be formally expressed as

$$\mathbf{r}_m = \sqrt{\rho}\sum_{k=1}^{K}\sqrt{\eta_k}\mathbf{g}_{m,k}d_k + \mathbf{w}_m, \tag{2}$$

where d_k denotes the quadrature amplitude modulation data symbols from the kth user with $\mathrm{E}\{d_k d_{k'}^*\} = \delta_{k,k'}$, and where the power control coefficients $0 \leq \eta_k \leq 1\ \forall k$ enforce a maximum power constraint ρ per UE. The mth received signal $\mathbf{r}_m$ can further be written in a more compact form as

$$\mathbf{r}_m = \sqrt{\rho}\cdot\mathbf{G}_m\cdot\boldsymbol{\eta}\cdot\mathbf{d} + \mathbf{w}_m, \tag{3}$$

where we have defined the $N \times K$ channel matrix $\mathbf{G}_m = [\mathbf{g}_{m,1},\ldots,\mathbf{g}_{m,K}]$, the $K \times K$ diagonal matrix $\boldsymbol{\eta}$ with the kth diagonal entry being equal to η_k and the $K \times 1$ vector of data symbols $\mathbf{d} = [d_1,\ldots,d_K]^T$.

With the maximum ratio combining, each AP first partially equalizes the received signal based on the local channel estimates as $\hat{\mathbf{g}}_{m,k}^H\mathbf{r}_m\ \forall m$, with $(.)^H$ being the Hermitian transpose. The combiner at the CPU (see Fig. 1) sums up the different partially equalized signals as

$$y_k = \sum_{m=1}^{M}\hat{\mathbf{g}}_{m,k}^H\mathbf{r}_m = \sum_{m\in\boldsymbol{A}_k}\hat{\mathbf{g}}_{m,k}^H\mathbf{r}_m, \tag{4}$$

where following a certain AP scheduling policy, the combiner will only consider contributions from the scheduled APs serving the user in the UL denoted by $\boldsymbol{A}_k$. This can have significant saving in the FH BW provided that the cardinality of the union of the subsets $\{\boldsymbol{A}_k\}$ is minimized, which is the main research challenge tackled in this paper. We further expand the equalized signal (4) as

$$\begin{aligned} y_k = {} & \sqrt{\rho}d_k\sqrt{\eta_k}\sum_{m\in\boldsymbol{A}_k}\hat{\mathbf{g}}_{m,k}^H\mathbf{g}_{m,k} \\ & + \sqrt{\rho}\sum_{k'=1,k'\neq k}^{K}d_{k'}\sqrt{\eta_{k'}}\sum_{m\in\boldsymbol{A}_k}\hat{\mathbf{g}}_{m,k}^H\mathbf{g}_{m,k'} \\ & + \sum_{m\in\boldsymbol{A}_k}\hat{\mathbf{g}}_{m,k}^H\mathbf{w}_m. \end{aligned} \tag{5}$$

Note that the derivations in [11] for UL assume a single antenna ($N = 1$) per AP while the derivations in [13] were restricted to DL. For the UL system whose signal model is described by (5), it can be shown that the SNR admits a closed-form expression that resembles the representation in [11] up to a scaling factor (N) and where we have taken into account the subset $\boldsymbol{A}_k$ of the serving APs per UEs as

$$\gamma_k = \frac{N\rho\left(\sum_{m\in\boldsymbol{A}_k}\alpha_{m,k}\right)^2\eta_k}{\rho\sum_{k'=1}^{K}(\sum_{m\in\boldsymbol{A}_k}\alpha_{m,k}\beta_{m,k'})\eta_{k'} + \sum_{m\in\boldsymbol{A}_k}\sigma_w^2\alpha_{m,k}}. \tag{6}$$

III. Joint Access Points Scheduling and Power Allocation

A. Problem Formulation

For the sake of brevity, we consider the CF massive MIMO architecture depicted in Fig. 1, wherein the M APs are connected to one single CPU although the derivations in principles can be easily extended to multiple serial FH links each accommodating non-overlapping AP subsets. These APs need to be powered and their traffic needs to be fronthauled. For the selected split processing with distributed beamforming, the UL FH traffic scales with $M \cdot K$ and typically requires higher quantization, which is much more demanding than the DL transmission. Minimizing the UL traffic BW to be fronthauled over a BW limited serial FH link can be done by limiting the number of the reported equalized samples $\{\hat{\mathbf{g}}_{m,k}^H\mathbf{r}_m\}$ or more formally by minimizing $\sum_k|\boldsymbol{A}_k|$ with $|\boldsymbol{A}_k|$ denoting the cardinality of the subset $\boldsymbol{A}_k$, and/or using lower precision to represent the samples $\{\hat{\mathbf{g}}_{m,k}^H\mathbf{r}_m\}$. In this paper, we focus on the minimization of the term $\sum_k|\boldsymbol{A}_k|$ in UL to solve the FH BW limitations. Note that limiting the number of processed users per AP can bring some other benefits in terms of hardware usage such as reduced peak-to-average power ratio.

For the subsequent analysis, let us introduce the so-called $M \times K$ association matrix $\mathbf{C}$ where the (m,k)th entry of the matrix is one ($c_{m,k} = 1$) if the kth user is served by the mth AP and zero otherwise. The set of APs serving the kth user can then be expressed as (see also the illustration example in Fig. 2) $\boldsymbol{A}_k = \{m|\, c_{m,k} = 1\}$ with cardinality $|\boldsymbol{A}_k| = \sum_{m=1}^{M}c_{m,k}$. The FH BW constraint can be then rewritten as $\sum_{k=1}^{K}|\boldsymbol{A}_k| = \sum_{m=1}^{M}\sum_{k=1}^{K}c_{m,k} \leq \hat{M}$, where the maximum FH BW $\hat{M}(\leq MK)$ denotes the system parameter specifying the upper-bound on the maximum number of equalized samples that can be fronthauled in the UL. We next propose to solve the problem of joint power allocation and AP scheduling under such FH BW constraint. To do this, we first rewrite the SNR in (6) using the definitions provided for the association matrix $\mathbf{C}$ as

$$\gamma_k = \frac{N(\sum_{m=1}^{M}c_{m,k}\alpha_{m,k})^2\eta_k}{\sum_{k'=1}^{K}(\sum_{m=1}^{M}c_{m,k}\alpha_{m,k}\beta_{m,k'})\eta_{k'} + \frac{\sigma_w^2}{\rho}\sum_{m=1}^{M}c_{m,k}\alpha_{m,k}}, \tag{7}$$

where the summation over the subset $\boldsymbol{A}_k$ in (6) is replaced with the new definition of the association matrix.

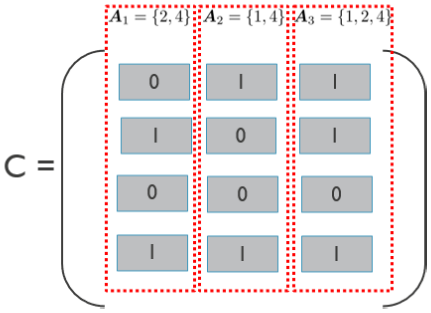

Fig. 2. Illustration of association matrix with $M = 4$ and $K = 3$.

B. Optimal Joint Power Allocation and AP Scheduling

The main challenge in maximizing the SNR in (7) is the dependency on both the parameters of the users and the allocated power to each user, which makes the whole problem non-concave. One way to facilitate the SNR maximization is to maximize the worst SNR as proposed in [11]. The resulting max-min optimization problem with the new FH constraint can be formulated as

$$\begin{aligned}
&\max_{\boldsymbol{\eta},\mathbf{C}} \min_{k} \gamma_k(\boldsymbol{\eta}, \boldsymbol{C})\\
&\text{st. } 0 \le \eta_k \le 1, \quad \forall k\\
&\sum_{m=1}^{M}\sum_{k=1}^{K} c_{m,k} \le \hat{M}\\
&c_{m,k} \in \{0,1\}, \quad \forall m,k, \qquad (8)
\end{aligned}$$

which can be equivalently rewritten using a newly defined slack variable t as

$$\begin{aligned}
&\max_{\boldsymbol{\eta},\mathbf{C},t} t\\
&\text{st. } \gamma_k(\boldsymbol{\eta}, \mathbf{C}) \ge t, \ \forall k\\
&\quad 0 \le \eta_k \le 1, \quad \forall k\\
&\sum_{m=1}^{M}\sum_{k=1}^{K} c_{m,k} \le \hat{M}\\
&c_{m,k} \in \{0,1\}, \quad \forall m,k. \qquad (9)
\end{aligned}$$

Given the existence of the entries of the association matrix that are binary variables, the above form can be presented as a mixed integer programming problem, and hence the problem admits a non-convex form [23]. Another source of non-convexity is due to the presence of unknown optimization variables $\{c_{m,k}\}$ and $\{\eta_k\}$ appearing as products in both the numerator and denominator of the SNR given by (7). The joint optimization of $\{\eta_k, c_{m,k}\, \forall m,k\}$ cannot be implemented in a complexity-efficient way at the first glance. Therefore**,** we propose, in the rest of the paper, to iteratively optimize the power $\boldsymbol{\eta}$ and the association matrix $\mathbf{C}$ as presented in Alg. 1, with the process being iterated until the convergence is attained. The proposed solutions are based on an iterative framework combined with convex relaxation that are standard and popular in signal processing community dealing with large-scale complex optimization problems.

