Flexible Multiple Base Station Association and Activation for Downlink Heterogeneous Networks

Kaiming Shen, *Student Member, IEEE*, Ya-Feng Liu, *Member, IEEE*, David Yiwei Ding, and Wei Yu, *Fellow, IEEE*

Abstract—This letter shows that the flexible association of possibly multiple base stations (BSs) with each user over multiple frequency bands, along with the joint optimization of BS transmit power that encourages the BSs to turn off at off-peak time, can significantly improve the performance of a downlink heterogeneous wireless cellular network. We propose a gradient projection algorithm for optimizing BS association and an iteratively reweighting scheme together with a novel proximal gradient method for optimizing power in order to find the optimal tradeoff between network utility and power consumption. Simulation results reveal significant performance improvement as compared to the conventional single-BS association.

Index Terms—Base station (BS) activation, BS association, group sparsity, proximal gradient method.

I. INTRODUCTION

HETEROGENEOUS network (HetNet), in which the pico-base stations (BSs) are deployed throughout the network to offload traffic from macro-BSs, is a key technique for densifying the BS deployments, thereby improving the throughput of wireless cellular network. A central question in the design of HetNet is the joint optimization of cell loading (i.e., user-BS association), transmit power, and frequency partitioning across the pico- and macro-BSs. In contrast to conventional HetNet design in which each user is associated with a single BS, this letter proposes a novel network utility maximization formulation that allows each user to associate with *multiple* BSs over multiple frequency bands.

The optimization of user-BS association is inextricably related to power control. Toward this end, this letter further formulates a BS transmit power spectrum optimization problem by incorporating a power consumption model that encourages turning off of BSs at off-peak time and illustrates that the proposed flexible multiple-BS association and BS activation scheme that takes into account signal strength, interference management, frequency allocation, and load balancing can achieve the right balance between network utility and power consumption in HetNets—although at a cost of larger overhead for coordinated scheduling and power control across the BSs.

This letter solves the proposed network optimization problem from a group sparse optimization perspective and uses block coordinate ascent to iteratively optimize user-BS association and BS transmit power. We show that the optimization of user-BS association across the frequencies can be formulated as a convex optimization under fixed power, and further propose an efficient gradient projection implementation based on its Lagrangian dual. Moreover, we propose a *proximal gradient method* for power optimization with a nonsmooth power penalty. The proposed approach is faster than greedy heuristics yet achieves comparable performance.

Wireless cellular networks are traditionally designed with each user assigned to the BS with the strongest signal, and the BS schedules all associated users according to some system objective. It is well known, however, that association based on maximum signal-to-interference-plus-noise ratio (SINR) may not be optimal, as it typically underutilizes pico-BSs with lower transmit powers. The optimal user-BS association problem has been treated from global optimization [1], [2] and game theoretic [3] perspectives. The most common heuristic used in practice is *cell-range expansion* [4], [5], which can be interpreted as a dual optimization method that improves upon max-SINR association by taking cell-loading into account [6], [7]. The main observation that underpins the above dual optimization interpretation is that giving equal time/bandwidth to the multiple users associated with a single BS is optimal for proportional fairness [6]. This observation is true, however, only when each user is associated with only one BS and all BSs transmit at fixed and flat power spectral density (PSD) across the frequencies. The main point of this letter is that providing the BSs with the flexibility of varying PSDs across the frequencies and giving the users the flexibility of associating with possibly multiple BSs over multiple frequency bands can significantly improve the network utility, albeit at the cost of larger scheduling overhead. Furthermore, significant power saving can be obtained assuming a realistic power consumption model.

K. Shen, D. Y. Ding, and W. Yu are with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: kshen@comm.utoronto.ca; davidyding19@gmail.com; weiyu@comm.utoronto.ca).

Y.-F. Liu is with the State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yafliu@lsec.cc.ac.cn).

