Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks

Jingya Li, Emil Björnson, Tommy Svensson, Thomas Eriksson, Mérouane Debbah

Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks

Jingya Li, *Student Member, IEEE*, Emil Björnson, *Member, IEEE*, Tommy Svensson, *Senior Member, IEEE*, Thomas Eriksson, and Mérouane Debbah, *Fellow, IEEE*

***Abstract*—This paper considers a downlink heterogeneous network, where different types of multiantenna base stations (BSs) communicate with a number of single-antenna users. Multiple BSs can serve the users by spatial multiflow transmission techniques. Assuming imperfect channel state information at both BSs and users, the precoding, load balancing, and BS operation mode are jointly optimized for improving the network energy efficiency. We minimize the weighted total power consumption while satisfying quality-of-service constraints at the users. This problem is nonconvex, but we prove that for each BS mode combination, the considered problem has a hidden convexity structure. Thus, the optimal solution is obtained by an exhaustive search over all possible BS mode combinations. Furthermore, by iterative convex approximations of the nonconvex objective function, a heuristic algorithm is proposed to obtain a suboptimal solution of low complexity. We show that although multicell joint transmission is allowed, in most cases, it is optimal for each user to be served by a single BS. The optimal BS association condition is parameterized, which reveals how it is impacted by different system parameters. Simulation results indicate that putting a BS into sleep mode by proper load balancing is an important solution for energy savings.**

***Index Terms*—Energy efficiency, heterogeneous networks, load balancing, precoding design.**

I. Introduction

THE rapid growth of data traffic in wireless networks impose great challenges on future wireless communication systems [2]–[4], in particular on improving the spectral efficiency as well as the energy efficiency. At the same time, the users are expecting that future networks will provide a uniform quality of service (QoS) over the coverage area. In many challenging scenarios, e.g., in shopping malls, dense urban environments, or during the occurrence of traffic jams, the users are non-uniformly distributed over the network [5]. One widely acknowledged cost- and energy-efficient approach to tackle these challenges is the concept of heterogeneous dense networks, where the traditional macro base stations (BSs) are complemented with a dense deployment of low-cost and low-power BSs [6]–[8]. By adding such a large number of small cells, the corresponding low-power BSs can offload traffic from the macro BSs, reduce the average distance between users and transmitters, and thereby improve the data rates and/or reduce the average transmit power. Since the data traffic load fluctuates greatly over the day [9], both macro and small cells might be needed at peak hours while there is an opportunity to turn off some BSs when there is little traffic in the corresponding coverage areas. *Load balancing* is the technique that maps the current traffic load to the available transmission resources, i.e., associates users with BSs. Mathematically speaking, the network would like to find the BS association that maximizes some performance metric, under the condition that the QoS requirements of all users are fulfilled.

Different from the traditional cellular networks, the densely deployed BSs will be heterogeneous in the number of antennas, transmit power, backhaul capacity and reliability, coverage area, etc. Moreover, the channel state information (CSI) at each BS is likely to be different and imperfect. In this complex scenario, a major research problem is to design low-complexity and robust coordinated multi-BS transmission schemes that minimize the total power consumption, while satisfying the QoS expectations of the users.

The total power consumption of the network can be modeled with a circuit part that depends on the transceiver hardware and a dynamic part that is a function of the transmitted signal power [10]–[13]. Adding more low-power BSs can reduce the dynamic power consumption due to the shorter propagation distances, but require more hardware; thus, it will increase the circuit power part. Note that the circuit power consumption also depends on the operational mode of each BS, i.e., whether the BS is active or in sleep mode. It has been shown that, putting a BS into sleep mode when there is nothing to transmit or receive is an important solution for energy savings [11]. Therefore, to actually improve the overall power efficiency of a heterogeneous network, the cooperation scheme, the BS operational modes, and the load balancing must be properly and jointly optimized.

Manuscript received September 6, 2014; revised March 2, 2015 and May 22, 2015; accepted May 31, 2015. Date of publication June 9, 2015; date of current version October 8, 2015. This paper was presented in part at the IEEE International Conference on Communications, London, U.K., June 2015 [1]. This work was supported in part by FP7 project ICT-317669 METIS, which is partly funded by the EU, by the project 621-2009-4555 Dynamic Multipoint Wireless Transmission and the International Postdoc Grant 2012-228 from the Swedish Research Council, and by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering). The associate editor coordinating the review of this paper and approving it for publication was J. Cheng.

J. Li, T. Svensson, and T. Eriksson are with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg 412 96, Sweden (e-mail: jingya.li@chalmers.se; tommy.svensson@chalmers.se; thomase@chalmers.se).

E. Björnson was with Supélec, Gif-sur-Yvette 91190, France, and also with the Department of Signal Processing, KTH Royal Institute of Technology, Stockholm 100 44, Sweden. He is now with the Department of Electrical Engineering (ISY), Linköping University, Linköping 581 83, Sweden (e-mail: emil.bjornson@liu.se).

M. Debbah is with CentraleSupélec, Gif-sur-Yvette 91190, France (e-mail: merouane.debbah@centralesupelec.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2443044

Simulation-based studies for load balancing in heterogeneous networks have been performed within 3GPP, and several biased-received-power based criteria were proposed to control the number of users associated with the low-power BSs [6], [7]. Moreover, load balancing was analyzed in [14]–[23] for systems where the BSs are distributed according to stochastic point processes. Using stochastic geometry tools, these works have compared how different BS association rules (e.g., the nearest-BS based, the highest-received-power based, the maximum signal-to-interference-and-noise ratio (SINR) based, and the biased-SINR based cell selection) affect the downlink SINR distribution [15]–[17] and the average achievable rate [17]–[21]. We note that the results in [14]–[19], [21], [22] are limited to BSs with single antennas, while contemporary and future networks use multiple antennas for downlink precoding. The papers [20] and [23] consider the practically important case of multi-antenna BSs, but these results are restricted to single-cell zero-forcing precoding with perfect CSI; in contrast, imperfect CSI and inter-cell interference coordination are essential properties of future heterogeneous networks. Moreover, shadowing has a great impact on the system performance of heterogeneous networks, but was not considered in [14], [15], [17], [18], [20]–[23], probably due to mathematical intractability.

The precoding design is of paramount importance in multi-antenna cellular networks, since it determines the achievable array gains and interference suppression [24]. Joint precoding and load balancing was studied in [25] for a homogeneous network, where all BSs are turned on and there is no explicit power constraints. In [26] and [27], the authors investigated joint load balancing and power control in heterogeneous networks with single-antenna BSs, where different algorithms were proposed to maximize the minimum rate subject to per-BS power constraints. Considering heterogeneous networks with multi-antenna BSs, joint load balancing and precoding algorithms were designed in [28]–[30] to maximize various system utilities. In [31], downlink linear precoding problems were studied jointly with BS selection. The objective was to either minimize the total transmit power or maximize the sum rate performance. The results in [31] show that by imposing certain sparsity patterns in the precoding vectors, the number of active BSs in the network can be effectively reduced. With the objective of improving network energy efficiency, radio resource optimization was studied in [32] for the downlink of an orthogonal frequency-division multiplexing (OFDM) system. In particular, the power allocation, subcarrier allocation and the number of activated transmit antennas were jointly optimized for maximization of the energy efficiency of data transmission (bit/Joule delivered to the users). However, the work in [32] did not optimize the precoding vectors and the results were limited to a single-cell scenario. In [33], using a stochastic geometry based model, the energy efficiency of both multi-cell homogeneous and heterogeneous networks was analyzed by considering active and sleep modes for macro BSs with fixed power control. Since both BSs and users are assumed to have a single antenna in [33], precoding design was not considered.

Joint precoding and load balancing design problem is typically a mixed-integer nonlinear programming problem, for which finding the global optimum is challenging [34]. Inspired by the compressive sensing literature, the reweighted l_1-norm technique has been adopted in [35]–[39], where different heuristic algorithms were proposed for solving joint precoding and BS clustering design problems. In [40], [41], group sparse optimization has been used to improve the energy efficiency of cloud radio access networks, where the weighted mixed l_1/l_p-norm minimization is used to induce group sparsity on the beamforming. The BSs are switched off based on the obtained group sparsity patterns. Note that in [37]–[41] the algorithms are designed based on the assumption of perfect CSI at both BSs and users. In this paper, we study joint precoding and load balancing optimization for energy efficient heterogeneous networks with imperfect CSI. The goal is to minimize the weighted total power consumption while satisfying QoS constraints at the users and transmit power constraints at the BSs. Although it is practically convenient and desirable to associate each user with only one BS per time-frequency resource block, our system model allows for serving users by multiple BSs. The paper investigates the following important system design questions: 1) Which and how many BSs should each user be associated with? 2) How should the precoding matrices be selected when having imperfect CSI? 3) How can we decide on the operational mode (active or sleep) for each BS? The contributions of this paper can be summarized as follows:

- We formulate the joint load balancing and precoding as a non-convex optimization problem. We show that for a given combination of BS modes, the considered optimization problem can be reformulated as a convex semi-definite problem. Thus, we obtain the global optimal solution by an exhaustive search over all possible BS mode combinations. The obtained global optimal solution serves as an upper bound for any other suboptimal precoding and load balancing solutions, e.g., the strategies proposed in [6], [7], [14]–[23].
- We derive the structure of the optimal solution, by investigating the structure of the dual problem. Our result verifies the intuition that, in most cases, it is optimal for each user to be served by a single BS. However, there are also occasions when multi-BS association is beneficial. Moreover, we show that the load balancing rules previously considered in [6], [7], [14]–[23] are not optimal when minimizing the total power consumption under per-BS transmit power constraints and per-user QoS constraints. The optimal BS association rule consists of comparing weighted channel norms, where the weighting matrix depends on channel uncertainty, power constraints, and QoS constraints.
- We propose an efficient iterative algorithm that resolves the non-convexity of the original optimization problem by iterative convex approximations of the power consumption functions. Each iteration solves a convex problem with a modified objective function. This convex objective function is updated in each iteration such that most of the BSs with small transmit powers in the solution are driven to sleep mode. We show that the idea behind the proposed algorithm is very similar to the reweighted l_1-norm minimization based methods used in [37]–[39].

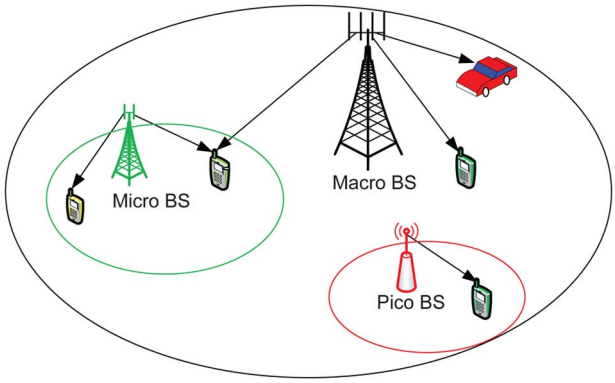

Fig. 1. Illustration of a downlink three-tier heterogeneous network consisting of macro, micro and pico BSs.