It is worth to note that the general architecture with S serial FH wires with M_S APs and a maximum FH BW $\hat{M}_S$ per serial FH wire where $SM_S = M$ can be easily handled by the optimization problem (8) by enforcing S linear FH BW constraints rather than one constraint as $\sum_{m=1}^{M_S}\sum_{k=1}^{K} c_{m,k}^{(s)} \le \hat{M}_S$, which increases the complexity. On the other hand, with more FH wires and in the extreme case in P2P architecture, iso $M_S = 1$, the FH BW constraints become less restrictive as we will have less APs connected per FH wire.

Algorithm 1 Iterative Power Allocation and AP Scheduling.

```
Init: $\kappa = 0$, set $\mathbf{C} = \mathbf{C}^{(0)}$ and call power allocation: $\boldsymbol{\eta}^{(0)} = optimize(\mathbf{C}^{(0)})$
  1) AP scheduling using Alg. 2 or Alg. 3: $\mathbf{C}^{(\kappa+1)} = optimize(\boldsymbol{\eta}^{(\kappa)})$.
  2) Power allocation: $\boldsymbol{\eta}^{(\kappa+1)} = optimize(\mathbf{C}^{(\kappa+1)})$.
  3) Increase the super-iteration index: $\kappa \leftarrow \kappa + 1$.
  4) Repeat steps (1)-(3) until convergence.
```

C. Power Allocation Optimization

While the objective in (12) is linear, the first inequality constraint ($\gamma_k(\boldsymbol{\eta},\boldsymbol{C}) \ge t$) does not admit a linear form. In addition, the last binary constraint in (12) makes the whole optimization problem non-convex.

The first inequality constraint $\gamma_k(\boldsymbol{\eta},\mathbf{C}) \ge t\ \forall k$ can be now rewritten into the following form

$$\Big[\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\beta_{m,k} - \frac{N}{t}\Big(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\Big)^2\Big]\eta_k + \sum_{k'\neq k} \Big(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\beta_{m,k'}\Big)\eta_{k'} \le -\frac{\sigma_w^2}{\rho}\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}, \quad \forall k, \tag{10}$$

which forms a function with respect to the optimization variables $\{\eta_k\}$ and for a given $(t, \mathbf{C})$. In this case, the problem can be solved using the bisection [11] and through solving a sequence of linear feasibility problems

$$\begin{aligned}
&\max_{\boldsymbol{\eta}} 0\\
&\text{st. } \left[\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\beta_{m,k} - \frac{N}{t}\Big(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\Big)^2\right]\eta_k\\
&\quad + \sum_{k'\neq k}\Big(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\beta_{m,k'}\Big)\eta_{k'}\\
&\quad \le -\frac{\sigma_w^2}{\rho}\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}, \ \forall k\\
&\quad 0 \le \eta_k \le 1, \quad \forall k. \qquad (11)
\end{aligned}$$

D. Access Point Scheduling

Given the binary nature of all the entries of the association matrix $\mathbf{C}$, we propose next two algorithms that both rely on the convex relaxation of $\mathbf{C}$ to write the original optimization problem in a convex form. Hence, given a set of power control coefficients generated by a bisection sketched by (11), the proposed AP scheduling algorithms strive both to solve the following optimization problem

$$\begin{aligned}&\max_{\mathbf{C},t} t\\&\text{st.}\gamma_k(\boldsymbol{\eta},\mathbf{C}) \geq t, \quad \forall k\\&\quad\sum_{m=1}^{M}\sum_{k=1}^{K} c_{m,k} \leq \hat{M}\\&\quad 0 \leq c_{m,k} \leq 1, \quad \forall m,k. \end{aligned}\tag{12}$$

Knowing that the denominator of the SNR can be re-written

$$\begin{aligned}&\rho\sum_{k'=1}^{K}\eta_{k'}(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\beta_{m,k'}) + \sigma_w^2\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}\\&\qquad = \rho\sum_{m=1}^{M} c_{m,k}\alpha_{m,k}(\sum_{k'=1}^{K}\beta_{m,k'}\eta_{k'} + \frac{\sigma_w^2}{\rho}),\end{aligned}\tag{13}$$

the inequality $\gamma_k(\boldsymbol{\eta},\mathbf{C}) \geq t$ can be rewritten in a more suitable form as $\forall k$

$$\sum_{m=1}^{M}(\alpha_{m,k}[\chi_m(\boldsymbol{\eta})+\frac{\sigma_w^2}{\rho}])c_{m,k} \leq \frac{N}{t}(\sum_{m=1}^{M}\alpha_{m,k}c_{m,k})^2\eta_k,\tag{14}$$

where we have defined the sum power received by the mth AP as follows

$$\chi_m(\boldsymbol{\eta}) \doteq \sum_{k=1}^{K}\beta_{m,k}\eta_k.\tag{15}$$

Noting that $\eta_k = 0$ is irrelevant as it means that the kth user will get no rate, the set of variables $\{c_{m,k}\}$ satisfying the inequality (14) do not form a convex set. Next, we propose different AP scheduling algorithms of the problem (12) for a given power distribution $\boldsymbol{\eta}$.

1) Iterative AP Balancing: Given the complexity of the MK-multidimensional optimization in (12), we propose to decouple the problem across the users by each time in an inner loop optimizing for one user k given the other users ($k' \neq k$) settings, and repeat the cycling across the users in an outer loop until convergence. Denoting the kth column of the association matrix as $\mathbf{c}_k$, the optimization problem solved in the inner loop writes

$$\begin{aligned}&\max_{\mathbf{c}_k,t} t\\&\text{st.}\gamma_{\text{u},k}(\boldsymbol{\eta},\mathbf{C}) \geq t, \quad \forall k\\&\quad\sum_{m=1}^{M} c_{m,k} \leq \hat{M} - \sum_{m=1}^{M}\sum_{k'\neq k} c_{m,k'}, \quad \forall k\\&\quad 0 \leq c_{m,k} \leq 1, \quad \forall m.\end{aligned}\tag{16}$$

This class of algorithms will be referred to as Iterative AP Balancing (IAPB). In essence, it is similar to the iterative spectrum balancing [21] and the space alternating generalized expectation maximization [22] algorithms with both having good convergence properties.

2) AP Scheduling Based on χ-Algorithm: To derive a convex form of the problem (12), we propose to replace the source of non-convexity, which are the coefficients $\chi_m(\boldsymbol{\eta}) = \sum_{k=1}^{K}\beta_{m,k}\eta_k$ in the inequality (14), by an upper bound as $\chi_m(\boldsymbol{\eta}) = \sum_{k=1}^{K}\beta_{m,k}\eta_k \leq \hat{\chi}(\boldsymbol{\eta})$ where the upper bound $\hat{\chi}(\boldsymbol{\eta})$ is a design parameter that is selected to be AP independent (for instance $\hat{\chi} = \max_m\{\chi_m(\boldsymbol{\eta})\}$). With the proposed upper bound, it can be easily shown that the set of variables $\{c_{m,k}\}$ satisfying the following inequality

$$\begin{aligned}&\frac{\sigma_w^2}{\rho}\sum_{m=1}^{M}\alpha_{m,k}c_{m,k} + \hat{\chi}(\boldsymbol{\eta})\sum_{m=1}^{M}\alpha_{m,k}c_{m,k}\\&\qquad\qquad \leq \frac{N}{t}(\sum_{m=1}^{M}\alpha_{m,k}c_{m,k})^2\eta_k, \forall k,\end{aligned}\tag{17}$$

also satisfies the inequality (14) although the opposite may not necessary hold. In other words, we have squeezed the feasible set of variables to make the original problem convex and, therefore, the parameter $\hat{\chi}$ should be carefully selected to avoid excessive feasible set reduction. The inequality (17) can be reformulated as $\forall k$

$$(\sum_{m=1}^{M}\alpha_{m,k}c_{m,k})\cdot(\hat{\chi}(\boldsymbol{\eta})+\frac{\sigma_w^2}{\rho} - \frac{N}{t}[\sum_{m=1}^{M}\alpha_{m,k}c_{m,k}]\eta_k) \leq 0.\tag{18}$$