II. PROBLEM FORMULATION

Consider a downlink heterogeneous cellular network consisting of K users and L BSs. The total frequency band W is divided into N equal bands. Each user makes a decision at each band which BS(s) it wishes to associate with. Multiple users associated with the same BS in the same band are served in a

time-division multiplex fashion. Let p_ℓ^n be the transmit power at BS ℓ in band n. If user k is served by BS ℓ in band n, then its long-term average spectral efficiency is

$$r_{k\ell}^n = \frac{W}{N} \log \left(1 + \frac{g_{k\ell}^n p_\ell^n}{\sigma^2 + \sum_{\ell' \neq \ell} g_{k\ell'}^n p_{\ell'}^n} \right) \tag{1}$$

where σ^2 is the additive noise power, and $g_{k\ell}^n$ is the long-term channel magnitude from BS ℓ to user k in band n.

The BS power consumption is modeled as comprised of two parts: the transmit power p_ℓ^n, and an *ON-power* ψ_ℓ, which incentivizes BS turning-off during off-peak time:

$$Q_\ell(\mathbf{p}_\ell) = \mathbf{e}_N^T \mathbf{p}_\ell + \psi_\ell \|\mathbf{p}_\ell\|_0 \tag{2}$$

where $\mathbf{p}_\ell = [p_\ell^1, \dots, p_\ell^N]^T$, $\mathbf{e}_N$ denotes the all-one (column) vector of length N, and $\|\mathbf{p}_\ell\|_0$ is the zero-norm of vector $\mathbf{p}_\ell$, i.e., $\|\mathbf{p}_\ell\|_0 = 0$ if $\mathbf{p}_\ell = \mathbf{0}$ and $\|\mathbf{p}_\ell\|_0 = 1$ otherwise. This BS power consumption model has been adopted in [8] and [9].

A main novelty of our problem formulation is that we allow each user to be associated with potentially different BSs in different frequency bands. Let $\mathbf{X}^n \in \mathbb{R}^{K \times L}$ denote the association matrix in band n between all users and all BSs. Since multiple users associated with the same BS and in the same band need to be served with time-division multiplexing, we use $x_{k\ell}^n$, the (k, ℓ)th entry of $\mathbf{X}^n$, to denote that the total portion of time user k is served by BS ℓ in band n. Then, the transmission rate of user k is

$$R_k = \sum_{\ell=1}^L \sum_{n=1}^N x_{k\ell}^n r_{k\ell}^n, \ k = 1, 2, \dots, K. \tag{3}$$

Let $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_L] \in \mathbb{R}^{N \times L}$ and let $\mathbf{X}$ denote the collection of $\{\mathbf{X}^n\}_{n=1}^N$. The objective is to maximize a network utility, chosen here as a tradeoff between the BS power consumption and the proportionally fair utility defined as the sum of log of the user long-term rates:

$$\max_{\mathbf{X}, \mathbf{P}} \quad \sum_{k=1}^K \log(R_k) - \lambda \sum_{\ell=1}^L Q_\ell(\mathbf{p}_\ell) \tag{4a}$$

$$\text{s.t.} \quad \mathbf{0} \leq \mathbf{p}_\ell \leq \bar{\mathbf{p}}_\ell, \ \ell = 1, 2, \dots, L \tag{4b}$$

$$\mathbf{X}^n \mathbf{e}_L \leq \mathbf{e}_K, \ n = 1, 2, \dots, N \tag{4c}$$

$$(\mathbf{X}^n)^T \mathbf{e}_K \leq \mathbf{e}_L, \ n = 1, 2, \dots, N \tag{4d}$$

$$\mathbf{X}^n \geq \mathbf{0}, \ n = 1, 2, \dots, N \tag{4e}$$

where $\lambda \geq 0$ is a constant factor for tradeoff purpose, and $\bar{\mathbf{p}}_\ell$ is the power budget of BS ℓ across all bands. In problem (4), $\mathbf{X}^n \mathbf{e}_L \leq \mathbf{e}_K$ enforces that the portion of time that each user is served by all associated BSs should be less than or equal to 1 in each band; $(\mathbf{X}^n)^T \mathbf{e}_K \leq \mathbf{e}_L$ enforces that the portion of time that each BS serves all associated users should also be less than or equal to 1 in each band. It is possible to prove that any $\mathbf{X}^n$ satisfying (4c)–(4e) is always realizable via coordinated scheduling across the BSs over multiple time slots [10]. Note that this formulation differs significantly from prior works on BS activation and BS association [6], [7], [10]–[12] in allowing each user to be served by multiple BSs in different bands. Furthermore, the BSs may vary PSDs in different bands.