- Numerical results are provided to show how putting BSs into sleep mode by proper load balancing is a key to energy savings in heterogeneous networks. The BS activation probability is shown to depend on the target QoS requirements, as well as the ratio between the circuit power consumed in the active mode and that consumed in the sleep mode.

The remainder of this paper is organized as follows: Section II introduces the system and signal model. In Section III, we analyze the optimal precoding and load balancing design. In Section IV, an iterative heuristic algorithm is proposed to obtain a suboptimal solution with low complexity. Section V provides a set of numerical results to illustrate our analytical results and the proposed algorithms. Finally, the main results of the paper are summarized in Section VI.

Notation: we use upper-case bold face letters, such as $\mathbf{E}$, for matrices and lower-case bold face letters, such as $\mathbf{h}$, for vectors. $\mathbf{W} \succeq 0$ represents that the matrix $\mathbf{W}$ is positive semidefinite. $|\mathcal{C}|$ denotes the cardinality of a set $\mathcal{C}$. The operator $\mathbb{E}\{\cdot\}$ stands for expectation. The notation $\sim$ denotes "distributed as," $\triangleq$ is used to mark definitions, $\|\cdot\|$ represents the Euclidean norm, and $\mathrm{Tr}(\cdot)$ is the matrix trace.

II. System and Signal Model

We consider the downlink of a heterogeneous network consisting of M BSs and K single-antenna users, as illustrated in Fig. 1. The heterogeneity lies in the assumption that the M BSs are different in terms of the number of transmit antennas, the power consumption characteristics, the channel propagation model, and the CSI quality. BSs with the same characteristics can be said to belong to the same tier or category (e.g., macro or small BS), but we stress that our system model supports anything from 1 to M tiers. The users are not pre-associated with any particular cell and are randomly distributed in the network coverage area.

BS v is assumed to have N_v antennas. The channel from BS v to user k is assumed to be flat-fading, and denoted by $\mathbf{h}_{k,v} \in \mathbb{C}^{N_v \times 1}$ for $v = 1, \ldots, M$ and $k = 1, \ldots, K$. In practice, these channels are imperfectly known at the BSs. This is modeled as $\mathbf{h}_{k,v} = \hat{\mathbf{h}}_{k,v} + \mathbf{e}_{k,v}$, where $\hat{\mathbf{h}}_{k,v}$ is the known estimate of $\mathbf{h}_{k,v}$ at BS v. The error vector $\mathbf{e}_{k,v} \backsim \mathcal{CN}(\mathbf{0}, \mathbf{E}_{k,v})$ is assumed to have zero-mean and a covariance matrix $\mathbf{E}_{k,v} \in \mathbb{C}^{N_v \times N_v}$. This is, for example, a good model of time-division duplex (TDD) systems where the channels are Rayleigh fading, $\mathbf{h}_{k,v} \sim \mathcal{CN}(\mathbf{0}, g_{k,v}\mathbf{I}_{N_v})$, and the BS uses uplink pilot signals for channel estimation. Here, $g_{k,v}$ denotes the average channel gain between BS v and user k, including pathloss and shadowing. If the minimum mean-squared error (MMSE) channel estimator is used [42]–[45], then estimation errors are zero-mean complex Gaussian distributed and the error covariance becomes

$$\mathbf{E}_{k,v} = \frac{g_{k,v}}{1 + \gamma_{k,v}^{p}} \mathbf{I}_{N_v} \tag{1}$$

where $\gamma_{k,v}^{p} = \frac{p g_{k,v}}{\sigma_k^2}$ denotes the pilot SNR, p is the total pilot power and σ_k^2 is the noise power. The users also need to acquire CSI, but only for the precoded channels; this is further discussed in Section II-B.

The received signal at user k is

$$y_k = \sum_{v=1}^{M} \mathbf{h}_{k,v}^{H} \mathbf{x}_v + n_k \tag{2}$$

where $\mathbf{x}_v \in \mathbb{C}^{N_v \times 1}$ is the transmitted signal from BS v and $n_k \backsim \mathcal{CN}(0, \sigma_k^2)$ is the independent additive receiver noise at user k.

A main goal of this paper is to determine the optimal association between users and BSs. It makes practical sense to only associate one BS with each user, but we will not make this limiting assumption at this point since we simply do not know if it is optimal. Instead, we assume that all BSs are able to transmit to all users at the same time-frequency resource block, and then our analysis will tell which and how many BSs that each user should be associated with. Motivated by the fact that tight phase synchronization between BSs is extremely difficult to achieve in practice, only linear spatial multiflow transmission is allowed [46]. This is a scheme for multiple access that allows each user to receive different parallel data streams from multiple BSs. These streams are detected sequentially at the user, based on conventional successive interference cancelation techniques [47]. Define $\mathcal{V} \triangleq \{1, 2, \ldots, M\}$ as the set of all BSs in the network, and let $\mathcal{V}_k \subseteq \mathcal{V}$ denote the set of BSs that provide data transmission to user k. Then, the set of users associated with BS v can be represented by $\mathcal{U}_v = \{k | v \in \mathcal{V}_k\}$. Let $s_{k,v} \backsim \mathcal{CN}(0, 1)$ be the coded independent information symbols for user k, transmitted from BS v.[1] Then, the desired signals for user k transmitted by BS v is $\mathbf{w}_{k,v} s_{k,v}$, where $\mathbf{w}_{k,v} \in \mathbb{C}^{N_v \times 1}$ is the linear precoding vector for user k at BS v. The aggregated transmitted signal from BS v is

$$\mathbf{x}_v = \sum_{k \in \mathcal{U}_v} \mathbf{w}_{k,v} s_{k,v}. \tag{3}$$

[1] Note that, the data symbols $s_{k,v}$ for user k are independent for different BS v. This spatial multiflow transmission scheme is different from the traditional network MIMO scheme, which assumes that the same user data s_k is transmitted from all BSs. The network MIMO scheme allows for coherent joint transmission from all BSs, however, it requires tight phase synchronization between all BSs.

TABLE I
POWER MODEL PARAMETERS FOR DIFFERENT BS TYPES

BS type	N_v	$P_{v,\max}$	$P_{\text{sleep},v}$	$P_{\text{active},v}$	Δ_v
MBS	4	39.8 W	75.0 W	130.0 W	4.7
SBS	2	6.3 W	39.0 W	56.0 W	2.6

A. Power Consumption Model

From (3), the expected transmit power from BS v can be calculated as

$$P_{\text{trans},v} = \sum_{k\in\mathcal{U}_v} \|\mathbf{w}_{k,v}\|^2 \mathbb{E}\left\{|s_{k,v}|^2\right\} = \sum_{k\in\mathcal{U}_v} \|\mathbf{w}_{k,v}\|^2. \quad (4)$$

In this paper, we adopt the linear approximated power consumption model proposed in [11, Eq. (4-3)] for 10 MHz bandwidth, where the total consumed power of BS v, for $v \in \mathcal{V}$, is

$$P_v = \begin{cases} N_v P_{\text{active},v} + \Delta_v P_{\text{trans},v}, & 0 < P_{\text{trans},v} \le P_{v,\max} \\ N_v P_{\text{sleep},v}, & P_{\text{trans},v} = 0, \end{cases} \quad (5)$$

where $P_{\text{active},v}$ is the hardware power consumption at BS v at the minimum non-zero transmit power, $P_{\text{sleep},v}$ denotes the sleep mode power consumption of BS v with $P_{\text{sleep},v} \le P_{\text{active},v}$. Note that $P_{\text{sleep},v} > 0$ in the sleep mode (due to the DC-DC power supply, mains supply, active cooling, maintaining backhaul connections, and enabling fast turn on control signaling) [10], [11]. Here, $P_{v,\max}$ is the peak transmit power constraint for BS v. The scaling factor, $\Delta_v \ge 1$, models the inefficiency of the power amplifier; that is, how much extra power that is consumed at BSs when the transmitted power is $P_{\text{trans},v}$. Some example values of $P_{\text{active},v}$, $P_{\text{sleep},v}$, $P_{v,\max}$ and Δ_v for different BS types can be found in [11, Table 8], and some of these are also given in Table I.

B. Aggregated Received SINR

Each user might receive multiple information symbols, thus we need an aggregated performance measure for each user. The natural choice is the sum spectral efficiency of the user when successive interference cancellation is applied.[2]

Lemma 1: Assume that user k knows the effective precoded channels $\mathbf{w}_{l,v}^H \hat{\mathbf{h}}_{k,v}$ (for all l and v). Then, a lower bound on the achievable ergodic sum spectral efficiency of user k is $R_k = \mathbb{E}\{\log_2(1+\gamma_k)\}$ where the expectation is with respect to the *aggregated* instantaneous SINR

$$\gamma_k = \frac{\sum_{v\in\mathcal{V}_k} \left|\hat{\mathbf{h}}_{k,v}^H \mathbf{w}_{k,v}\right|^2}{I_k + E_k + \sigma_k^2} \quad (6)$$

with

$$I_k \triangleq \sum_{v\in\mathcal{V}} \sum_{\substack{l\in\mathcal{U}_v \\ l\neq k}} \mathbf{w}_{l,v}^H \left(\hat{\mathbf{h}}_{k,v}\hat{\mathbf{h}}_{k,v}^H + \mathbf{E}_{k,v}\right) \mathbf{w}_{l,v} \quad (7)$$

being the co-user interference and

$$E_k \triangleq \sum_{v\in\mathcal{V}_k} \mathbf{w}_{k,v}^H \mathbf{E}_{k,v} \mathbf{w}_{k,v} \quad (8)$$

is the effective estimation errors on the channels related this user.

Proof: The achievable sum spectral efficiency is obtained, similar to [47], [48], by decoding the Gaussian information sequences from the different BSs in a sequential manner, using conventional successive interference cancellation. Since the users only know the effective channels $\mathbf{w}_{l,v}^H \hat{\mathbf{h}}_{k,v}$ and not the true channels $\mathbf{w}_{l,v}^H \mathbf{h}_{k,v}$, the channel uncertainty is handled by computing a lower bound on the mutual information, using the approach from [49] where all signals that are uncorrelated with $\mathbf{w}_{k,v}^H \hat{\mathbf{h}}_{k,v} s_{k,v}$ are treated as Gaussian noise (which is the worst case in terms of mutual information). This applies for both inter-user interference and the part of the desired signals that are conveyed over the zero-mean channel estimation error vectors. □

This lemma provides a lower bound on the achievable capacity, since the latter is unknown under imperfect CSI. We note that Lemma 1 assumes that the users know the effective precoded channels. In practice, the users can estimate these effective channels using downlink pilots, and get estimates of $\mathbf{w}_{l,v}^H \mathbf{h}_{k,v}$ that are at least as accurate $\mathbf{w}_{l,v}^H \hat{\mathbf{h}}_{k,v}$. Hence, it might be possible to achieve higher spectral efficiencies than in Lemma 1. Nevertheless, the aggregated SINR in (6) is the most convenient one for precoding design, since the BSs can only utilize their own CSI in the optimization.