As $\alpha_{m,k} > 0$ and $c_{m,k} \geq 0$, $\forall m,k$, the above inequality may be satisfied by considering either, (i) $\sum_{m=1}^{M}\alpha_{m,k}c_{m,k} = 0$, or (ii) $\hat{\chi}(\boldsymbol{\eta}) + \frac{\sigma_w^2}{\rho} - \frac{N}{t}(\sum_{m=1}^{M}\alpha_{m,k}c_{m,k})\eta_k \leq 0$. The solution for the former case is $c_{m,k} = 0, \forall m,k$, which is a trivial solution, while the latter case results in the following linear constraint for each user k

$$-\sum_{m=1}^{M_S}\alpha_{m,k}c_{m,k} \leq \frac{1}{\eta_k}(-\frac{\sigma_w^2}{\rho} - \hat{\chi}(\boldsymbol{\eta}))\frac{t}{N}.\tag{19}$$

The bisection can then be used to solve a sequence of linear feasibility problems as

$$\begin{aligned}&\max_{\mathbf{C}} 0\\&\text{st.} -\sum_{m=1}^{M}\alpha_{m,k}c_{m,k} \leq \frac{1}{\eta_k}(-\frac{\sigma_w^2}{\rho} - \hat{\chi}(\boldsymbol{\eta}))\cdot\frac{t}{N}, \quad \forall k\\&\quad\sum_{m=1}^{M}\sum_{k=1}^{K} c_{m,k} \leq \hat{M}\\&\quad 0 \leq c_{m,k} \leq 1, \quad \forall m,k.\end{aligned}\tag{20}$$

The approximation made in χ-algorithm may degrade the performance of the optimization algorithm, especially, when there is a large spread in the large scale fading $\beta_{m,k}$ across the APs for a given UE. In the case of collocated massive MIMO with a shadowing effect that is only AP-dependent, $\beta_k = \beta_{m,k}$, $\forall m$, the choice of $\chi(\boldsymbol{\eta})$ becomes relatively easy $\sum_{k'=1}^{K}\beta_{m,k'}\eta_{k'} = \sum_{k'=1}^{K}\beta_{k'}\eta_{k'} = \chi(\boldsymbol{\eta})$. In this case, no additional suboptimality is introduced when simplifying the problem from (14) to (19).

Further simplifications are possible by cycling across the users and, then, optimizing for one column of the association matrix given the other columns, in which case the algorithm will be referred to as iterative χ balancing (IχB) that pertains to the IAPB class introduced in subsection III-D.1. The kth optimization problem can be therefore written as follows:

$$
\begin{aligned}
&\max_{\mathbf{c}_k} 0\\
&\text{st.} -\sum_{m=1}^{M}\alpha_{m,k}c_{m,k} \le \frac{1}{\eta_k}\left(-\frac{\sigma_w^2}{\rho}-\hat{\chi}(\boldsymbol{\eta})\right)\cdot\frac{t}{N}, \quad \forall k\\
&\sum_{m=1}^{M}c_{m,k} \le \hat{M}-\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}, \quad \forall k\\
&0 \le c_{m,k} \le 1, \quad \forall m,k. \qquad (21)
\end{aligned}
$$

The proposed AP scheduling algorithm referred to as IχB is summarized in Alg. 2.

Algorithm 2 Bisection for IχB-based AP Scheduling Algorithm.

```
Init: $\iota = 0$, set a tolerance $\epsilon$, $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$, $\mathbf{C} = \mathbf{C}^{(0)}$,
  1) Cycle across UEs: for UE $k$
      a) select feasible $t_{min}$ and $t_{max} > t_{min}$
      b) Set $t = \frac{t_{min}+t_{max}}{2}$, solve the convex feasibility problem (21)
      c) If problem feasible, then $t_{min} = t$ else $t_{max} = t$
      d) If $t_{max} - t_{min} < \epsilon$ go to the next UE, otherwise go back to (1.b)
  2) $\iota \leftarrow \iota + 1$
  3) If convergence go to (4) otherwise go back to (1)
  4) Round the association matrix: $\mathbf{C} \leftarrow \lceil \mathbf{C} \rfloor$ and stop the optimization
```

3) AP Scheduling Driven by Channel Hardening: Motivated by the fact that the non-convex set in (14) appears due to the presence of a quadratic term in the numerator of the SNR, we aim at minimizing the interference appearing in the denominator of the SNR function subject to a lower bound $\hat{D}_k(\boldsymbol{\eta})$ on the effective hardened channel in the numerator of the SNR. The newly-formed optimization problem can be accordingly written as

$$
\begin{aligned}
&\min_{\mathbf{C},k}\mathcal{I}_k(\mathbf{C}) = \sum_{m=1}^{M}(\chi_m(\boldsymbol{\eta})+\frac{\sigma_w^2}{\rho})\alpha_{m,k}c_{m,k}\\
&\text{st. } \mathcal{D}_k(\mathbf{C}) = \sum_{m=1}^{M}c_{m,k}\alpha_{m,k} \ge \hat{D}_k(\boldsymbol{\eta}), \ \forall k\\
&\sum_{m=1}^{M}\sum_{k=1}^{K}c_{m,k} \le \hat{M}\\
&0 \le c_{m,k} \le 1, \quad \forall m,k. \qquad (22)
\end{aligned}
$$

Note that the objective function can be selected using other aggregation functions such as a weighted sum of the per-UE interference as $\mathcal{I}(\mathbf{C}) = \sum_k w_k\mathcal{I}_k(\mathbf{C})$. To further decouple the KM-multidimensional optimization to K independent M-multidimensional optimization problems, an Iterative Channel Hardening Balancing (IHB) procedure is proposed that belongs to the class of IAPB from section III-D.1. The new IHB algorithm operates by iterating over the users each time focusing on optimizing for one user (say the kth column $\mathbf{c}_k$ of the association matrix) while fixing the remaining columns as

$$
\begin{aligned}
&\min_{\mathbf{c}_k}\mathcal{I}_k(\mathbf{C}) = \sum_{m=1}^{M}(\chi_m(\boldsymbol{\eta})+\frac{\sigma_w^2}{\rho})c_{m,k}\alpha_{m,k}\\
&\text{st.} \sum_{m=1}^{M}c_{m,k}\alpha_{m,k} \ge \hat{D}_k(\boldsymbol{\eta}), \quad \forall k\\
&\sum_{m=1}^{M}c_{m,k} \le \hat{M}-\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}, \quad \forall k\\
&0 \le c_{m,k} \le 1, \quad \forall m,k. \qquad (23)
\end{aligned}
$$

Algorithm 3 IHB-based AP Scheduling Algorithm.

```
Init: $\iota = 0$, set a tolerance $\epsilon$, $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$, $\mathbf{C} = \mathbf{C}^{(0)}$, $\hat{D}_k^{(0)} \le \sum_{m=1}^{M}[c_{m,k}]^{(0)}\alpha_{m,k}\ \forall k$
  1) $\hat{D}_k = \hat{D}_k^{(0)}\ \forall k$
  2) Cycle across users: for each user $k$
      a) Solve the convex problem (23)
      b) If problem feasible, increase $\hat{D}_k$: $\hat{D}_k \leftarrow \hat{D}_k + \Delta_D$ then go to (2.a)
      c) If problem not feasible then go to the next user
  3) If $\sum_k \left|\hat{D}_k - \hat{D}_k^{(0)}\right| > \epsilon$, $\iota \leftarrow \iota + 1$ and go back to (2) otherwise go to (4)
  4) Round the association matrix: $\mathbf{C} \leftarrow \lceil \mathbf{C} \rfloor$ and stop the optimization
```

The algorithm is summarized in Alg. 3 where the steps to tune the lower-bound $\hat{D}_k$ are also discussed in further details. The optimization of kth column of the association matrix is repeated until there is no possible increase of the hardened channel magnitude, in which case the algorithm processes the next column in the pipeline. Since the utility of interest is the received SNR, there is no guarantee that the obtained feasible solution through Alg. 3 improves the SNR in subsequent iterations, at least for small MIMO systems. One can accept the solution if the resulting SNR proportional to $\hat{D}_k(\boldsymbol{\eta})^2/\mathcal{I}_k(\mathbf{C})$ is also improved. In case of SNR degradation, the solution will be rejected before moving to the next user. We have noticed that for reasonable MIMO and certainly massive MIMO, the final solution always improves the SNR (see the discussion on the convergence analysis in section IV).