Algorithm 1: An Iteratively Reweighting Algorithm for Solving Problem (4).

Step 1. Choose a positive sequence $\{\tau(t)\}$. Set $t = 0$ and $w_\ell(0) = 1$ for all $\ell = 1, 2, \dots, L$.
Step 2. Solve problem (5) with $w_\ell = w_\ell(t), \ell = 1, 2, \dots, L$ to obtain $\mathbf{P}(t)$ and $\mathbf{X}(t)$.
Step 3. Update the weights by

$$w_\ell(t+1) = \frac{1}{\|\mathbf{p}_\ell(t)\|_2 + \tau(t)}, \ \ell = 1, 2, \dots, L, \tag{7}$$

set $t = t + 1$, and go to **Step 2.**

III. PROPOSED ALGORITHMS

A. Iteratively Reweighted Sparse Optimization

Since the power model (2) includes a BS ON-power that incentivizes BS turning-off, it is advantageous in the overall problem formulation (4) to find a group sparse solution $\mathbf{P}$, in which the zero columns correspond to BSs that are turned off. The paper adopts a nonsmooth mixed ℓ_2/ℓ_1 approach with a penalty function $\sum_{\ell=1}^L \|\mathbf{p}_\ell\|_2$ to promote group sparsity [13]–[15]. In this case, (4) can be approximated by the following weighted mixed ℓ_2/ℓ_1 problem:

$$\max_{\mathbf{X}, \mathbf{P}} \quad f(\mathbf{X}, \mathbf{P}) - \lambda \sum_{\ell=1}^L \psi_\ell w_\ell \|\mathbf{p}_\ell\|_2 \tag{5a}$$

$$\text{s.t.} \quad \text{(4b)–(4e)} \tag{5b}$$

where

$$f(\mathbf{X}, \mathbf{P}) = \sum_{k=1}^K \log(R_k) - \lambda \mathbf{e}_L^T \mathbf{P} \mathbf{e}_N \tag{6}$$

and $\{w_\ell\}$ are weights. This leads to the following iteratively reweighting algorithm for solving problem (4) as shown on the top of the page, which consists of solving a sequence of weighted approximation problems (5) where the weights used for the next iteration are computed from the value of the current solution, as in (7) in Algorithm 1.

We remark that as shown in [16], (7) is an efficient and effective way of updating the weights (for the next iteration based on the solution of the current one). The idea is to use smaller weights to penalize BSs with larger $\{\|\mathbf{p}_\ell(t)\|_2\}$. In particular, the weights (7) are chosen to be (approximately) inversely proportional to $\{\|\mathbf{p}_\ell(t)\|_2\}$, and $\{\tau(t)\}$ is introduced to avoid numerical instability. Empirically, it is better to choose $\{\tau(t)\}$ to be a decreasing sequence than simply setting them to be a small constant value. It has been observed in numerical simulations that Algorithm 1 equipped with (7) converges very fast and often terminates within several iterations.

B. Iterative BS Association and Power Optimization

It remains to solve the subproblem (5). We observe that since the variables $\mathbf{X}$ and $\mathbf{P}$ are separable in the constraints of (5), we can use the block coordinate gradient ascent (BCGA) algorithm [17]–[19] to optimize BS association and power iteratively. The proposed BCGA algorithm alternatively updates $\mathbf{X}$ and $\mathbf{P}$, one at a time with the other being fixed. In the following, we propose a gradient projection step to update $\mathbf{X}$ and a proximal gradient

ascent step [20], [21] to update $\mathbf{P}$ followed by a projection step. We use s to denote the iteration index in the BCGA algorithm and $(\mathbf{X}(s), \mathbf{P}(s))$ to denote the iterates at the sth iteration.

1) Update $\mathbf{X}$ by Gradient Projection: At the $(s + 1)$th iteration, the optimization of $\mathbf{X}$ with $\mathbf{P}$ fixed as $\mathbf{P}(s)$ is a convex optimization, because the objective of (5) is logarithm of a linear function of $\mathbf{X}$ and the constraints are linear. We propose gradient projection as an efficient method for optimizing $\mathbf{X}$. The gradient update is

$$\tilde{\mathbf{X}}^n(s) = \mathbf{X}^n(s) + \alpha^n(s)\nabla_{\mathbf{X}^n} f(\mathbf{X}(s), \mathbf{P}(s)) \qquad (8)$$

where $\alpha^n(s)$ is some appropriately chosen step size.