C. Problem Formulation

The focus of this paper is on the joint design of load balancing (i.e., the UE association in $\mathcal{U}_v$) and precoding vectors ($\mathbf{w}_{k,v}$) for $v = 1, \ldots, M$ and $k = 1, \ldots, K$, which is an optimization that takes place at every channel realization. To this end, the goal is to minimize the weighted total power consumption (for any given channel realization) while satisfying a set of SINR constraints (or, equivalently, spectral efficiency constraints) for each user and a set of transmit power constraints for each BS. These constraints are referred to as the QoS constraints. With (4), (5), and (6) in hand, the optimization problem can be formulated as

$$\begin{aligned} \underset{\{\mathcal{U}_v\},\{\mathbf{w}_{k,v}\}}{\text{minimize}} \quad & \sum_{v=1}^{M} a_v P_v \\ \text{subject to} \quad & \gamma_k \ge \Gamma_k, \qquad \forall k \\ & P_{\text{trans},v} \le P_{v,\max}, \qquad \forall v \end{aligned} \quad (9)$$

where $\Gamma_k > 0$ is the target SINR value for user k. By satisfying this QoS target for every channel realization, the ergodic spectral efficiency is $R_k \ge \log_2(1+\Gamma_k)$. In this paper, we assume that the weights $a_v > 0$ are given. These weights can be used to balance the power consumptions of different BSs. For the rest of the paper, we assume that the problem (9) has at least one feasible solution, which is reasonable in dense networks with an over-provisioning of access points. In practice, if no feasible

[2]The power consumption at the user side might depend on how many symbols that the user receives, but this paper has an operator perspective where only the power consumptions at BSs is considered—this is the dominating factor in the downlink.

solution exists, the SINR constraints have to be relaxed either by decreasing the target SINRs or by removing users [25].

III. Optimal Precoding and Load Balancing

In this section, we solve the optimization problem in (9). As a first step, we show that the set variables $\mathcal{U}_v$ can be eliminated by optimizing over all precoding vectors.

Lemma 2: The original problem (9) is equivalent to[3]

$$\begin{aligned} \underset{\{\mathbf{w}_{k,v}\}}{\text{minimize}} \quad & \sum_{v=1}^{M} a_v P_v \\ \text{subject to} \quad & \gamma_k \geq \Gamma_k, \qquad \forall\, k \\ & \sum_{k=1}^{K} \|\mathbf{w}_{k,v}\|^2 \leq P_{v,\max}, \qquad \forall\, v \end{aligned} \tag{10}$$

where P_v can be rewritten as a function of $\mathbf{w}_{k,v}$ by substituting $P_{\text{trans},v} = \sum_{k=1}^{K} \|\mathbf{w}_{k,v}\|^2$ into (5), and γ_k is reformulated as

$$\gamma_k = \frac{\sum_{v=1}^{M} \left| \hat{\mathbf{h}}_{k,v}^H \mathbf{w}_{k,v} \right|^2}{I_k + E_k + \sigma_k^2} \tag{11}$$

with I_k rewritten as

$$I_k \triangleq \sum_{v=1}^{M} \sum_{\substack{l=1 \\ l \neq k}}^{K} \mathbf{w}_{l,v}^H \left(\hat{\mathbf{h}}_{k,v} \hat{\mathbf{h}}_{k,v}^H + \mathbf{E}_{k,v} \right) \mathbf{w}_{l,v} \tag{12}$$

and E_k replaced by

$$E_k \triangleq \sum_{v=1}^{M} \mathbf{w}_{k,v}^H \mathbf{E}_{k,v} \mathbf{w}_{k,v}. \tag{13}$$

Proof: Note that if BS j does not serve a particular user k (i.e., $k \notin \mathcal{U}_j$ and $j \notin \mathcal{V}_k$), then all terms that would have contained $\mathbf{w}_{k,j}$ in the SINR of (6) and the transmit power (4) are missing. This is equivalent to setting $\mathbf{w}_{k,j} = \mathbf{0}$ and adding said terms (which then are zero). Hence, the sets $\mathcal{U}_v$ and $\mathcal{V}_k$ are fully determined by checking which of the precoding vectors are zero:

$$\mathcal{U}_v = \left\{ k | \mathbf{w}_{k,v} \neq 0, k \in \{1, \ldots, K\} \right\}, \tag{14}$$

$$\mathcal{V}_k = \{ v | \mathbf{w}_{k,v} \neq 0, v \in \mathcal{V} \}. \tag{15}$$

The sets $\mathcal{U}_v$ can therefore be removed as optimization variables from (9), if we add the missing terms in (4) and (6). The corresponding equivalent problem is the one stated in this lemma. □

This lemma shows that we do not need to optimize the BS association sets $\mathcal{U}_v$ since these are implicitly determined by checking which precoding vectors that are non-zero. Note that although the expressions for P_v, γ_k, I_k, and E_k in Lemma 2 are different from the expressions in Section II, the values are identical for every selection of precoding vectors $\{\mathbf{w}_{k,v}\}$. As will be shown later, even if all BSs are allowed to transmit to all users at the same time-frequency resource block, in most cases, at the optimal point, each user k will be connected to only one BS.

[3]Here, "equivalent" means that the minimal value of problems (9) and (10) is the same and that the solution to (9) can be obtained from the solution to (10).

The optimization problem (10) is not convex. In particular, the power consumption function in (5) leads to a hard combinatorial problem [50]. Moreover, the SINR constraints of (10) do not have a standard convex form. In the following, we first show that, for each combination of BS modes (active or sleep), problem (10) can be reformulated as a convex problem. Then, the global optimum can be found by solving this convex problem for all 2^M combinations of modes.

Define $\mathbf{w}_k \triangleq \left[\mathbf{w}_{k,1}^T, \mathbf{w}_{k,2}^T, \ldots, \mathbf{w}_{k,M}^T \right]^T \in \mathbb{C}^{\left(\sum_{v=1}^{M} N_v\right) \times 1}$ as the aggregated precoding vector for user k from all BSs. We notice that the received SINR, γ_k in (11), can be expressed as

$$\gamma_k = \frac{\mathbf{w}_k^H \hat{\mathbf{R}}_k \mathbf{w}_k}{\sum_{\substack{l=1 \\ l \neq k}}^{K} \mathbf{w}_l^H \left(\hat{\mathbf{R}}_k + \mathbf{E}_k \right) \mathbf{w}_l + \mathbf{w}_k^H \mathbf{E}_k \mathbf{w}_k + \sigma_k^2}.$$

using the block-diagonal matrices

$$\mathbf{E}_k \triangleq \text{diag}(\mathbf{E}_{k,1}, \mathbf{E}_{k,2}, \ldots, \mathbf{E}_{k,M}) \tag{16}$$

$$\hat{\mathbf{R}}_k \triangleq \begin{bmatrix} \hat{\mathbf{R}}_{k,1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_{k,2} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \vdots & \mathbf{0} & \hat{\mathbf{R}}_{k,M} \end{bmatrix} \tag{17}$$

with the diagonal blocks $\hat{\mathbf{R}}_{k,v} \triangleq \hat{\mathbf{h}}_{k,v} \hat{\mathbf{h}}_{k,v}^H \in \mathbb{C}^{N_v \times N_v}$ for $v = 1, \ldots, M$.

Similarly, the power constraints in (10) are written in terms of $\mathbf{w}_k$ as $\sum_{k=1}^{K} \mathbf{w}_k^H \boldsymbol{Q}_v \mathbf{w}_k$, where

$$\boldsymbol{Q}_v \triangleq \text{diag}(\mathbf{Q}_{1,v}, \mathbf{Q}_{2,v}, \ldots, \mathbf{Q}_{M,v}) \tag{18}$$

$$\mathbf{Q}_{i,v} \triangleq \begin{cases} \mathbf{I}_{N_v}, & \text{if } i = v \\ \mathbf{0}_{N_v \times N_v}, & \text{otherwise.} \end{cases} \tag{19}$$

With this notation, the optimization problem (10) looks like a classical precoding optimization problem of the type in [51], but with the important difference that $\hat{\mathbf{R}}_k$ has rank M and not rank 1 as in the case with one BS per user. Hence, we cannot use the second-order cone techniques from [51], but the following semi-definite relaxation approach.[4]

Lemma 3: Let z_v be the BS mode indicator for $v \in \mathcal{V}$: $z_v = 1$ if BS v is active, and $z_v = 0$ if BS v is in sleep mode. Define $\mathbf{W}_k \triangleq \mathbf{w}_k \mathbf{w}_k^H \succeq 0$. Consider the following semi-definite relaxation of (10) for fixed BS modes:

$$\begin{aligned} \underset{\{\mathbf{W}_k \succeq 0\}}{\text{minimize}} \quad & \sum_{k=1}^{K} \text{Tr}(\mathbf{A}\mathbf{W}_k) + \mathcal{J}(\mathbf{z}) \\ \text{subject to} \quad & \text{Tr}(\hat{\mathbf{R}}_k \mathbf{W}_k) - \Gamma_k \sum_{\substack{l=1 \\ l \neq k}}^{K} \text{Tr}\left((\hat{\mathbf{R}}_k + \mathbf{E}_k) \mathbf{W}_l \right) \\ & - \Gamma_k \text{Tr}(\mathbf{E}_k \mathbf{W}_k) \geq \Gamma_k \sigma_k^2, \qquad \forall\, k \\ & \sum_{k=1}^{K} \text{Tr}(\boldsymbol{Q}_v \mathbf{W}_k) \leq z_v P_{v,\max}, \qquad \forall\, v \end{aligned} \tag{20}$$

[4]Semi-definite relaxation means that the optimization variables changed to $\mathbf{W}_k \triangleq \mathbf{w}_k \mathbf{w}_k^H \succeq 0$ instead of $\mathbf{w}_k$. This would require an additional rank constraint, $\text{rank}(\mathbf{W}_k) = 1, \forall k$, but this one is dropped as a relaxation.

where $z_v \in \{0, 1\}, \forall v$ and

$$\mathbf{A} \triangleq \text{diag}\left(a_1\Delta_1\mathbf{I}_{N_1}, a_2\Delta_2\mathbf{I}_{N_2}, \ldots, a_M\Delta_M\mathbf{I}_{N_M}\right) \quad (21)$$

$$\mathcal{J}(\mathbf{z}) = \sum_v a_v N_v \left(P_{\text{active},v} z_v + P_{\text{sleep},v}(1 - z_v)\right). \quad (22)$$

The problem (20) is a convex semi-definite program and it always has a rank one solution, if the problem is feasible.