The optimization framework formed for both the IχB and the IHB algorithms is applied over a certain column of the association matrix, and is affected by the column distribution of the other users, which can be also seen in the inequality $\sum_{m=1}^{M}c_{m,k} \le \hat{M}-\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}$. In other words, if the association entries corresponding to all columns other than the kth one are determined such that the term $\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}$ admits a large value, then the search space for the kth column is tightened. This is because $\sum_{m=1}^{M}c_{m,k}$ is upper bounded by $\hat{M}-\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}$ that significantly shrinks when $\hat{M}-\sum_{m=1}^{M}\sum_{k'\neq k}c_{m,k'}$ goes lower. To decouple the optimization framework across all the users, we propose to upper-bound the association budget per user as $\sum_{m=1}^{M} c_{m,k} \leq \hat{M}_k$ where $\hat{M} = \sum_k \hat{M}_k$. This will then guarantee that the constraint $\sum_{m=1}^{M}\sum_{k=1}^{K} c_{m,k} \leq \hat{M}$ is always fulfilled. There are two main benefits associated with the proposed framework. First, we have K independent optimization problems that could be solved independently. In addition, given $\{\hat{M}_k\}$, no user will be penalized by the AP-UE association of the other users. This enables to determine the association matrix in a more computationally efficient way.

4) Large-Scale-Based AP Scheduling Algorithms: The proposed IHB depends on the initial distribution of the association matrix. Indeed, the optimized lower bound of the hardened channel $\hat{D}_k(\boldsymbol{\eta})$ is upper bounded by $\hat{D}_k^{(\sigma)}$. As a matter of fact, if the initial guess is such that $\hat{D}_k(\boldsymbol{\eta}) = \hat{D}_k^{(\sigma)}$, then the IHB does not proceed in the subsequent iterations. Motivated by this reasoning and driven by further simplifications, we propose a one-shot AP scheduling algorithm based on the large-scale fading in the same way as proposed in [13], however, for the UL and with the maximum FH budget per UE being taken into consideration. This is further motivated by the form appearing in the numerator of (7) (or the effective hardened channel of the kth user) that is proportional to $(\sum_{m=1}^{M} c_{m,k}\alpha_{m,k})^2$. Assuming sufficiently accurate channel estimates, we can safely assume $\alpha_{m,k} \simeq \beta_{m,k}$, otherwise one can use the variance $\{\alpha_{m,k}\}$ rather than $\{\beta_{m,k}\}$. The large scale fadings are sorted in a descending way as $\beta_{\sigma(1),k} \geq \beta_{\sigma(2),k} \cdots \geq \beta_{\sigma(M),k}$ and the largest $\hat{M}_k$ coefficients determine the scheduled APs of the kth user as $c_{\sigma(m),k} = 1$ for $m \leq \hat{M}_k$ and $c_{\sigma(m),k} = 0$ otherwise. In this case, the upper bound $\hat{D}_k^{(\sigma)} = \sum_{m=1}^{\hat{M}_k} \alpha_{\sigma(m),k}$ is achievable and the AP scheduling is simplified and is decoupled from the power allocation. The proposed algorithm will be referred to as the Largest Large-Scale Fading (LLSF) algorithm.

IV. Convergence Analysis

The iterative nature of the algorithms presented by Alg. 2 and Alg. 3 along with the intermediate rounding stages to produce binary association entries may result in undesired responses such as oscillation or divergence of the objective function in subsequent iterations. First, we sketch the proof of convergence by presenting the following propositions:

Proposition 1: Let $f(x,y)$ denote an objective function with the set of constraints $g_i(x,y) \leq 0$. Assume $\{x_n, y_n\}$ as a set of candidate optimal solutions obtained by an iterative optimization algorithm at the nth iteration. Supposing the marginal objective and constraints ($f(x,y_n)$, $f(x_n,y), g_i(x,y_n)$, $g_i(x_n,y)$) are convex for all the values of x_n and y_n, respectively, the iterative optimization algorithm guarantees that the objective function monotonically decreases $f(x_{n+1},y_{n+1}) \leq f(x_n,y_n)$ while all the constraints are satisfied and the solution is feasible.

Proof: Let $\{x_n, y_n\}$ be the optimal solution at the nth iteration. Given the convexity of $f(x,y_n)$, the solution $x_{n+1} = \arg\{\min_x f(x,y_n)\}$ can be obtained by solving a convex optimization problem as $\min_x f(x,y_n)$ st. $g_i(x,y_n) \leq 0, \forall i$. The Karush–Kuhn–Tucker (KKT) condition [23] states that the optimal solution meets both the feasibility as well as the inequality $f(x_{n+1},y_n) \leq f(x,y_n)$, $\forall x$ that leads to $f(x_{n+1},y_n) \leq f(x_n,y_n)$ considering $x = x_n$. It is also known from the assumptions that $f(x_{n+1},y)$ is convex and, therefore, admits an optimal solution y_{n+1} that results in the inequality $f(x_{n+1},y_{n+1}) \leq f(x_{n+1},y_n)$. It is then known that $f(x_{n+1},y_{n+1}) \leq f(x_n,y_n)$ for the feasible solution $\{x_n,y_n\}$ and the conclusion can be generalized to all the values of n by induction. ■

Proposition 2: For the objective function $f(x,y)$ and a given optimal solution $\{x_n,y_n\}$ at the nth step of the iterative optimization algorithm, let $\bar{g}_i(x,y_n)$ correspond to the convex-relaxed representation of the original constraints $g_i(x,y_n)$. Assume $\{x_{n+1},y_n\}$ as the optimal solution of the following problem:

$$\begin{aligned}&\min_x f(x,y_n).\\ &st\ \bar{g}_i(x,y_n) \leq 0, \quad \forall i.\end{aligned} \tag{24}$$

Considering $f(x_{n+1},y)$ and $g_i(x_{n+1},y)$ being both convex for all the values of x_{n+1}, the objective function is monotonically decreasing if the solution $\{x_{n+1},y_n\}$ is feasible for $g_i(x,y_n), \forall i$.

Proof: Considering $\{x_n,y_n\}$as the solution obtained at the nth iteration, and given the convexity of the optimization problem in (24), the KKT condition can be adopted to obtain $f(x_{n+1},y_n) \leq f(x_n,y_n)$. As the solution $\{x_{n+1},y_n\}$ also holds for $g_i(x,y_n)$ $\forall i$, the inequality $f(x_{n+1},y_n) \leq f(x_n,y_n)$ is still valid as the solution is feasible for the main optimization problem. Since $g_i(x_{n+1},y)$ is convex, the optimal solution $\{x_{n+1},y_{n+1}\}$ satisfies the inequality $f(x_{n+1},y_{n+1}) \leq f(x_{n+1},y_n)$. Hence the objective function monotonically decreases over subsequent iterations following the procedure above. ■

Using the outcomes of propositions (1) and (2), the following lemma sketches the convergence proof for algorithms (2) and (3).

Lemma 1: The iterative optimization frameworks given by algorithms (2) and (3) provide monotonically decreasing/increasing objectives, respectively, with the set of obtained solutions being feasible for the main optimization problem in (9).

Proof: We first show the objective function monotonically increases over subsequent iterations. Both problems posed by Alg. (2) and Alg. (3) can be re-structured to the general form of (24) assuming $x = \mathbf{C}$ and $y_n = \boldsymbol{\eta}_n$. Also, all the constraints admit a linear form except for the binary equality applied to the association entries. Therefore, the relaxed optimization problem follows the form in (24) admitting $\bar{g}_i$ as the relaxed constraint on the association matrix. As the relaxed form is written as a linear program and the marginal problem is also convex for the optimal power allocation problem ($f(\mathbf{C}_n,\boldsymbol{\eta})$), Proposition 1 states that the objective (upper bound in bi-section) monotonically increases, or $f(\mathbf{C}_{n+1},\boldsymbol{\eta}_{n+1}) \geq f(\mathbf{C}_n,\boldsymbol{\eta}_n)$. Also, it is known that the solution of a linear program lies on the vertices of the boundary points, which indicates the optimal solution to (2) and (3) satisfies $c_{m,k} \leq \epsilon$ or $c_{m,k} \geq 1-\epsilon$, $\forall m,k$ and with ϵ being a small positive number. Indeed, the solution of the problems (20) and (3) is feasible for the original (without relaxation) problem. Then, Proposition 2 indicates that the objective monotonically increases for the new feasible solution $\{\mathbf{C}_{n+1}, \boldsymbol{\eta}_n\}$. As $\mathbf{C}_{n+1}$ meets the constraints in both (2) and (3), and the feasibility of $\boldsymbol{\eta}_n$ has been already maintained in the power allocation phase, the new solution $\{\mathbf{C}_{n+1}, \boldsymbol{\eta}_n\}$ is feasible for the original problem (9). ■