The projection step is separable among different bands and in each of band $n = 1, 2, \ldots, N$ takes the following form:

$$\min_{\mathbf{X}} \quad \frac{1}{2}\|\mathbf{X}^n - \tilde{\mathbf{X}}^n(s)\|_F^2 \qquad (9a)$$

$$\text{s.t.} \quad (4c)\text{–}(4e) \qquad (9b)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The projection step is also a convex optimization problem, but instead of solving it directly, we show that solving its Lagrangian dual is much easier. Let $\mathbf{y}^n$ and $\mathbf{z}^n$ be the Lagrange multipliers associated with the linear inequality constraints $\mathbf{X}^n\mathbf{e}_L \le \mathbf{e}_K$ and $(\mathbf{X}^n)^T\mathbf{e}_K \le \mathbf{e}_L$, respectively. The Lagrangian dual of problem (9) can be written as

$$\min_{\mathbf{y}^n, \mathbf{z}^n} \quad \frac{1}{2}\|\Theta_s(\mathbf{y}^n, \mathbf{z}^n)\|_F^2 + \mathbf{e}_K^T\mathbf{y}^n + \mathbf{e}_L^T\mathbf{z}^n \qquad (10a)$$

$$\text{s.t.} \quad \mathbf{y}^n \le \mathbf{0} \qquad (10b)$$

$$\mathbf{z}^n \le \mathbf{0} \qquad (10c)$$

where $\Theta_s(\mathbf{y}^n, \mathbf{z}^n) = \max\{\tilde{\mathbf{X}}^n(s) + \mathbf{y}^n\mathbf{e}_L^T + \mathbf{e}_K(\mathbf{z}^n)^T, \mathbf{0}\}$. The objective function in problem (10) is convex and continuously differentiable over $(\mathbf{y}^n, \mathbf{z}^n)$ and its gradient is

$$\left(\Theta_s(\mathbf{y}^n, \mathbf{z}^n)\mathbf{e}_L + \mathbf{e}_K, \Theta_s(\mathbf{y}^n, \mathbf{z}^n)^T\mathbf{e}_K + \mathbf{e}_L\right)^T.$$

Since the dual problem (10) involves only the nonpositive constraint and the projection of any point onto the nonpositive orthant is simple to compute, we can easily apply the gradient projection method to solve the dual problem (10). After obtaining the solution $(\mathbf{y}^n, \mathbf{z}^n)$ of the dual problem (10), we can recover the solution of the original problem (9) as

$$\mathbf{X}^n(s + 1) = \Theta_s(\mathbf{y}^n, \mathbf{z}^n). \qquad (11)$$

Simulation results show that the above algorithm for solving the projection problem (9) is much faster than directly using solver CVX [22]. The above algorithm is an extension of the DualBB algorithm proposed in [23, Sec. 4.3] from the square matrix case to the nonsquare matrix case and from the linear equality constraint case to the linear inequality constraint case.

2) Update $\mathbf{P}$ by Proximal Gradient: At the $(s + 1)$th iteration, the optimization of $\mathbf{P}$ with $\mathbf{X}$ fixed as $\mathbf{X}(s + 1)$ is nonconvex but is separable among different BSs. However, because of the nonsmooth ℓ_2-norm term in the objective (6) of problem (5), we need to use the proximal gradient method [20], [21]. First, compute the gradient update

$$\tilde{\mathbf{p}}_\ell(s + 1) = \mathbf{p}_\ell(s) + \beta_\ell(s + 1)\nabla_{\mathbf{p}_\ell} f(\mathbf{X}(s + 1), \mathbf{P}(s)) \quad (12)$$

where $\beta_\ell(s + 1)$ is some appropriately chosen step size, e.g., 0.5. The proximal gradient method [20], [21] updates $\mathbf{p}_\ell$ by

---

Algorithm 2: A Block Coordinate Gradient Ascent Algorithm for Solving Problem (5).