Proof: For any fixed combination of BS modes $\mathbf{z} = [z_1, \ldots, z_M]$, $\mathcal{J}(\mathbf{z})$ in (22) is fixed. Then, the problem (20) is on the form of (P2) in [52]. Based on [52, Theorem 1], this type of optimization problems always has optimal solutions with rank one if it is feasible. □

Based on this lemma, we solve the original precoding and load balancing problem as follows.

Theorem 1: The global optimum to (9) is obtained by solving (20) for each of the 2^M mode combinations ($z_v = 0$ or $z_v = 1$ for each v) and selecting the solution that provides the lowest weighted total power consumption.

To summarize, Lemma 3 shows that semi-definite relaxation is tight for the problem at hand. For each fixed mode $\mathbf{z}$, we can solve (20) using standard convex optimization software, such as CVX [53] or YALMIP [54]. By doing this for all 2^M mode combinations, the global optimum to (9) is obtained. We stress that (9) optimizes the precoding, load balancing (i.e., BS association), and BS modes jointly. The global optimum to (9) is a benchmark for any suboptimal heuristic load-balancing and precoding algorithms; for example, the ones proposed in [6], [7], [14]–[16], [18], [20], [21], [23].

A. Structure of the Optimal Load Balancing

Theorem 1 shows how to solve the joint precoding and load balancing optimization problem (9) using convex optimization techniques. Although it provides the truly optimal solution, it brings little insight on the structure of the optimal load balancing. In the following, we will analyze the dual problem of (20) and thereby shed light on the optimal BS association.

Recall from Lemma 3 that (20) is a semi-definite optimization problem. This problem is convex and satisfies Slater's condition, which implies strong duality [55, Sec. 5.2.3]. The dual problem has the same optimal objective value as the original problem. Define $\mathcal{A} \triangleq \{v | z_v = 1, v \in \mathcal{V}\}$ as the set of active BSs, and $\mathcal{S} \triangleq \{v | z_v = 0, v \in \mathcal{V}\}$ as the set of BSs in the sleep mode. The Lagrangian of (20) is

$$\begin{aligned}
&\mathcal{L}\left(\{\mathbf{W}_k, \lambda_k, \mu_i, \nu_j\}\right) \\
&= \mathcal{J}(\mathbf{z}) + \sum_{k=1}^{K} \text{Tr}(\mathbf{A}\mathbf{W}_k) - \sum_{k=1}^{K} \lambda_k \text{Tr}(\hat{\mathbf{R}}_k \mathbf{W}_k) \\
&\quad + \sum_{k=1}^{K} \lambda_k \Gamma_k \sum_{\substack{l=1 \\ l\neq k}}^{K} \text{Tr}\left((\hat{\mathbf{R}}_k + \mathbf{E}_k)\mathbf{W}_l\right) + \sum_{k=1}^{K} \lambda_k \Gamma_k \left(\text{Tr}(\mathbf{E}_k \mathbf{W}_k) + \sigma_k^2\right) \\
&\quad + \sum_{i\in\mathcal{A}} \mu_i \left(\sum_{k=1}^{K} \text{Tr}(\boldsymbol{Q}_i \mathbf{W}_k) - P_{i,\max}\right) + \sum_{j\in\mathcal{S}} \nu_j \sum_{k=1}^{K} \text{Tr}(\boldsymbol{Q}_j \mathbf{W}_k)
\end{aligned} \quad (23)$$

where $\lambda_k, \mu_i, \nu_j \geq 0$ are the Lagrange multipliers associated to the k-th user's SINR constraint, the power constraint for BS i in set $\mathcal{A}$, and the power constraint for BS j in set $\mathcal{S}$, respectively. The dual problem to (20) is an unconstrained maximization of the dual function, defined as

$$g\left(\{\lambda_k, \mu_i, \nu_j\}\right) = \underset{\{\mathbf{W}_k\}}{\text{minimize}}\, \mathcal{L}\left(\{\mathbf{W}_k, \lambda_k, \mu_i, \nu_j\}\right). \quad (24)$$

Define

$$\mathbf{B}_k \triangleq \mathbf{A} + \lambda_k \Gamma_k \mathbf{E}_k + \sum_{\substack{l=1 \\ l\neq k}}^{K} \lambda_l \Gamma_l (\hat{\mathbf{R}}_l + \mathbf{E}_l) + \sum_{i\in\mathcal{A}} \mu_i \boldsymbol{Q}_i + \sum_{j\in\mathcal{S}} \nu_j \boldsymbol{Q}_j, \quad (25)$$

which is a block-diagonal matrix whose v-th block is

$$\begin{aligned}
\mathbf{B}_{k,v} \triangleq\ & a_v \Delta_v \mathbf{I}_{N_v} + \lambda_k \Gamma_k \mathbf{E}_{k,v} + \sum_{\substack{l=1 \\ l\neq k}}^{K} \lambda_l \Gamma_l (\hat{\mathbf{R}}_{l,v} + \mathbf{E}_{l,v}) \\
&+ \sum_{i\in\mathcal{A}} \mu_i \boldsymbol{Q}_{i,v} + \sum_{j\in\mathcal{S}} \nu_j \boldsymbol{Q}_{j,v}.
\end{aligned} \quad (26)$$

From (23), it is easy to show that $g(\{\lambda_k, \mu_i, \nu_j\}) = \mathcal{J}(\mathbf{z}) + \sum_{k=1}^{K} \lambda_k \Gamma_k \sigma_k^2 - \sum_{i\in\mathcal{A}} \mu_i P_{i,\max}$, if $\mathbf{B}_k - \lambda_k \hat{\mathbf{R}}_k \succeq 0$ for all $k = 1, \ldots, K$; otherwise, $g(\{\lambda_k, \mu_i, \nu_j\}) = -\infty$. Hence, the dual problem of (20) becomes

$$\begin{aligned}
\underset{\{\lambda_k, \mu_i, \nu_j \geq 0\}}{\text{maximize}} \quad & \mathcal{J}(\mathbf{z}) + \sum_{k=1}^{K} \lambda_k \Gamma_k \sigma_k^2 - \sum_{i\in\mathcal{A}} \mu_i P_{i,\max} \\
\text{subject to} \quad & \mathbf{B}_k - \lambda_k \hat{\mathbf{R}}_k \succeq 0, \quad \forall k.
\end{aligned} \quad (27)$$

Lemma 4: Let $\{\lambda_k^*, \mu_i^*, \nu_j^*\}$ denote the optimal Lagrange multipliers to (27), and let $\mathbf{B}_{k,v}^*$ be the value of $\mathbf{B}_{k,v}$ in (26) for these multipliers. The optimal precoding vectors are

$$\mathbf{w}_{k,v}^* = \begin{cases} \alpha_{k,v} \left(\mathbf{B}_{k,v}^*\right)^{-1} \hat{\mathbf{h}}_{k,v}, & \text{if } \lambda_k^* = \frac{1}{\hat{\mathbf{h}}_{k,v}^H \left(\mathbf{B}_{k,v}^*\right)^{-1} \hat{\mathbf{h}}_{k,v}}, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (28)$$

where $\alpha_{k,v} \geq 0$ is a scaling factor.

Proof: Since for any fixed $\mathbf{z}$, strong duality holds for (20) and the solution has rank one as $\mathbf{W}_k^* = \mathbf{w}_k^* (\mathbf{w}_k^*)^H$, the optimal $\mathbf{w}_k^*$ can be calculated by setting the first-order derivative of the Lagrangian in (23) with respect to $\mathbf{w}_k$ to zero; that is,

$$\left.\frac{\partial \mathcal{L}\left(\left\{\mathbf{W}_k, \lambda_k^*, \mu_i^*, \nu_j^*\right\}\right)}{\partial \mathbf{w}_k}\right|_{\mathbf{w}_k^*} = 2\left(\mathbf{B}_k^* - \lambda_k^* \hat{\mathbf{R}}_k\right)\mathbf{w}_k^* = \mathbf{0} \quad (29)$$

from which we have the condition

$$\mathbf{B}_{k,v}^* \mathbf{w}_{k,v}^* = \lambda_k^* \hat{\mathbf{h}}_{k,v} \hat{\mathbf{h}}_{k,v}^H \mathbf{w}_{k,v}^*, \qquad \forall v. \quad (30)$$

Hence,

$$\mathbf{w}_{k,v}^* = \alpha_{k,v} \left(\mathbf{B}_{k,v}^*\right)^{-1} \hat{\mathbf{h}}_{k,v} \quad (31)$$

for all k and v, where $\alpha_{k,v} \triangleq \lambda_k^* \hat{\mathbf{h}}_{k,v}^H \mathbf{w}_{k,v}^*$ is a scalar. Recall that we assume that the problem (9) has at least one feasible

solution. Thus, $\lambda_k^* > 0$ for all k. If we now multiply (29) by $(\mathbf{w}_k^*)^H$ from the left, we obtain the equivalent condition

$$2\left(\mathbf{w}_k^*\right)^H\left(\mathbf{B}_k^* - \lambda_k^*\hat{\mathbf{R}}_k\right)\mathbf{w}_k^* = 0 \quad \Leftrightarrow$$
$$2\left(\mathbf{w}_{k,v}^*\right)^H\left(\mathbf{B}_{k,v}^* - \lambda_k^*\hat{\mathbf{h}}_{k,v}\hat{\mathbf{h}}_{k,v}^H\right)\mathbf{w}_{k,v}^* = 0. \tag{32}$$

By plugging (31) into (32), we obtain the condition

$$\alpha_{k,v}^2\left(\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v} - \lambda_k^*\left(\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}\right)^2\right) = 0 \tag{33}$$

which is satisfied when either $\lambda_k^*\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v} = 1$ or $\alpha_{k,v} = 0$. These two cases correspond to the two cases in (28). □

Lemma 4 gives the structure of the optimal precoding vectors. In particular, it helps us to understand the optimal BS association (i.e., which precoding vectors $\mathbf{w}_{k,v}^*$ that are non zero).

Theorem 2: The optimal BS association for user k falls into one of the following two cases:

1) It is only served by one BS v^*, with $v^* = \arg\max_v \left(\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}\right)$, that is, $\mathcal{V}_k = \{v^*\}$;
2) It is served by a set of BSs $\mathcal{V}_k = \left\{v^* | v^* = \arg\max_v \left(\hat{\mathbf{h}}_{k,v}^H \times \left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}\right)\right\}$ where $|\mathcal{V}_k| > 1$.