It has been so far shown that the joint power/association matrix optimization using (2) and (3) results in a monotonically increasing/decreasing objective solution, respectively, with the solution being also feasible for (9). We also need to prove that the objective in (9) improves over subsequent iterations. The analysis is done for any of the given algorithms separately:

IHB Convergence: Alg. 3 aims at minimizing the interference (the denominator in the objective of (9)). The first constraint in Alg. 3 followed by the boundary increase in step (2.b) from Alg. 3 ensures that the numerator of the objective function in (9) grows over subsequent iterations, since the interference optimization is done over all the users. Nevertheless, the interference may also go up compared to the previous iteration leading to compensate the increase in the numerator of (9). Now, considering large values of M ($M \to \infty$), it can be shown that the power allocation problem results in an almost-uniform solution for the powers ($\eta_k = 1$). Also, given the second constraint in Alg. 3 and for a sufficiently large M, there are sufficiently large number of non-zero association entries in the optimal solution. The objective in (9) now admits the asymptotic form $\gamma_k \xrightarrow[M\to\infty]{} \frac{O(M^2)}{O(M)}$, which indicates the numerator grows at a much higher rate than the interference. This also implies that the whole problem becomes interference-free for a large value of M, which is in-line with the proof given by [11]. As it has been already shown that the numerator of objective in (9) grows at a much higher rate than the interference, when M admits a large value, increasing the boundary $\hat{D}_k$ in step (2.b) of Alg. 3 ensures that the objective function grows over different iterations. Therefore, while the procedures in Alg. 3 provide a feasible solution for the original problem of (9), the interference asymptotically converges to zero with the numerator of the objective growing over iterations. This implies that the iterative solution of Alg. 3 provides a feasible solution to (9) that ensures the SNR grows at a rate proportional to M.

IχB Convergence: Alg. 2 aims at solving the original optimization in (2) using bi-section where the term $\chi_m(\eta)$ is replaced with its upper bound $\hat{\chi}$. As a result, it can be inferred that the interference (denominator of the objective in (9)) becomes larger as $\sum_m c_{m,k}\alpha_{m,k}\chi_m(\eta) \leq \sum_m c_{m,k}\alpha_{m,k}\hat{\chi}$. Indeed, Alg. 2 intends to maximize a lower bound on the objective of (2). Now, lemma 1 states that the algorithm provides a feasible solution that guarantees the objective is monotonically increasing. Given the fact that the objective of (9) always exceeds the solution of Alg. 2, the SNR in (9) also increases over iterations. Note that the maximum user-to-AP assignment constraint given by $\hat{M}$ in Alg. 2 controls the best achievable SNR and, therefore, the iterative algorithm converges to a maximum bound after some iterations.

V. Complexity Analysis

The time complexity of the main iterative algorithm proposed to tackle (8) is studied by investigating the computation time for any of the power allocation and AP-scheduling stages, separately.

Proposition 3: For the AP-scheduling problem formulated by Alg. 2 and Alg. 3, the time-complexity admits the form $K\hat{T}_{AP}O_{LP}\left(f_{AP}(M)\right)$ with $\hat{T}_{AP}$ being the maximum number of iterations before the convergence, O_{LP} as the time complexity of the linear programming, and $f_{AP}(M)$ being a certain function of M.

Given the above proposition and considering the power allocation procedure described by (11), the overall complexity of the iterative solution as described by Alg. 1 can be obtained by the following proposition:

Proposition 4: The time complexity of the iterative power allocation and AP-scheduling for (8) is $\kappa_{max}\left[\hat{T}_{PWC}O_{LP}\left(f_{PWC}(K)\right) + K\hat{T}_{AP}O_{LP}\left(f_{AP}(M)\right)\right]$ where κ_{max} denotes the maximum number of super-iterations in Alg. 1 before the convergence, $\hat{T}_{PWC}$ being the maximum number of iterations before the convergence of power allocation, and $f_{PWC}(K)$ is a certain function of the number of users K.

It was shown by Lemma 1 that the iterative algorithms provide monotonically decreasing/increasing objectives, which implies both algorithms 2 and 3 converge within a certain number of iterations κ_{max} that is typically very small (see for instance later in this paper the outage probability results reported in Fig. 10). The time complexity of linear programming (O_{LP}) is specified depending on the type of the algorithm employed to tackle the LP problem. While the simplex method gives an exponential time-complexity at the worst-case and linear as the average, the solvers that use interior-point method as the base optimization framework show a variety of time-complexities depending on the underlying methodology to solve the problem. The theoretical complexity of the interior point methods has been lowered to $O(n)$ (the reader is referred to [25] for more details).

Considering the interior-point method as the base backbone to tackle the linear programming, the following lemma summarizes the computational complexity of the proposed algorithm in this paper:

Lemma 2: The time complexity of the iterative power allocation and AP-scheduling for (8) with a back-end interior-point method for LP optimization is $\kappa_{max}\left[\hat{T}_{PWC}O\left(K\right) + K\hat{T}_{AP}O\left(M\right)\right]$.

It is now known from Lemma 2 that the iterative algorithm scales linearly with both the number of users and APs. To verify the finding of Lemma 2, an empirical study is conducted to analyze the complexity with respect to both the number of users (K) and access points (M), and the results are reported in Fig. 3 for different combinations of M and K. For any of the proposed algorithms in this paper, we do the following two steps:

- Fixed M: we choose $M = 128$ and run the algorithm for different values of $K \in \{2^m, m \in 2, \ldots, 6\}$. Results of 100 Monte-Carlo runs are collected for each value of

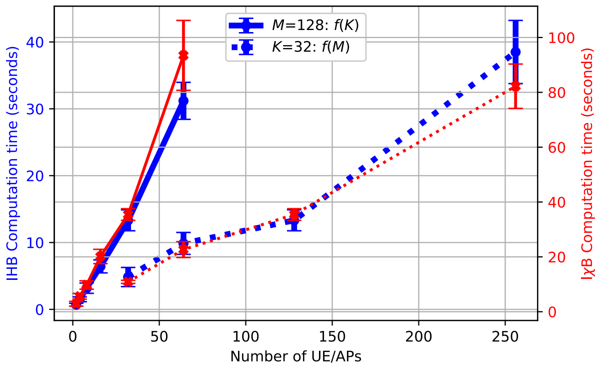

Fig. 3. Mean and standard deviation of the computation time for $\hat{M} = 60\%$ and different combinations of M and K.

m, and the average and standard-deviation for the total computation-time are calculated.

- Fixed K: similar to the previous case, the value of $K = 32$ is chosen and the experiment is done for $M \in \{2^m, m \in 5, \ldots, 8\}$.

All the results are produced by running the experiments on the same Intel Core i5-8350U CPU with the clock-frequency of 1.7 GHz. The figure indicates that the computational time scales linearly with both the number of UEs and APs. While the linear scalability is observed in both IχB and IHB algorithms, the IHB algorithm shows lower average timing compared to the other framework. We will also show in section VI-E how the aforementioned linear scalability may outperform state-of-art algorithms for solving general non-linear mixed-integer programming where the computational complexity is polynomial.

VI. Numerical Results

We evaluate the performance of the proposed algorithms on a simulated scenario. The scenario setup is first described and the performance of any of the proposed algorithms in this paper is analyzed under this scenario. We also assess the convergence of the algorithms that are further compared with the one shot LLSF heuristic and the exhaustive NMIP solver. The maximum FH budget $\hat{M}$ is defined as a fraction of MK: $\hat{M} = x\%$ means that $\hat{M} = {}^{x}\!/_{100} \cdot MK$.