Step 1. Chose an initial point $(\mathbf{X}(0), \mathbf{P}(0))$. Set $s = 0$.
Step 2. Use the gradient projection method to solve problem (10) and use (11) to recover the solution $\mathbf{X}^n(s + 1)$ of problem (9) for all $n = 1, 2, \ldots, N$.
Step 3. Use (12), (15) and (16) to obtain $\mathbf{p}_\ell(s + 1)$ for all $\ell = 1, 2, \ldots, L$. Set $s = s + 1$ and go to **Step 2.**

---

solving the following problem for each BS $\ell = 1, 2, \ldots, L$:

$$\min_{\mathbf{p}_\ell} \quad \frac{1}{2}\|\mathbf{p}_\ell - \tilde{\mathbf{p}}_\ell(s + 1)\|_2^2 + t_\ell(s + 1)\|\mathbf{p}_\ell\|_2 \qquad (13)$$

where

$$t_\ell(s + 1) = \lambda\psi_\ell w_\ell\beta_\ell(s + 1). \qquad (14)$$

This problem has a closed-form solution [21, Sec. 6.5.1]:

$$\hat{\mathbf{p}}_\ell(s + 1) = \max\left\{1 - \frac{t_\ell(s + 1)}{\|\tilde{\mathbf{p}}_\ell(s + 1)\|_2}, 0\right\}\tilde{\mathbf{p}}_\ell(s + 1) \quad (15)$$

which is sometimes called block soft thresholding. Since $\mathbf{p}_\ell$ must satisfy the power budget constraint, we further project $\hat{\mathbf{p}}_\ell(s + 1)$ onto the feasible set and obtain the next iterate

$$\mathbf{p}_\ell(s + 1) = \max\{\min\{\hat{\mathbf{p}}_\ell(s + 1), \bar{\mathbf{p}}_\ell\}, \mathbf{0}\}. \qquad (16)$$

We summarize the proposed BCGA algorithm for solving problem (5) as Algorithm 2 above. The initial point can be chosen as the converged $(\mathbf{X}, \mathbf{P})$ with the previous update of w_ℓ. Global convergence of Algorithm 2 can be established, with appropriately chosen step sizes $\alpha^n(s)$ and $\beta_\ell(s + 1)$ (e.g., by backtracking line search), as in [17]–[19].

C. Greedy Algorithm for Solving Problem (4)

This section outlines a greedy algorithm for solving problem (4), which serves as a benchmark for comparing with Algorithm 1. The idea is to *sequentially* determine whether or not to turn off an ON-BS with all the other BSs' ON–OFF statuses being fixed. More specifically, at each iteration, the greedy algorithm picks an ON-BS ℓ' and solves problem

$$\max_{\mathbf{X}, \mathbf{P}} \quad f(\mathbf{X}, \mathbf{P}) \qquad (17a)$$

$$\text{s.t.} \quad (4b)\text{–}(4e) \qquad (17b)$$

$$\mathbf{p}_\ell = \mathbf{0}, \ \ell \in \mathcal{L}' \cup \{\ell'\} \qquad (17c)$$

where $\mathcal{L}'$ is the set of OFF-BSs at the current iteration. The algorithm turns off BS ℓ' if the objective value of problem (4) at the solution of problem (17) is larger than that before turning off of BS ℓ' and otherwise keeps it on. Note that problem (17) is a smooth version of (4), so $\mathbf{X}$ and $\mathbf{P}$ can both be optimized by gradient projection as in Algorithm 2. The greedy algorithm for solving problem (4) is summarized as Algorithm 3 on the next page.