Proof: We know from (28) in Lemma 4 that user k is associated with BSs v only if

$$\lambda_k^* = \frac{1}{\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}}. \tag{34}$$

Dual feasibility requires that $\mathbf{B}_k - \lambda_k\hat{\mathbf{R}}_k \succeq 0$ for all k, or equivalently that $\mathbf{u}_{k,v}^H\left(\mathbf{B}_{k,v}^* - \lambda_k^*\hat{\mathbf{R}}_{k,v}\right)\mathbf{u}_{k,v} \geq 0$ for all vectors $\mathbf{u}_{k,v}$. By selecting $\mathbf{u}_{k,v} = \left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}$, this conditions becomes

$$\left(\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v} - \lambda_k^*\left(\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}\right)^2\right) \geq 0$$
$$\Rightarrow \lambda_k^* \leq \frac{1}{\hat{\mathbf{h}}_{k,v}^H\left(\mathbf{B}_{k,v}^*\right)^{-1}\hat{\mathbf{h}}_{k,v}}. \tag{35}$$

Hence, the equality in (34) can only be achieved for the BSs that have the largest value on $\hat{\mathbf{h}}_{k,v}^H(\mathbf{B}_{k,v}^*)^{-1}\hat{\mathbf{h}}_{k,v}$. This can be one or multiple BSs, as reflected by the theorem. □

Theorem 2 proves that single-BS association is optimal in most cases, although our system model supports spatial multiflow transmission from multiple BSs (a similar result was obtained in [19] in for single-antenna BSs). The optimal BS association for user k is the one with the largest value of $\hat{\mathbf{h}}_{k,v}^H(\mathbf{B}_{k,v}^*)^{-1}\hat{\mathbf{h}}_{k,v}$. We notice that $\mathbf{B}_{k,v}^*$ in (26) is the weighted sum of several terms; the spatial directions of interfering channels, the noise variance, the channel uncertainty, and the matrices from the power constraints. These terms are weighted by the different Lagrange multipliers, which means that the QoS and power constraints that are hard to satisfy will have a large impact on $\mathbf{B}_{k,v}^*$ and vice versa. The BS association rule is based on the norm of the channel $\|\hat{\mathbf{h}}_{k,v}^H\|^2$ from BS v, which is then weighted through $\mathbf{B}_{k,v}^*$. The weighing will punish BSs with smaller power budget, lower estimation quality, and/or many users with high QoS targets.

As seen from Case 2 in Theorem 2, it may happen that multiple BSs are associated with a certain user. This occurs when the most appropriate BS does not have the power resources to satisfy the QoS target, thus another BS needs to help out. This result stands in contrast to [25] where single-BS association always occurs since there are no power constraints. The probability of multi-BS association is evaluated in Section V.

The optimal BS association rule is clearly a complicated function of the channel quality, estimation quality, power constraints, and QoS constraints. This stands in contrast to heuristic association rules (e.g., the nearest-BS based, the highest-received-power based, the max-SINR based, the biased-received-power based and the biased-SINR based load balancing criteria), which are generally not optimal in terms of maximizing the energy efficiency under per-BS transmit power constraints and per-user QoS constraints. These heuristic association rules have been studied under various conditions (different from our system model); see for example [6], [7], [14]–[16], [20]–[23], [33]. Hopefully, these heuristics can evolve in future works, based on insights on the optimal BS association from Theorem 2.

IV. Iterative Heuristic Algorithm Design

In this section, we tackle the non-convex problem (9) by iterative convex approximations of the power consumption functions. In particular, each iteration solves a problem with a modified objective function, which is convex. This convex objective function is updated in each iteration such that most of the BSs with small transmit powers in the solution are driven to sleep mode. The proposed algorithm will find a suboptimum to the original problem in (9).

Note that $0 \leq P_{\text{trans},v} \leq P_{v,\max}$ for each BS v, $v \in \mathcal{V}$. Thus, the total consumed power of BS v, P_v in (5), can be relaxed with its convex envelope, $P_v^{\text{c.e.}}$ over the interval $[0, P_{v,\max}]$, where

$$P_v^{\text{c.e.}}(P_{t,v}) \triangleq N_v P_{\text{sleep},v} + \Delta_v' P_{\text{trans},v} \tag{36}$$

with

$$\Delta_v' \triangleq \frac{N_v(P_{\text{active},v} - P_{\text{sleep},v})}{P_{v,\max}} + \Delta_v \tag{37}$$

which is the largest convex function smaller than or equal to P_v over the interval. Replacing P_v with $P_v^{\text{c.e.}}$, problem (9) and (10) are relaxed to

$$\begin{aligned} \underset{\{\mathbf{w}_{k,v}\}}{\text{minimize}} \quad & \sum_{v=1}^{M} a_v P_v^{\text{c.e.}} \\ \text{subject to} \quad & \gamma_k \geq \Gamma_k, \qquad \forall k \\ & \sum_{k=1}^{K} \|\mathbf{w}_{k,v}\|^2 \leq P_{v,\max}, \qquad \forall v. \end{aligned} \tag{38}$$

The idea, which is based on replacing an indicator function of a bounded variable with its convex envelope, is often referred to as the l_1-norm relaxation, where sparse solutions can be obtained. The relaxed problem (38) can be reformulated as a convex optimization problem

$$\begin{aligned}\underset{\{\mathbf{W}_k\succeq 0\}}{\text{minimize}} \quad & \sum_{k=1}^{K}\text{Tr}(\mathbf{A}'\mathbf{W}_k)+\sum_{v=1}^{M}a_vN_vP_{\text{sleep},v}\\ \text{subject to} \quad & \text{Tr}(\hat{\mathbf{R}}_k\mathbf{W}_k)-\Gamma_k\sum_{\substack{l=1\\ l\neq k}}^{K}\text{Tr}\left((\hat{\mathbf{R}}_k+\mathbf{E}_k)\mathbf{W}_l\right)\\ & -\Gamma_k\text{Tr}(\mathbf{E}_k\mathbf{W}_k)\geq\Gamma_k\sigma_k^2, \quad \forall\, k\\ & \sum_{k=1}^{K}\text{Tr}(\boldsymbol{Q}_v\mathbf{W}_k)\leq P_{v,\max}, \quad \forall\, v \end{aligned} \tag{39}$$

where $\mathbf{A}'$ is a modified block diagonal matrix of $\mathbf{A}$, with Δ_v replaced by Δ_v' for each block v. Note that based on Lemma 3, the rank-one constraints are dropped without loss of optimality. Compared to the original problem (10), the relaxed problem (39) has the same feasible set, but a modified objective function. The optimal value of (39) is a lower bound on the optimal value of the original problem (10).

The proposed iterative heuristic algorithm is as follows:

1) $i:=0$; Initialize $\mathbf{W}_k^{(0)}$ for $k=1,\ldots,K$ by solving the relaxed convex problem (39).
2) $i:=i+1$; Obtain the transmit power of each BS v as $P_{\text{trans},v}^{(i-1)}=\sum_{k=1}^{K}\text{Tr}(\boldsymbol{Q}_v\mathbf{W}_k^{(i-1)})$.
   Define $\hat{P}_v^{(i)}(P_{\text{trans},v})\triangleq N_vP_{\text{sleep},v}+\Delta_v^{(i)}P_{\text{trans},v}$, where
$$\Delta_v^{(i)}\triangleq\frac{N_v(P_{\text{active},v}-P_{\text{sleep},v})}{P_{\text{trans},v}^{(i-1)}+\delta}+\Delta_v. \tag{40}$$
   Solve the modified optimization problem
$$\begin{aligned}\underset{\{\mathbf{W}_k\succeq 0\}}{\text{minimize}} \quad & \sum_{k=1}^{K}\text{Tr}\left(\mathbf{A}^{(i)}\mathbf{W}_k\right)+\sum_{v=1}^{M}a_vN_vP_{\text{sleep},v}\\ \text{subject to} \quad & \text{Tr}(\hat{\mathbf{R}}_k\mathbf{W}_k)-\Gamma_k\sum_{\substack{l=1\\ l\neq k}}^{K}\text{Tr}\left((\hat{\mathbf{R}}_k+\mathbf{E}_k)\mathbf{W}_l\right)\\ & -\Gamma_k\text{Tr}(\mathbf{E}_k\mathbf{W}_k)\geq\Gamma_k\sigma_k^2, \quad \forall\, k\\ & \sum_{k=1}^{K}\text{Tr}(\boldsymbol{Q}_v\mathbf{W}_k)\leq P_{v,\max}, \quad \forall\, v \end{aligned} \tag{41}$$
   where $\mathbf{A}^{(i)}$ is the modified block diagonal matrix of $\mathbf{A}$, with Δ_v replaced by $\Delta_v^{(i)}$ for each block v.
3) Let $\mathbf{W}_k^{(i)}$ be the solution to this problem.
4) If $P_{\text{trans},v}^{(i-1)}$ and $P_{\text{trans},v}^{(i)}$ are approximately[5] equal for each v, return $\mathbf{W}_k^*:=\mathbf{W}_k^{(i)}$.
   Otherwise, go back to step 2).

[5]There are many different ways to define "approximately equal," such as $\max_v|P_{\text{trans},v}^{(i-1)}-P_{\text{trans},v}^{(i)}|\leq\varepsilon$ and $\sum_{v=1}^{M}|P_{\text{trans},v}^{(i-1)}-P_{\text{trans},v}^{(i)}|\leq\varepsilon$. The latter is used as a stopping criterion in our simulation with $\varepsilon=10^{-6}$.

Note that δ in (40) is a non-negative small value, which can be interpreted as a soft threshold for deciding when a BS is set to sleep mode. Define $P_{\text{trans},v}^*\triangleq\sum_{k=1}^{K}\text{Tr}(\boldsymbol{Q}_v\mathbf{W}_k^*)$. Thus, for $P_{\text{trans},v}^*\gg\delta$, we have $\hat{P}_v(P_{\text{trans},v}^*)\triangleq N_vP_{\text{sleep},v}+\left(\frac{N_v(P_{\text{active},v}-P_{\text{sleep},v})}{P_{\text{trans},v}^*+\delta}+\Delta_v\right)P_{\text{trans},v}^*\approx N_vP_{\text{active},v}+\Delta_vP_{\text{trans},v}^*=P_v(P_{\text{trans},v}^*)$, and BS v is in the active mode; while for $P_{\text{trans},v}^*=0$, $\hat{P}_v(P_{\text{trans},v}^*)\triangleq N_vP_{\text{sleep},v}$ and BS v is under the sleep mode.

For each iteration as shown in step 2), when $P_{\text{trans},v}^{(i-1)}$ is small, the modified $\Delta_v^{(i)}$ in (40) becomes large, i.e., the derivative of the power consumption function $\hat{P}_v^{(i)}(P_{\text{trans},v})$ increases. Therefore, the modified optimization problem (41) will push the small $P_{\text{trans},v}^{(i-1)}$ to zero; that is, the BSs with small transmit powers in the solution to the previous problem are driven to sleep mode. This leads to sparse solutions of $\mathbf{W}_k^*$.