A. Setup

Similar to [5], we consider a piazza topology wherein the APs are placed along the perimeter of DxD square as depicted in Fig. 4. The large scale fading coefficient $\beta_{m,k}$ depends on the path-loss $Ploss_{m,k}$ and the shadow fading $SF_{m,k}$ as $\beta_{m,k} = Ploss_{m,k} \times SF_{m,k}$. For the path-loss model, we use the same three slope model [11] that depends on the distance $d_{m,k}$ between the mth AP and kth UE as

$$(Ploss_{m,k})_{\text{dB}} = -L_{cte} \begin{cases} -35\log_{10}(d_{m,k}), & \text{if } d_{m,k} > d_1 \\ -15\log_{10}(d_1) - 20\log_{10}(d_{m,k}), & \text{if } d_0 < d_{m,k} \leq d_1 \\ -15\log_{10}(d_1) - 20\log_{10}(d_0), & \text{if } d_{m,k} \leq d_0, \end{cases} \tag{25}$$

where the constant L_{cte} depends on the carrier frequency f_c (in MHz), the AP antenna height h_{AP} (in m) and the UE antenna height h_{UE} (in m) as

$$L_{cte} = 46.3 + 33.9 \cdot \log_{10}(f_c) - 13.82 \cdot \log_{10}(h_{AP}) - (1.1 \cdot \log_{10}(f_c) - 0.7) \cdot h_{UE} + (1.56 \cdot \log_{10}(f_c) - 0.8)\,.$$

We assume uncorrelated shadow fading with log-normal distribution i. e. $(SF_{m,k})_{\text{dB}} \sim \mathcal{N}(0, \sigma_{\text{sh}}^2)$. For the chosen parameters $f_c = 1.9$ GHz, $h_{AP} = 10$ m, and $h_{UE} = 1.65$ m, $L_{cte} \simeq 143.148$ dB. The noise power is given by $\sigma_w^2 = B \cdot k_B \cdot T_0 \cdot n_F$ where the Boltzmann constant $k_B = 1.381 \cdot 10^{-23}$ J/K, the noise temperature $T_0 = 290$ K and $n_F = 9$ dB is the noise figure. The UL throughput of the kth user, taking the TDD frame structure into account, writes $S_k = B\frac{\tau_{\text{ud}}}{\tau_c}\log_2(1+\gamma_k)$ where the coherence time $T_c = 1$ ms, the coherence bandwidth $B_c = 200$ kHz, and the TDD frame length is $\tau_c = B_c T_c = 200$ samples. We consider a scenario with $K = 20$ UEs and $M = 100$ APs equipped with $N = 1$ antenna each. With $\tau_{\text{up}} = 20$ samples for UL training and 50% UL and DL split ratio, $\tau_{\text{ud}} = 90$ samples are dedicated for UL payload transmission with a maximum transmit power $\rho = 100$ mWatts.

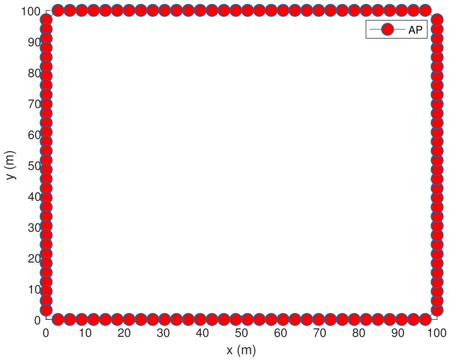

Fig. 4. The Piazza topology.

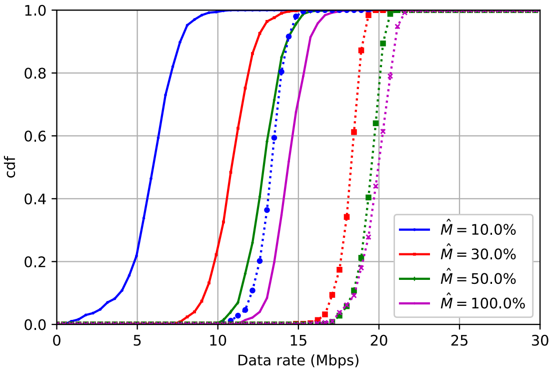

Fig. 5. IHB data rate cdf for $10 \leq \hat{M} \leq 100\%$: thick (resp. dashed) line is for Full (resp. Optimized) power.

B. Performance Comparison of the IHB and the χ-Algorithms

We show in this subsection the resulting data rate performance improvement of the proposed iterative algorithms after the convergence. The cumulative distribution function (cdf) of the data rate for different maximum budgets $\hat{M}$ ($10 \leq \hat{M} \leq 100\%$) is shown in Fig. 5 for IHB algorithm where the results

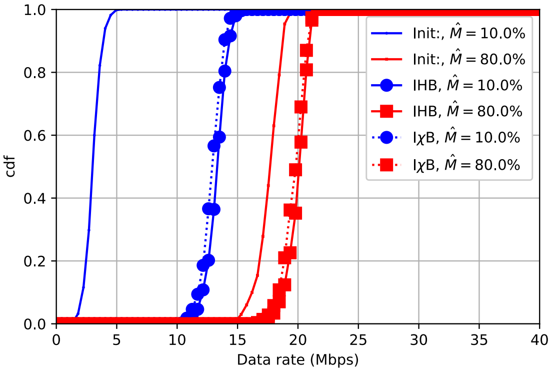

Fig. 6. IHB and IχB data rate cdf for different super iterations.

are depicted for no power-control referred to as '*Full power*' ($\eta_k = 1$) and optimal power control referred to as '*Optimized power*'. As a matter of a fact, using power control solutions resulting from the optimization algorithm, yields significant performance increase against full power transmission for a given optimized association matrix (similar behavior has been observed for IχB). It is also observed that the amount of data-rate improvement is boosted when the maximum FH budget $\hat{M}$ decreases. For instance, for $\hat{M} = 10\%$ (resp. $\hat{M} = 80\%$) of the overall maximum budget MK, the outage data rate improvement is roughly more than 150% (resp. 45%) for target outage probabilities $p_{outage} \leq 20\%$.

However, optimizing the power alone does not maintain the best performance and the appropriate association matrix configuration is required to provide the best throughput as suggested by Fig. 6. The legend '*Init*' in Fig. 6 refers to the performance resulting from the non-optimized association matrix, however, with the optimal power distribution. The figure clearly indicates that the iteration between power control and AP scheduling is required to improve the performance, especially, with stringent requirements on the FH BW. For instance, comparing results obtained for $\hat{M} = 10\%$ versus $\hat{M} = 80\%$, both IHB and IχB yield about 400% outage data rate increase for outage probabilities less than $p_{outage} \leq 20\%$ with $\hat{M} = 10\%$. This gain on the other hand decreases as the upper bound on FH BW $\hat{M}$ increases. As an example, with $\hat{M} = 80\%$, Fig. 6 shows less than 20% outage data rate improvement for outage probabilities $p_{outage} \leq 20\%$.

Based on these results, the joint AP scheduling and power control is mandatory and the impact of the joint optimization becomes more evident as the FH BW goes down. Furthermore, using the appropriate AP scheduling can overweight the use of suboptimal power control, a fact that can be inferred when comparing e.g. the curves labeled '*Init, $\hat{M} = 10\%$*' and '*Full power, $\hat{M} = 10\%$*' from respectively Figs. 6 and 5.

It is also essential to visualize how many UEs are served by each AP in the final solution as this will indicate the uniformity of the load distribution among the APs. This can be quantified by computing the cdf of the number of served UEs per AP after convergence with the results being sketched in Fig. 7 for both IHB and IχB algorithms. Results in Fig. 7 show that the proposed algorithms tend to uniformly allocate the UEs across the APs, especially, for the IHB algorithm and at higher values of $\hat{M}$. Note that the aforementioned observation is expected as this '*natural*' behavior allows each UE to use more APs hence increased diversity, especially, when the upper-bound $\hat{M}$ increases. The gap between the two algorithms becomes wider for $\hat{M} = 80\%$ where IHB is outperforming the IχB. For instance, the probability of more than 9 served UEs per AP is respectively 90% and 100% for the IHB and the IχB algorithms. Such an improvement becomes tighter for smaller values of $\hat{M}$. For instance for $\hat{M} = 30\%$, the probability of more than 9 served UEs per AP drops to roughly 2% and 5% for, respectively, IHB and IχB. Although the performance improvement degrades by decreasing $\hat{M}$, with more constrained FH BW e.g. $\hat{M} \leq 10\%$, the degree of freedom in selecting the APs is rather limited and both algorithms converge to much similar distribution.

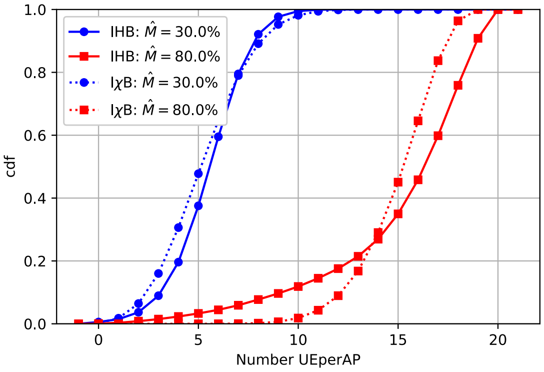

Fig. 7. cdf of the number of served UEs per AP for IHB and IχB.