The greedy Algorithm 3 is a simple heuristic, with some possible drawbacks. First, the greedy algorithm can only turn off (at most) one BS at each iteration, which makes it inefficient and also makes it easy to get stuck in a locally optimal solution. In contrast, the proposed iteratively reweighting Algorithm 1 is capable of turning off multiple BSs simultaneously at each iteration; the proposed algorithm often terminates within only a few iterations. Note that the performance of Algorithm 3 is
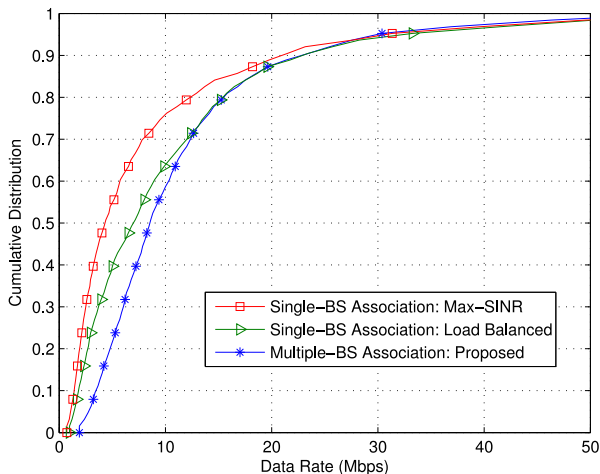
Fig. 1. Rate CDFs of single-BS versus multiple-BS association.

Fig. 2. Greedy versus proposed iteratively reweighting for BS turning-off.

Algorithm 3: A Greedy Algorithm for Solving Problem (4).

Step 1. Set $\mathcal{L}' = \emptyset$, i.e., all BSs are on.
Step 2. Apply Algorithm 2 to optimize $(\mathbf{X}, \mathbf{P})$ for all the ON-BSs in (17).
Step 3. Sequentially turn off each ON-BS ℓ' and set $\mathcal{L}' = \mathcal{L}' \cup \{\ell'\}$ if doing so increases the objective value of problem (4). Go to **Step 2.**

---

sensitive to the sequence of which BS to test in Step 2. In fact, the first few BSs in the testing sequence are more likely to be deactivated than the rest, particularly when the initial utility value is low. A random sequence might lead to quite suboptimal turning-off decisions.

IV. SIMULATION RESULTS

We simulate a seven-cell HetNet deployed on a regular hexagon wrapped around topology, with one macro-BS and three pico-BSs per cell. A total of 63 users are uniformly distributed throughout the network. The downlink distance-dependent pathloss is modeled as $128.1 + 37.6 \log_{10} d$, where d represents the BS-to-user distance in kilometers; the shadowing effect is modeled as a zero-mean Gaussian random variable with 8-dB standard deviation. The total bandwidth is 10 MHz, divided into 16 equal bands. The background noise PSD is -169 dBm/Hz; the max transmit PSD of macro-BS is -27 dBm/Hz; the max transmit PSD of pico-BS is -47 dBm/Hz. The ON-power of macro-BS is 1450 W; the ON-power of pico-BS is 21.32 W. Two single-BS association schemes are included as benchmarks: 1) max-SINR scheme, which assigns each user to the BS with the highest SINR, and then performs power control under fixed association; 2) load balanced scheme [7], which uses a proportional fairness objective to associate users with BSs followed by power control under fixed association. The allocation variable $\mathbf{X}$ in these two benchmark schemes is also iteratively optimized with $\mathbf{P}$ using Steps 2 and 3 of Algorithm 2.

Fig. 1 compares the cumulative distribution function (CDF) of user rates for multiple-BS and single-BS association schemes without power consumption penalty (i.e., when $\lambda = 0$). It is observed that the proposed multiple-band multiple-BS association approach significantly outperforms the single-BS association baselines, particularly in the low-rate regime, e.g., the proposed algorithm doubles the 10th percentile rate as compared to
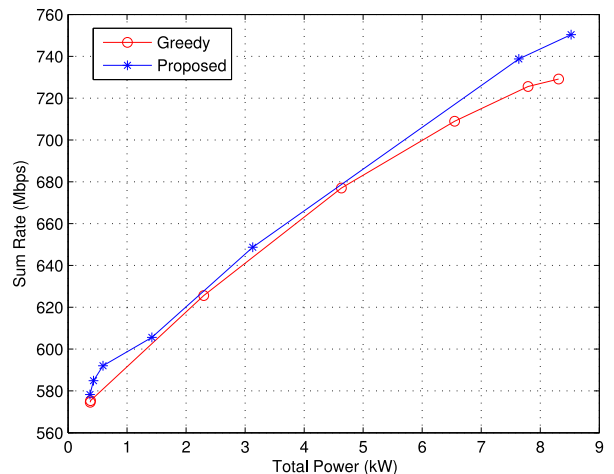
single-BS association. We remark that all the user-BS channels are assumed to be flat-fading across the 16 bands in the simulation, so the reported performance gain is entirely due to multiple-BS association but not frequency diversity.