Lemma 5: The proposed iterative heuristic algorithm always converges.

Proof: The objective function of problem (41) is on the form of the objective function in [50, Eq. (21)], which always gives convergence; that is, with $0\leq P_{\text{trans},v}\leq P_{v,\max}$ a convex, compact set, and $\delta>0$, we can show that $P_{\text{trans},v}^{(i)}-P_{\text{trans},v}^{(i-1)}\to 0$ for $v=1,\ldots,M$. A proof of convergence for this type of heuristic algorithms is given in [50, Appendix B]. □

Note that, upon convergence, the partial derivative with respect to $P_{\text{trans},v}$ of the function minimized in the last iteration is given by

$$\frac{N_v(P_{\text{active},v}-P_{\text{sleep},v})}{P_{\text{trans},v}^*+\delta}+\Delta_v, \tag{42}$$

which is equal to the derivative of the function

$$f(P_{\text{trans},v})=\sum_{v=1}^{M}\alpha_v\log(P_{\text{trans},v}+\delta)+\sum_{v=1}^{M}\Delta_vP_{\text{trans},v} \tag{43}$$

at $P_{\text{trans},v}=P_{\text{trans},v}^*$, where $\alpha_v\triangleq N_v(P_{\text{active},v}-P_{\text{sleep},v})$. From the equality of the first-order conditions for optimality, we see that the iterative procedure finds a local minimum of $f(P_{\text{trans},v})$. The log-sum function $\sum_{v=1}^{M}\alpha_v\log(P_{\text{trans},v}+\delta)$ is used as a smooth surrogate for the circuit power consumption part of the objective function. Therefore, our proposed heuristic algorithm is very similar to the weighted l_1-norm minimization methods, where the weighting factors are chosen based on the log-sum surrogate function of the l_0-norm [56].

V. Numerical Results

Numerical results are presented in this section to illustrate our analytical results and the proposed algorithms. The purpose of this section is not to provide a large-system analysis, but to compare the heuristic algorithm from Section IV with the optimal solution from Theorem 1, for which the complexity of mode selection grows quickly with the number of BSs.

The propagation environment is a simplified version of the dense urban information society model (TC2) used in the METIS project [57], as illustrated in Fig. 2. The model consists of four square-shaped buildings of dimensions 120 × 120 m, each with 6 floors. A macro BS (MBS) is complemented with 4

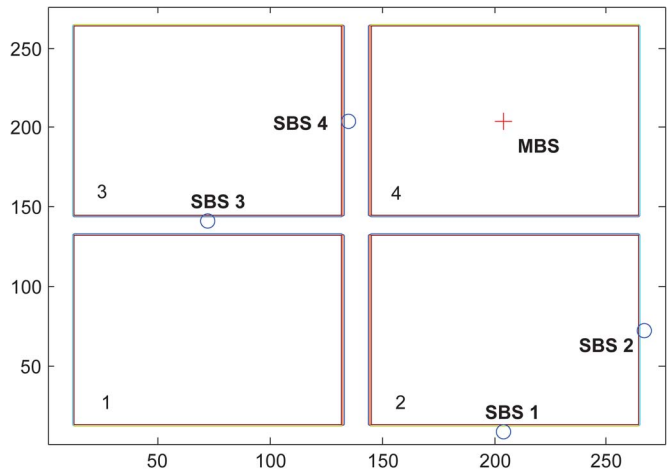

Fig. 2. The MBS (cross) and SBSs (circles) deployment considered in Section V.

small cell BSs (SBSs). The MBS has 4 transmit antennas, and the SBSs have 2 transmit antennas each. Load-balancing is particularly important in the lightly loaded cases that occur during the majority of the day [9], because then there is an opportunity to turn off BSs and associate users with other BSs than the closest one. Hence, in most of the simulations, we consider five users that are randomly and uniformly dropped in the network, whereof 4 users are indoors and 1 user is outdoors in every user drop. The system bandwidth is 10 MHz. Here, we adopt the indoor and outdoor propagation models, PS#1–PS#4, identified in METIS. More details regarding network deployment and propagation modes can be found in [57, Table 3.7 and Section 8.1]. We assume independent Rayleigh small-scale fading. The MMSE channel estimation errors are calculated based on (1) with the total pilot power $p = P_{v,\max}/2$. Table I shows the power model parameters and is based on [11, Tables 6 and 8].

Three different joint precoding and load balancing schemes are compared in the scenario depicted in Fig. 2. We name these three schemes as "Optimal," "Heuristic" and "All Active" respectively. The "Optimal" scheme obtains the global optimal solution as described in Theorem 1, by an exhaustive search over all 2^5 possible BS mode combinations. The "Heuristic" scheme follows the algorithm proposed in Section IV, and the value of the soft threshold δ is set to 10^{-4}. The "All Active" scheme is used as our performance baseline, which solves the optimization problem (9) by assuming that all BSs are active, i.e., the BS mode indicator $z_v = 1$ for all BSs $v \in \mathcal{V}$. For each scheme, the performance is averaged over 1000 independent user drops that provide feasible solutions for our optimization problem (9). For each user drop, the algorithms are evaluated over 50 independent channel realizations. The weights a_v are set to 1 for all BSs.

Define the dynamic part of total power consumption as the total RF power ($\sum_{v=1}^{M} a_v \Delta_v P_{\text{trans},v}$), and the remaining part of the total power consumption as the circuit power ($\sum_v a_v \times N_v P_{\text{active},v} z_v + \sum_v a_v N_v P_{\text{sleep},v}(1 - z_v)$). Figs. 3 and 4 demonstrate the total RF power and the total power consumption as a function of target spectral efficiency per user, respectively. As expected, the total power consumption and the RF power increase as the target spectral efficiency increases. Fig. 3 shows

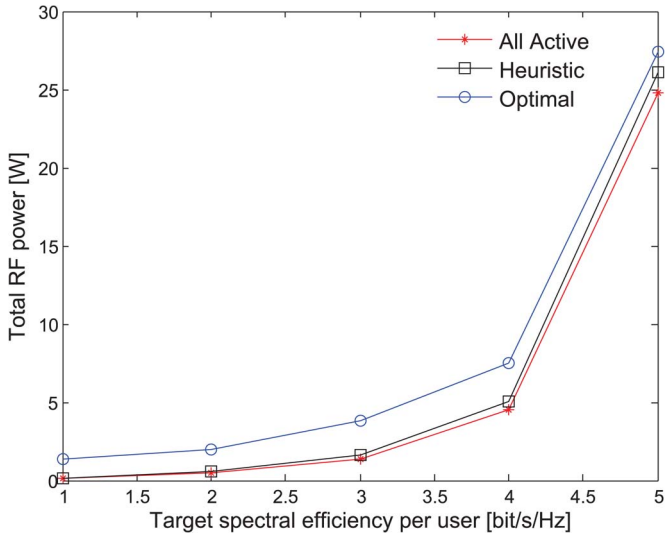

Fig. 3. Total RF power (the dynamic part $\sum_{v=1}^{M} a_v \Delta_v P_{\text{trans},v}$) vs. target spectral efficiency per user (R_k).

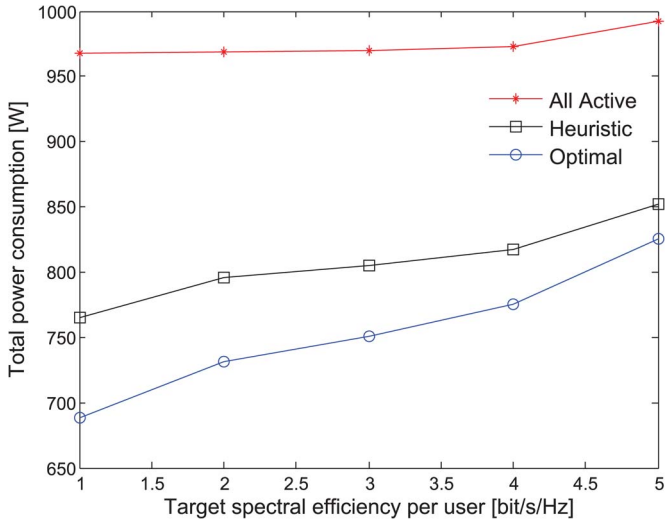

Fig. 4. Total power consumption ($\sum_{v=1}^{M} a_v P_v$) vs. target spectral efficiency per user (R_k).

that the RF power for the "All Active" scheme is less than that of the "Heuristic" and "Optimal" schemes. This is expected since all BSs are active in the "All Active" scheme, whereas for the "Heuristic" and "Optimal" schemes, some BSs are put into sleep mode. With more BSs being active, the "All Active" scheme provides better energy-focusing and less propagation losses between the users and the transmitters, and will therefore reduce the total RF power. However, as can be seen from Fig. 4, compared to the "All Active" scheme, the "Heuristic" and "Optimal" schemes can substantially reduce the total power consumption, especially when the target QoS is small. This is because the circuit power consumption under the sleep mode is much lower compared to the one under the active mode, i.e., $P_{\text{sleep},v} \ll P_{\text{active},v}$. For the "All Active" scheme, the increase in the circuit part from the extra power consumed by activating BSs clearly outweighs the decrease in the dynamic part. This implies that putting a BS into sleep mode by proper load balancing is an important solution for energy savings in heterogeneous networks.

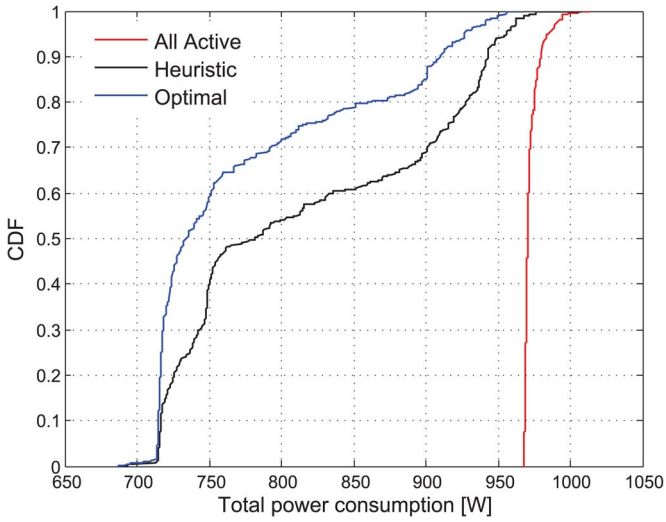

Fig. 5. The CDF of total power consumption.