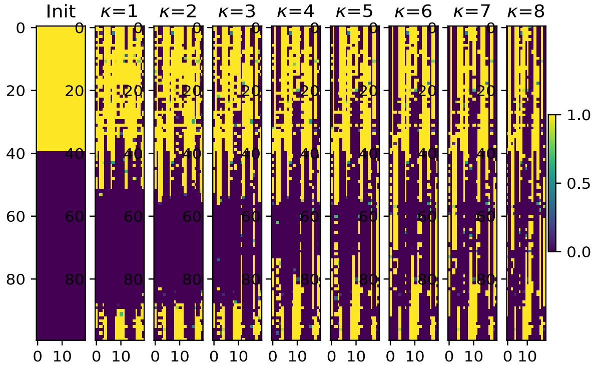

Fig. 8. Association matrix versus the super-iteration index κ.

C. Convergence Analysis of the IHB and the IχB Algorithms

We first explore how the joint optimization algorithm proposed in this paper finds a subpotimal association matrix over subsequent iterations. To do this, for the maximum budget $\hat{M}$ =40%, we show an example of the (relaxed) association matrix $\mathbf{C}_\kappa$ in Fig. 8 and the corresponding power allocation $\boldsymbol{\eta}_\kappa$ in Fig. 9 resulting from IHB versus the super-iteration κ between the power allocation and the AP scheduling. Note that the results in Fig. 8 start from a deterministic association matrix (see title '*Init*'). The figure shows that the IHB converges to a stable set of feasible solutions of the search space in a few iterations (essentially 5 iterations). The convergence of $\mathbf{C}$ dictates the convergence of $\boldsymbol{\eta}$ hence the iterative process stops as described in section IV. This conclusion seems rather intuitive as the IHB relies on hardening the effective channel.

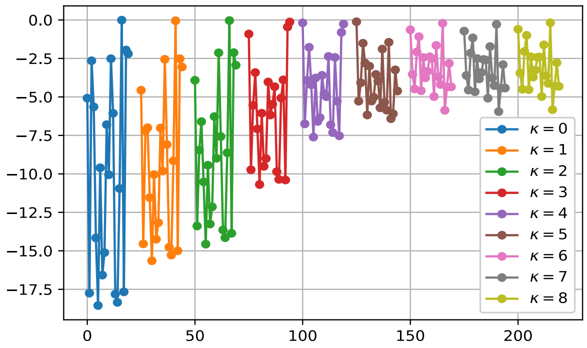

Fig. 9. Power allocation versus the super-iteration index κ.

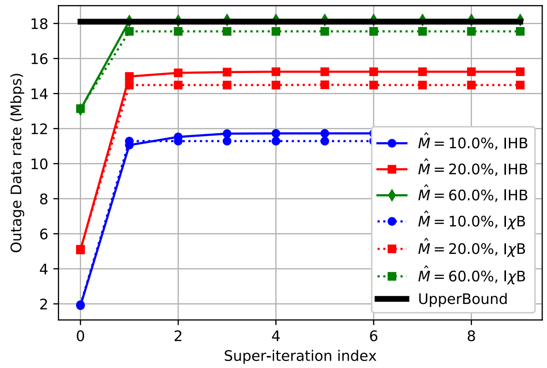

Fig. 10. 5% outage data rate resulting from IHB and IχB versus the super-iteration index κ.

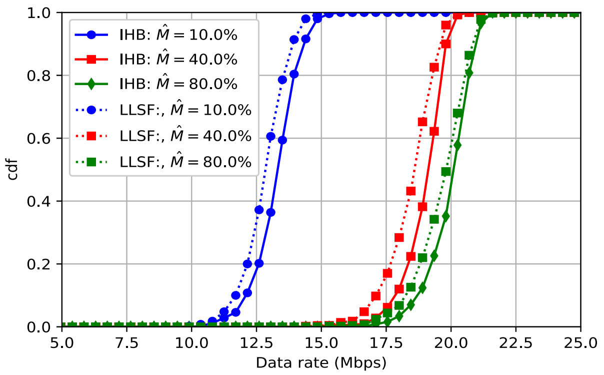

Fig. 11. Data rate cdf resulting from the LLSF and IHB algorithms.

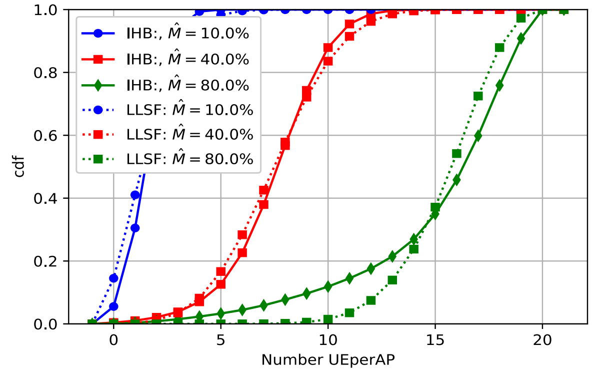

Fig. 12. The number of served UEs per AP cdf resulting from the LLSF and IHB algorithms.

Any (significant) improvement in the optimized hardened channel will (significantly) impact the power distribution. When initializing the association matrix randomly, the number of required super-iterations is reduced (see discussion later regarding Fig. 10).

The simulation results reported in Fig. 10 on the other hand, indicate that the IχB exhibits much lower number of super-iterations than the IHB algorithm irrespective of the initial configuration (figure with deterministic initialization not included). This is mainly due to the fact that the modified upper-bound of the optimization problem (20), which is proportional to $(-\sigma_w^2/\rho - \hat{\chi}(\boldsymbol{\eta}))/\eta_k \simeq -\hat{\chi}(\boldsymbol{\eta})/\eta_k$, seems to be a good a approximation and is less sensitive to the power distribution.

As another useful metric, in Fig. 10, the 5% outage data rates attained after applying any of the proposed algorithms are compared. It can be observed that both the IHB and IχB algorithms converge to an asymptotic rate within few iterations with the IHB algorithm slightly outperforming IχB for all the budgets. As reported previously in Fig. 6, significant outage data rate is observed with stringent FH BW. The gain for instance after the first super-iteration is 1100% and 17% for respectively $\hat{M} = 10\%$ and $\hat{M} = 60\%$. Hence, the gain from the super-iteration process becomes rather limited as $\hat{M}$ increases, akin to the law of diminishing returns. For instance with $\hat{M} = 60\%$ and one super-iteration, the upper-bound on outage rate is reached. In other words, only a subset of APs is needed to serve a given user as has also been reported for DL in [5], [13].

D. Comparison of the Proposed IHB and LLSF Algorithms

In Figs. 11 and 12, we compare the performance of the proposed LLSF and the IHB algorithms. Beside the data rate improvement achieved by using the systematic joint optimization algorithm proposed in this paper (see Fig. 11), the proposed heuristic has in overall two main drawbacks compared to IHB. First, the IHB results, as opposed to the LLSF, in a more uniform load across the APs (see Fig. 12), which brings several hardware advantages such as reduced power consumption and peak-to-average-power-ratio. For instance the probability of more than 9 served UEs per AP is respectively 90% and 100% for the IHB and the LLSF algorithms. Secondly, in dynamic environments with patterns such as joining/splitting, it is desirable to leave some headroom per AP such that the joining users (close to the hotspots) can get a decent throughput.

This is further highlighted in Fig. 13 showing the distribution of the number of served UEs per AP for an example hotspot corner case when the UEs are located in a $10 \times 10\text{m}^2$ surface centered in the middle of the square in Fig. 4. Irrespective of the maximum budget $\hat{M}$, the IHB yields much more uniform distribution compared to the LLSF algorithm while, in terms of the rate performance, the former iterative algorithm slightly outperforms LLSF (figure not included).

Although the proposed iterative algorithms generate a worst case data rate distribution that is slightly better than the LLSF with some increased complexity, they offer a systematic way to jointly optimize the power and association matrix that is agnostic to the radio access topology.

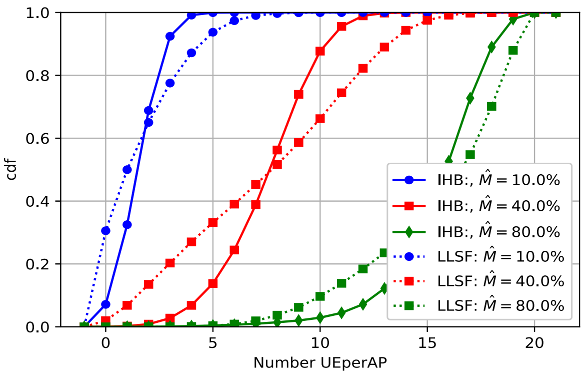

Fig. 13. Comparison of the LLSF and IHB algorithms in terms of the number of served UEs per AP distributions for a 10×10 m^2 centered hotspot configuration.