We now evaluate the effectiveness of the proposed algorithms for power optimization and in turning off BSs for power saving. The different points in Fig. 2 are obtained by solving (4) with different λ values. It is observed that the proposed algorithm slightly outperforms the greedy method. It should be emphasized that the greedy algorithm is highly sensitive to the initial starting point and the testing sequence of BSs (as in Step 3 of Algorithm 3), which are carefully tuned in this simulation. In particular, we choose to sequentially pick the ON-BSs first across the macro-BSs, and then across the pico-BSs, and terminate the algorithm until all ON-BSs are picked twice. This testing sequence is believed to be near optimal in this particular setting because the network model has only two tiers of BSs and their ON-power levels differ significantly. In a more complex network, however, it would be more difficult to find a (near-)optimal BS testing sequence. Another drawback of the greedy heuristic is its high complexity, as it involves testing of candidate BSs one at a time, while our proposed algorithm is capable of turning off multiple BSs at the same time through coordinating the proximal updates across the ON-BSs. In fact, the proximal update in our simulation has a fast convergence, typically within 30 iterations. As the greedy algorithm requires a centralized evaluation of the total network utility after every BS-deactivation attempt, its computational complexity would be prohibitively high when a large number of BSs are deployed in the network.

V. CONCLUSION

This letter illustrates the benefit of providing the BSs with the flexibility of varying transmit PSDs and possibly being turned off, and providing the users with the flexibility of associating with multiple BSs across multiple frequency bands in a downlink HetNet. We show that the network utility maximization problem with a power consumption penalty can be solved efficiently using a combination of reweighted sparse optimization, block coordinate ascent, and proximal gradient method. By jointly optimizing the PSD levels across the bands, the multiple-BS association, and the user frequency allocation, the proposed algorithm enables effective balancing between network throughput and power consumption.

REFERENCES

[1] 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, Mar. 2015.

[2] L. P. Qian, Y. J. Zhang, Y. Wu, and J. Chen, "Joint downlink base station association and power control via benders' decomposition," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1651–1665, Mar. 2013.

[3] V. N. Ha and L. B. Le, "Distributed base station association and power control for heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 282–296, Jul. 2013.

[4] D. Lopez-Perez, X. Chu, and I. Guvenc, "On the expanded region of picocells in heterogeneous networks," *IEEE J. Sel. Top. Signal Process.*, vol. 6, no. 3, pp. 281–294, Jun. 2012.

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

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

[7] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[8] A. Abbasi and M. Ghaderi, "Energy cost reduction in cellular networks through dynamic base station activation," in *Proc. IEEE Ann. Int. Conf. Sens. Commun. Netw.*, Jun. 2014, pp. 363–371.

[9] S. Ping, A. Aijaz, O. Holland, and A. H. Aghvami, "Green cellular access network operation through dynamic spectrum and traffic load management," in *Proc. IEEE 24th Ann. Int. Symp. Pers. Indoor Mobile Radio Commun.*, Sep. 2013, pp. 2791–2796.

[10] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, Mar. 2016.

[11] 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.

[12] 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.

[13] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[14] 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.

[15] Y.-F. Liu, M. Hong, and E. Song, "Sample approximation-based deflation approaches for chance SINR constrained joint power and admission control," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4535–4547, Jul. 2016.

[16] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[17] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[18] Y. Xu and W. Yin, "A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[19] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Math. Program.*, vol. 117, no. 1–2, pp. 387–423, 2009.

[20] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. New York, NY, USA: Springer, 2011, pp. 185–212.

[21] N. Parikh and S. P. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[22] M. Grant, S. Boyd, and Y. Ye, CVX: Matlab software for disciplined convex programming, 2008. [Online]. Available: http://cvxr.com/cvx/

[23] B. Jiang, Y.-F. Liu, and Z. Wen, "LP-norm regularization algorithms for optimization over permutation matrices," *SIAM J. Optim.*, vol. 26, no. 4, pp. 2284–2313, 2016.