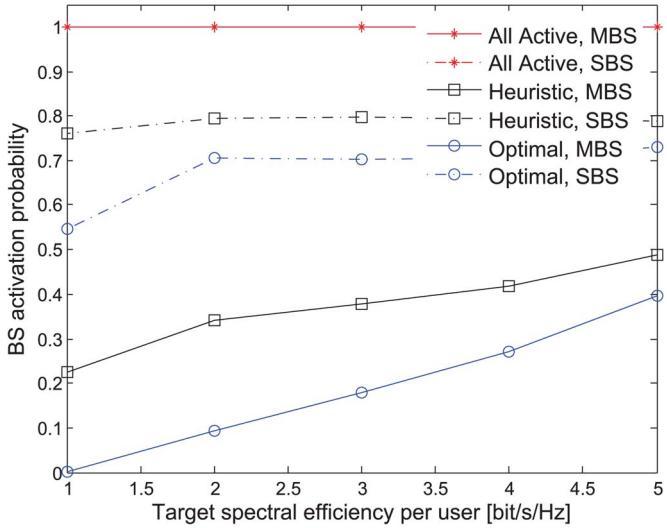

Fig. 6. BS activation probability vs. target spectral efficiency per user (R_k).

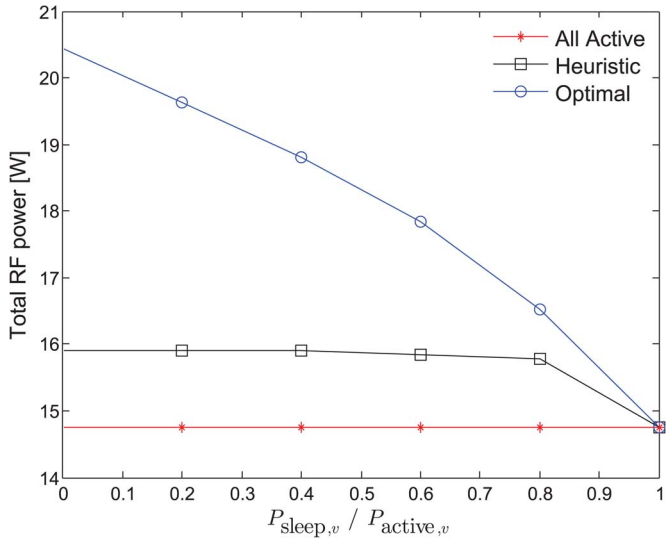

Fig. 7. Total RF power vs. $\eta = P_{\text{sleep},v}/P_{\text{active},v}$.

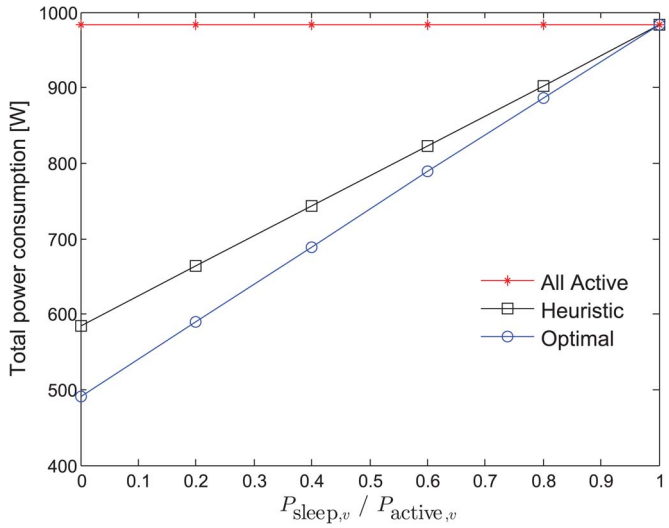

Fig. 8. Total power consumption vs. $\eta = P_{\text{sleep},v}/P_{\text{active},v}$. The target spectral efficiency per user R_k is 3 bit/s/Hz.

Fig. 5 plots the cumulative distribution function (CDF) of the total power consumption for the considered three schemes. The target spectral efficiency per user R_k is 4 bit/s/Hz. We observe that compared to the "All Active" scheme, 20% of the total power consumption can be saved by the "Optimal" scheme with 70% probability and by the "Heuristic" scheme with 55% probability. For some user drops, the energy consumption can be reduced by 30% for both the "Optimal" and "Heuristic" schemes.

Fig. 6 demonstrates the BS activation probability versus the target spectral efficiency per user. Here, the activation probability of the SBS is averaged over the probabilities of the four SBSs depicted in Fig. 2. We see that for the "All Active" scheme, the activation probabilities of the MBS and SBS are always one, since all BSs are always active in this scheme. Moreover, as anticipated, for both the "Heuristic" and "Optimal" schemes, the BS activation probabilities of the MBS and SBS increase as the target spectral efficiency per user increases. This is because in order to satisfy the raised QoS expectations of all users, the probability that a BS becomes active should increase so as to provide better energy-focusing and less propagation losses. Over the considered range of target spectral efficiency per user, the "Optimal" scheme has lower activation probability for the MBS and higher activation probability for the SBS as compared to the "Heuristic" scheme. Note that the circuit power consumed under the active mode $P_{\text{active},v}$ for the MBS is much higher than that of the SBSs. Thus, as shown in Fig. 4, the "Optimal" scheme results in better energy saving as compared to the "Heuristic" scheme.

Figs. 7–9 investigate the impact of the ratio $\eta \triangleq P_{\text{sleep},v}/P_{\text{active},v}$ on the overall energy efficiency for different schemes. The values of $P_{\text{active},v}$ are fixed to 130 W and 56 W for the MBS and SBSs respectively. The target spectral efficiency R_k is fixed to 3 bit/s/Hz. In Figs. 7 and 8, the total RF power and the total power consumption are plotted as a function of the ratio η, respectively. It can be seen from Fig. 7 that the RF power of the "Optimal" and "Heuristic" schemes decreases as the ratio η (or equivalently $P_{\text{sleep},v}$) increases, especially when the ratio η is large (close to 1). This is because it is better to turn on more BSs, to reduce the RF power, when the difference between the active

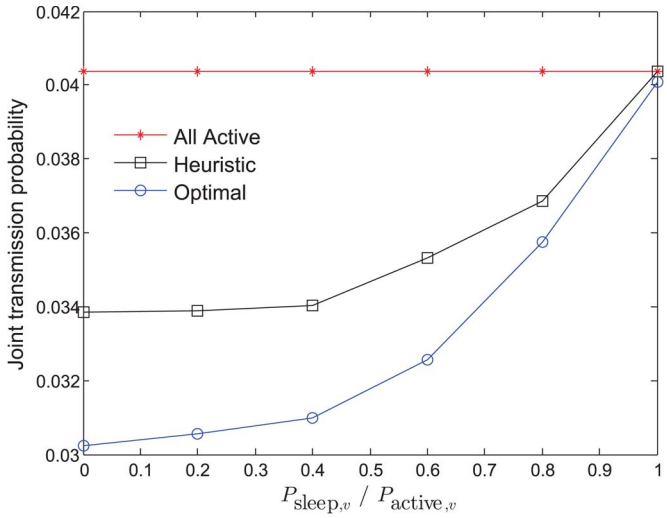

Fig. 9. Joint transmission probability vs. $\eta = P_{\text{sleep},v}/P_{\text{active},v}$. The target spectral efficiency per user R_k is 3 bit/s/Hz.

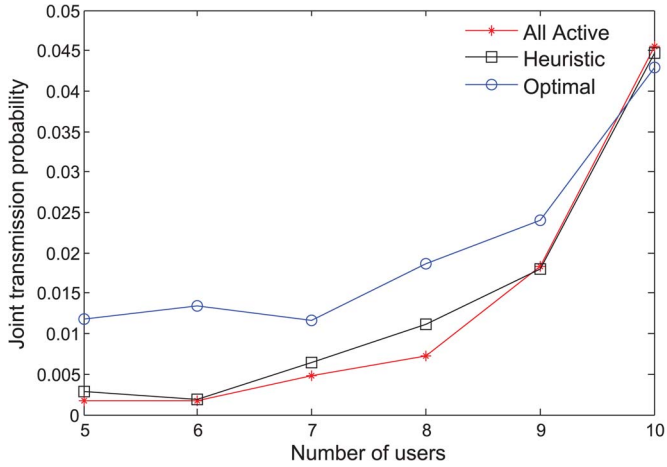

Fig. 10. Joint transmission probability vs. the number of users. The target spectral efficiency per user R_k is 1 bit/s/Hz.

and sleep modes decreases. The BS activation probability increases more for the "Optimal" scheme, compared to the "Heuristic" scheme. Hence, we observe that the total RF power reduces more significantly for the "Optimal" scheme. From Fig. 8, we see that the total power consumption increases almost linearly as η increases. This is mainly due to the increase of $P_{\text{sleep},v}$.

Although the system allows all BSs to transmit to all users simultaneously at the same time-frequency resource block, Fig. 9 shows that the probability that a user is served by multiple BSs is less than 4.2% for all the considered schemes over the entire range of η when the target spectral efficiency R_k is fixed to 3 bit/s/Hz. Not shown here, the joint transmission probability has also been evaluated over different targets of spectral efficiency, i.e., for $R_k = \{1, 2, 3, 4, 5\}$ bits/s/Hz, while the ratio η is fixed according to Table I. For these cases, simulation shows that the probability of multi-BS joint transmission is less than 4% over the considered range of R_k. Fig. 10 shows the joint transmission probability as function of the number of users, for a target spectral efficiency of 1 bit/s/Hz. The probability increases with the number of users, since it is harder to satisfy the QoS targets, but it is still in the range of a few percentages. These observations are in line with Theorem 2. From Fig. 9, we also observe that, for the "Optimal" and "Heuristic" schemes, the joint transmission probability increases as the ratio η increases. This is expected since by increasing η, the BS activation probability increases. Thus, the joint transmission probability also increases. Compared to the "Heuristic" algorithm, the "Optimal" scheme has a lower BS activation probabilities, and therefore it also has a lower joint transmission probability.

VI. Conclusion

This paper analyzed the energy efficiency in heterogeneous networks. More specifically, the downlink precoding vectors, load balancing (i.e., user-BS association), and BS operational modes were jointly optimized to minimize the weighted total power consumption. In order to verify how many BSs that should serve a user at the optimal load balancing solution, each user can be served by multiple BSs using spatial multiflow transmission. We proved that the optimal BS association rule consists of comparing weighted channel norms, where the weighting matrices depend on channel uncertainty, power constraints and QoS constraints. Moreover we proved that, in most cases, it is optimal for each user to be served by a single BS. Multiple BSs only serve a user when the primary BS does not have the power resources to deliver the full QoS, in which case neighboring BSs can cooperate in order to provide the full QoS. An iterative heuristic algorithm was proposed to find a suboptimal solution of relatively low complexity and it achieves good performance in relation to the optimal scheme. Our numerical results showed that the total power consumption can be greatly reduced by putting a BS into sleep mode using proper load balancing.