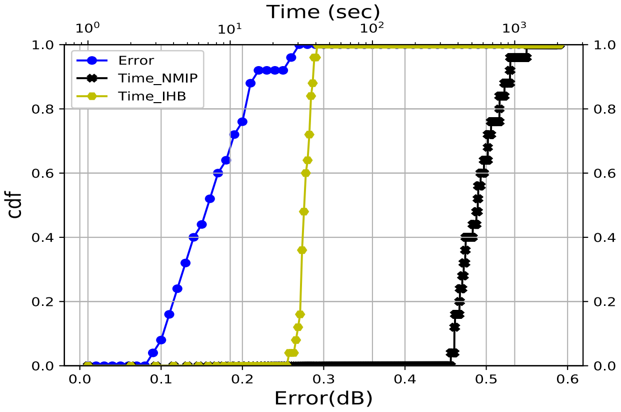

Fig. 14. The cdf of the difference in the optimal objectives for the IHB algorithm and the NMIP solution of [26], $K = 64$ and $M = 128$.

E. Analysis of Optimality

The original problem in (9) is non-convex and, therefore, has several locally optimal solutions. It was shown by Lemma 1 that the proposed framework in this paper converges to a sub-optimal solution, especially, for large-scale systems with a relatively big number of APs and/or UEs. To evaluate the proximity of the solution to the global bound, we compare the results of the algorithms in this paper to an exhaustive Nonlinear Mixed Integer Programming (NMIP) solver [26]. The joint AP-scheduling and power-allocation problem is now simulated with $M = 128$ and $K = 64$ users. The results from two algorithms are obtained using 1) IHB algorithm given by Algorithm 3, and 2) NMIP solver by [26] applied to the original problem in (9). For both cases, the optimization is applied over 100 randomly generated settings and, knowing that the solution from [26] is nearly a global optimum, the SNR difference to the solution of the NMIP solver is calculated and the resulting cdf is depicted in Fig. 14. Beside the error analysis, we also present the computational-time cdf for any of the above-mentioned algorithms in Fig. 14 where all the experiments were running on the same Intel Core i5-8350U CPU with the clock-frequency of 1.7 GHz.

Results in Fig. 14 reveal two important observations. First, 90% of settings show less than 0.25 dB difference in the optimal objective. In addition, the IHB algorithm completes the full optimization task in less than 28 sec for more than 90% of settings. This number jumps over 925 sec when NMIP solver is applied to the same cases consumed by the IHB algorithm. Indeed, although the NMIP solver allegedly attains slightly better optimal objective compared to the IHB algorithm, it suffers from the expensive computational complexity, especially when the number of UEs begins to grow. On the other hand, while IHB algorithm proposes a solution that scales linearly with the number of UEs/APs, the solver does not scale linearly as the computational time changes in a near polynomial order with respect to the umber of users. This last observation will make the realization of the NMIP solver infeasible for massive MIMO applications with large number of UEs and APs.

VII. Conclusion

In this paper, we developed a generic optimization framework for joint power control and AP scheduling of UL CF massive MIMO systems. Driven by the complexity reduction, two sub-optimal low-complexity convex algorithms involving linear programming are proposed fitting nicely in the bisection that aims at maximizing the worst case SNR. The proposed algorithms, benchmarked with a heuristic based on the largest large scale fading coefficients and with a popular exhaustive nonlinear mixed integer programming solver, reveal very competitive (scalable) performance-complexity tradeoffs. Simulation results confirm that, with stringent FH BW requirements, proper power control and AP scheduling is essential to improve the performance. The proposed algorithms are applicable to any UL CF massive MIMO with serial (P2MP), individual (P2P), or multi-serial FH architectures, and they provide quasi-optimal performances and minimize the FH BW requirements.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] J. Mietzner, R. Schober, L. Lampe, W. Gerstacker, and P. Hoeher, "Multiple-antenna techniques for wireless communications–a comprehensive literature survey," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 2, pp. 87–105, 2nd Quart., 2009.

[3] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[4] T. E. Bogale and L. B. Le, "Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges," *IEEE Veh. Technol. Mag.*, vol. 11, no. 1, pp. 64–75, Mar. 2016.

[5] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 197, Aug. 2019.

[6] P. Popovski *et al.*, "Wireless access ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[7] Z. Chen and E. Bjornson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.

[8] T. L. Marzetta, E. G. Larsson, H. Yang, and Q. H Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[9] A. Puglielli *et al.*, "A scalable massive MIMO array architecture based on common modules," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1310–1315.

[10] P. Frenger, J. Hederen, M. Hessler, and G. Interdonato, "Improved antenna arrangement for distributed massive MIMO," U.S. Patent Appl. 2018 103 897, Jan. 21, 2020. [Online]. Available: https://patentscope.wipo.int/search/en/WO2018103897

[11] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[12] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[13] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[14] G. Dong, H. Zhang, S. Jin, and D. Yuan, "Energy-efficiency-oriented joint user association and power allocation in distributed massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5794–5808, Jun. 2019.

[15] T. Van Chien, E. Bjornson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798–6812, Oct. 2020.

[16] G. Femenias, N. Lassoued, and F. Riera-Palou, "Access point switch ON/OFF strategies for green cell-free massive MIMO networking," *IEEE Access*, vol. 8, pp. 21788–21803, 2020.

[17] J. Garcia-Morales, G. Femenias, and F. Riera-Palou, "Energy-efficient access-point sleep-mode techniques for cell-free mmWave massive MIMO networks with non-uniform spatial traffic density," *IEEE Access*, vol. 8, pp. 137587–137605, 2020.

[18] A. Ashikhmin, L. Li, and T. L. Marzetta, "Interference reduction in multi-cell massive MIMO systems with large-scale fading precoding," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6340–6361, Sep. 2018.

[19] S. Blandino, G. Mangraviti, C. Desset, A. Bourdoux, P. Wambacq, and S. Pollin, "Multi-user hybrid MIMO at 60 GHz using 16-antenna transmitters," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 2, pp. 848–858, Feb. 2019.

[20] T. L. Marzetta, "How much training is required for multiuser MIMO?" in *Proc. Fortieth Asilomar Conf. Signals, Syst. Comput.*, Oct. 2006, pp. 359–363.

[21] R. Cendrillon and M. Moonen, "Iterative spectrum balancing for digital subscriber lines," in *Proc. IEEE ICC*, May 2005, pp. 1937–1941.

[22] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.

[23] S. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] L. P. Qian, Y. Wu, H. Zhou, and X. Shen, "Joint uplink base station association and power control for small-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5567–5582, Sep. 2017.

[25] D. Goldfarb and M. J. Todd, *Handbooks in Operations Research and Management Science: Chapter II Linear programming*. Amsterdam, The Netherlands: Elsevier, 1989, pp. 73–170.

[26] *Introduction to Ipopt: A Tutorial for Downloading, Installing, and Using Ipopt*, Open-Source Manual, Apr. 2015. [Online]. Available: https://coin-or.github.io/

[27] A. Bourdoux, B. Come, and N. Khaled, "Non-reciprocal transceivers in OFDM/SDMA systems: Impact and mitigation," in *Proc. RAWCON*, Aug. 2003. [Online]. Available: https://ieeexplore.ieee.org/document/1227923

Mamoun Guenach (Senior Member, IEEE) graduated from EMI, Morocco, in 1995. He received the Ph.D. degree from UCL, Belgium, in 2002. He was a Post-Doctoral Researcher with Ghent University, from 2002 to 2006, where he has been a part-time Visiting Professor since 2015. He was a member of Nokia Bell Labs Technical Staff, from 2006 to 2019. In 2019, he joined imec, Leuven, as a Research Scientist. His main research interests include coding, modulation, and equalization for wired and wireless communication technologies.

Ali A. Gorji received the B.Sc. and M.Sc. degrees from the Amirkabir University of Technology, Iran, in 2005 and 2008, respectively, and the Ph.D. degree from McMaster University, Canada, in 2012. He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. His research interests are in the areas of statistical signal processing, detection and estimation theory, target tracking, and their applications in radar, sensor systems, and robotics.

André Bourdoux (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from the Université Catholique de Louvain-la-Neuve, Belgium, in 1982. In 1998, he joined IMEC, where he is currently a Principal Member of Technical Staff with the Internet-of-Things Research Group. He holds several patents in these fields. He has authored or coauthored over 160 publications in books and peer-reviewed journals and conferences. His research interests include advanced signal processing and machine learning for wireless physical layer and high-resolution 3D/4D radars.