References

[1] J. Li, E. Björnson, T. Svensson, T. Eriksson, and M. Debbah, "Optimal design of energy-efficient HetNets: Joint precoding and load balancing," in *Proc. IEEE ICC*, Jun. 2015, pp. 1–6.
[2] Y. Chen *et al.*, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.
[3] Z. Hasan *et al.*, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, Nov. 2011.
[4] E. Björnson, E. Jorswieck, M. Debbah, and B. Ottersten, "Multi-objective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Process. Mag., Special Issue Signal Process. 5G Revolution*, vol. 31, no. 6, pp. 14–23, Nov. 2014.
[5] "Scenarios, requirements and KPIs for 5G mobile and wireless system," Mobile wireless commun. Enablers Twenty-twenty Inf. Soc. (METIS), Sweden, EU FP7 INFSO-ICT-317669 METIS, D1.1, Apr. 2013.
[6] B. Bjerke, "LTE-advanced and the evolution of LTE deployments," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 4–5, Oct. 2011.
[7] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2012.
[8] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.
[9] G. Auer *et al.*, "Cellular energy efficiency evaluation framework," in *Proc. IEEE VTC*, May 2011, pp. 1–6.
[10] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
[11] "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," Energy Aware Radio Netw. Technol., Tokyo, Japan, EU FP7 INFSO-ICT-247733 EARTH, D2.3, Jan. 2012.
[12] E. Björnson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proc. IEEE ICT*, 2013, pp. 1–5.

[13] J. Li, M. Matthaiou, S. Jin, and T. Svensson, "Energy efficiency analysis of rank-1 Ricean fading MIMO channels," in *Proc. IEEE Int. Workshop SPAWC*, Jun. 2014, pp. 349–353.
[14] S. Singh, F. Baccelli, and J. Andrews, "On association cells in random heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 1, pp. 70–73, Feb. 2014.
[15] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
[16] P. Madhusudhanan, J. Restrepo, Y. Liu, and T. Brown, "Downlink coverage analysis in a heterogeneous cellular network," in *Proc. IEEE GLOBECOM*, Dec. 2012, pp. 4170–4175.
[17] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.
[18] S. Singh and J. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.
[19] Q. Ye *et al.*, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[20] A. K. Gupta, H. S. Dhillon, S. Vishwanath, and J. G. Andrews, "Downlink multi-antenna heterogeneous cellular network with load balancing," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 18–25, Nov. 2014.
[21] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.
[22] Y. Lin and W. Yu, "Optimizing user association and frequency reuse for heterogeneous network under stochastic model," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 2045–2050.
[23] H. Dhillon, M. Kountouris, and J. Andrews, "Downlink MIMO HetNets: Modeling, ordering results and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5208–5222, Oct. 2013.
[24] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations Trends Commun. Inf. Theory*, vol. 9, no. 2/3, pp. 113–381, 2013.
[25] R. Stridh, M. Bengtsson, and B. Ottersten, "System evaluation of optimal downlink beamforming with congestion control in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 743–751, Apr. 2006.
[26] R. Sun and Z.-Q. Luo, "Globally optimal joint uplink base station association and power control for max-min fairness," in *Proc. IEEE ICASSP*, May 2014, pp. 454–458.
[27] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1040–1054, Jun. 2015.
[28] M. Hong, R. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.
[29] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3214–3228, Jun. 2013.
[30] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, "Optimal joint base station assignment and beamforming for heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 8, pp. 1950–1961, Apr. 2014.
[31] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," IEEE Trans. Signal Process., vol. 62, no. 15, pp. 3939–3952, Aug. 2014.
[32] D. Ng, E. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.
[33] Y. S. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.
[34] J. S. Arora, M. W. Huang, and C. C. Hsieh, "Methods for optimization of nonlinear problems with discrete variables: A review," *Structural Optim.*, vol. 8, no. 2/3, pp. 69–85, Oct. 1994.
[35] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statist. Soc.: Ser. B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.
[36] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[37] J. Zhao, T. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.
[38] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network MIMO system via reweighted power minimization," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 1962–1967.
[39] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.
[40] Y. Shi, J. Zhang, and K. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.
[41] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.
[42] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.
[43] M. Biguesh and A. B. Gershman, "Downlink channel estimation in cellular systems with antenna arrays at base stations using channel probing with feedback," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 9, pp. 1330–1339, 2004.
[44] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1807–1820, Mar. 2010.
[45] T. Wild, "Comparing downlink coordinated multi-point schemes with imperfect channel knowledge," in *Proc. IEEE VTC*, Sep. 2011, pp. 1–5.
[46] H. Holma and A. Toskala, *LTE Advanced: 3GPP Solution for IMT-Advanced*, 1st ed. Hoboken, NJ, USA: Wiley, 2012.
[47] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[48] M. Xu, D. Guo, and M. L. Honig, "Two-cell downlink noncoherent cooperation without transmitter phase alignment," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.
[49] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–9446, May 2000.
[50] M. S. Lobo, M. Fazel, and S. Boyd, "Portfolio optimization with linear and fixed transaction costs," *Ann. Oper. Res.*, vol. 152, no. 1, pp. 341–365, Jul. 2007.
[51] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Eds. Boca Raton, FL, USA: CRC Press, 2001.
[52] E. Björnson, N. Jalden, M. Bengtsson, and B. Ottersten, "Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6086–6101, Dec. 2011.
[53] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0 Beta," Sep. 2012.
[54] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. IEEE CACSD*, Sep. 2004, pp. 284–289.
[55] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[56] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted *l*-1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[57] "Simulation guidelines," Mobile wireless commun. Enablers Twenty-twenty Inf. Soc. (METIS), Sweden, EU FP7 INFSO-ICT-317669 METIS, D6.1, Oct. 2013.

Jingya Li (S'11) received the B.E. degree in communication engineering and the Master's degree in communication and information systems from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008 and 2010, respectively, and the Ph.D. degree in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 2015. Since August 2014, she has been with Ericsson Research, Ericsson AB, Gothenburg, where she is working as an Experienced Researcher. Her research interests are in the broad areas of cooperative wireless communications, with a focus on radio resource optimization, interference management, and hardware-constrained communications.

Dr. Li was a recipient of the Best Paper Award at the 2015 IEEE International Conference on Communications (ICC), London, U.K.

Emil Björnson (S'07–M'12) received the M.S. degree in engineering mathematics from Lund University, Lund, Sweden, in 2007 and the Ph.D. degree in telecommunications from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2011. From 2012 to July 2014, he was a Joint Postdoc at Suplec, Gif-sur-Yvette, France, and at KTH Royal Institute of Technology. He is currently an Assistant Professor and Docent at the Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden.

He is the first author of the textbook *Optimal Resource Allocation in Coordinated Multi-Cell System* (Foundations and Trends in Communications and Information Theory, 2013). He is also dedicated to reproducible research and has made a large amount of simulation code publicly available. His research interests include multi-antenna cellular communications, radio resource allocation, energy efficiency, massive MIMO, and network topology design. He was a recipient of the 2014 Outstanding Young Researcher Award from IEEE ComSoc EMEA, the 2015 Ingvar Carlsson Award, and five best paper awards for novel research on multi-cell multi-antenna communications: ICC 2015, WCNC 2014, SAM 2014, CAMSAP 2011, and WCSP 2009.

Tommy Svensson (S'98–M'03–SM'10) received the Ph.D. degree in information theory from Chalmers in 2003. He is currently an Associate Professor in communication systems with Chalmers University of Technology, Gothenburg, Sweden, where he is leading the research on air interface and wireless backhaul networking technologies for future wireless systems. He was with Ericsson AB where he worked on core networks, radio access networks, and microwave transmission products. He was involved in the European WINNER and ARTIST4G projects that made important contributions to the 3GPP LTE standards and the recently finished EU FP7 METIS and the emerging EU H2020 mm MAGIC projects targeting solutions for 5G. He has coauthored two books and more than 110 journal and conference papers. His main research interests are in the design and analysis of physical-layer algorithms, multiple access, resource allocation, cooperative systems, moving networks and satellite networks. He is the Chairman of the IEEE Sweden joint Vehicular Technology/Communications/Information Theory Societies chapter and coordinator of the Communication Engineering Master's Program at Chalmers.

Thomas Eriksson received the Ph.D. degree in information theory from Chalmers University of Technology, Gothenburg, Sweden, in 1996. In 1997 and 1998, he was with AT&T Labs—Research, Murray Hill, NJ, USA, and in 1998 and 1999, he was with Ericsson Radio Systems AB, Kista, Sweden. Since 1999, he has been with Chalmers University of Technology, where he is a Professor of communication systems. Furthermore, he was a Guest Professor at Yonsei University, Korea, in 2003–2004. He is currently the Vice Head of the Department of Signals and Systems, Chalmers University of Technology, with responsibility for undergraduate and master education. He has authored and co-authored more than 200 journal and conference papers and is the holder of seven patents. His research interests include communication, data compression, and modeling and compensation of non-ideal hardware components (e.g., amplifiers, oscillators, and modulators in communication transmitters and receivers, including massive MIMO).

Mérouane Debbah (F'15) received the M.Sc. and Ph.D. degrees from the Ecole Normale Supérieure de Cachan, Cachan, France. He worked for Motorola Labs, Saclay, France, from 1999 to 2002 and the Vienna Research Center for Telecommunications, Vienna, Austria, until 2003. From 2003 to 2007, he was with the Mobile Communications Department, Institut Eurecom, Sophia Antipolis, France, as an Assistant Professor. Since 2007, he has been a Full Professor at CentraleSupelec, Gif-sur-Yvette, France. From 2007 to 2014, he was the Director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he has been the Vice-President of the Huawei France R&D Center and the Director of the Mathematical and Algorithmic Sciences Lab. His research interests include fundamental mathematics, algorithms, complex systems analysis and optimization, statistics, information, and communication sciences research. He is a WWRF Fellow and a member of the Academic Senate of Paris-Saclay. He is an Associate Editor-in-Chief of the journal *Random Matrix: Theory and Applications* and was an Associate and Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, respectively, in 2011–2013 and 2013–2014. He was a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He was the recipient of the Mario Boella Award in 2005, the 2007 IEEE GLOBECOM Best Paper Award, the Wi-Opt 2009 Best Paper Award, the 2010 Newcom++ Best Paper Award, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC Best Paper Award, the 2015 ICC Best Paper Award, as well as the Valuetools 2007, Valuetools 2008, CrownCom2009, Valuetools 2012, and SAM 2014 Best Student Paper awards. He received the IEEE Glavieux Prize Award in 2011 and the Qualcomm Innovation Prize Award in 2012. In 2015, he received jointly the 2015 IEEE Communications Society Leonard G. Abraham Prize and the 2015 IEEE Communications Society Fred W. Ellersick Prize.