Managing Analog Beams in mmWave Networks

Yasaman Ghasempour*, Narayan Prasad†, Mohammad Khojastepour† and Sampath Rangarajan†
*Rice University †NEC Labs America; e-mail: Ghasempour@rice.edu, {prasad, amir, sampath}@nec-labs.com

***Abstract*—In this paper we consider multi-cell mmWave networks wherein each cell equipped with a large antenna array can employ an analog precoder (or a group of analog beams) to serve its associated users, while each such user can employ a single analog beam. A key problem over such a network is to determine the set of users that each cell should serve (a.k.a. user association), the group of beams it should employ, as well as their attributes such as how often and with how much power should each beam be used. This problem becomes harder since the choice of beam at any user is coupled to the cell it is assigned to and the latter's choice of beams. Moreover, practical considerations demand that each transmitting and receiving beam and their attributes be selected from finite codebooks. We develop novel solutions to this seemingly intractable problem. We adopt the generalized Quality-of-Service (QoS) Proportional Fairness (PF) network utility which can balance efficiency with fairness, and is particularly relevant for coverage constrained mmWave systems, since QoS constraints demand provisioning a minimum rate for each user. We prove that, remarkably, the user association problem under this QoS-PF utility can be formulated as a constrained submodular set function maximization problem. Consequently, it can be optimally solved (upto an additive constant) using distributed algorithms. We then propose a simple distributed algorithm to optimize the choice of beams and their attributes, and prove that it converges to a social equilibrium even in the presence of a non-ideal communication channel among cells.**

I. Introduction

The key components that are expected to provide bulk of the throughput improvements in 5G networks are *Massive MIMO* and *mmWave*. The former involves employing large antenna arrays, which with ideal channel state information (CSI), promises to dramatically improve multiplexing gains [1]. On the other hand, the latter seeks to harvest large chunks of hitherto unused spectrum bands. It is also recognized that exploiting mmWave for access necessitates highly directional radiation of signal energy to overcome the high propagation loss, which in turn is made possible by beamforming using large antenna arrays [2]. However, supporting a fully flexible (or digital) beamforming is considered prohibitive (in lieu of the very high bit rates that the ADCs must operate under) and as a result hybrid architectures have gained prominence. These architectures involve the use of a group of analog beams driven by a limited number of RF chains at each node and have received significant recent attention [3].

In this paper we focus on mmWave networks employing hybrid precoding in a multi-cell setting. An important aspect then is interference management among cells via coordination. However, owing to the very small channel coherence time at high carrier frequencies, achieving coordination among cells that seeks to exploit short term CSI imposes stringent constraints on latency in CSI gathering and dissemination, as well as constraints on reliability of inter-cell communication links. Such strict limits are quite unlikely to be met which makes the former coordination schemes ill suited and even detrimental. Fortunately, the large-scale fading parameters such as path loss and spatial correlations (which depend on angles of arrival and departure, path delays and array geometries) change at much coarser time scales, and can indeed be gainfully employed in coarse-level coordination schemes. This observation has been exploited in recent works, such as the *JSDM* scheme where angular spectra of users in a cell is used to partition them into non-overlapping groups [4], [5], as well as in schemes proposing cell-specific precoding and two time-scale resource management [7]. We add to this body of work by analyzing multi-cell mmWave systems with two important practical considerations, namely, restricting analog beams and attributes to finite codebooks and non-ideal communication links between cells. Our key contributions in this work are:

- We prove that the user association problem to optimize the generalized QoS-PF utility can be formulated as a submodular set function maximization problem. This result opens up a variety of distributed and centralized algorithms which can be used to determine an approximately optimal user association. We note here that research on user association or load balancing is well established [8] and of particular practical interest since it requires limited coordination among cells. Indeed, combinations of load balancing with several resource management schemes have also received wide attention [9]–[15]. However, no analytical result on load balancing to optimize the QoS-PF utility have so far been derived. In this context, we note that QoS-PF utility (entailing minimum rate constraints) is very relevant for mmWave since it ensures provisioning for coverage. Our proof methodology for showing submodularity is also novel and has wider applications. Indeed, as a by-product we obtain the result that the generalized water-filling problem (with arbitrary non-identical weights) is a submodular maximization problem. We note that [17] demonstrated the submodularity of waterfilling (when used to maximize Shannon rates over power allocations in a multi-channel setting) with identical weights. As noted by the authors of [17], extending their intricate proof to the generalized case with arbitrary weights was intractable. Our proof solves this open problem by devising an alternate novel approach.
- We propose a simple distributed algorithm to optimize the analog beam parameters. The proposed method provably converges to a social equilibrium even in the presence of backhaul erasures. In this context, we note that randomized algorithms have been used for discrete optimization in a multi-cell setting [18]. The novelty of our approach compared to [18], is that it is fully aligned to the 3GPP signalling framework for backhaul communications [16] and offers robustness against non-idealities in such communication.

In the following, we present the two key results along with proofs. Due to space constraints we defer further details on practical implementation as well as simulation results to [20].

II. Problem Formulation

We consider the downlink in a multi-cell network and let $\mathcal{U}$ denote the set of users with cardinality $|\mathcal{U}| = K$. Each user has multiple receive antennas, one receive RF chain and can choose any vector from a finite codebook for analog receive beamforming. Let $\mathcal{M}$ denote the set of TPs, where each TP has N_{TX} & S_{TX} transmit antennas and RF chains, respectively. Further, let $\mathcal{B}$ denote a codebook of transmit analog beams. For each TP $m \in \mathcal{M}$, let $\Gamma_{m,b} \in \mathcal{S}$ denote the duty cycle of the b^{th} beam for TP m, where $\mathcal{S}$ denotes a pre-defined finite set of such duty cycles. In particular, each $\Gamma_{m,b}$ lies in the unit interval $[0, 1]$ and denotes the fraction of the frame duration for which the b^{th} beam is activated in the m^{th} TP. Thus, $\Gamma_{m,b} = 0$ implies that the b^{th} beam is not activated at all by the m^{th} TP, whereas $\Gamma_{m,b} = 1$ implies that the b^{th} beam is activated for the entire frame duration by the m^{th} TP.[1]

[1]Note that we do not optimize the precise positions (or subframes) in a frame where a beam is activated by a TP, since doing so without the knowledge of instantaneous fading seen on that subframe will not be useful.

$\mathbf{\Gamma} = [\Gamma_{m,b}]_{m\in\mathcal{M},b\in\mathcal{B}}$ denotes the collection of all chosen duty cycles, whereas $\mathbf{\Gamma}_{(m)} = [\Gamma_{m,b}]_{b\in\mathcal{B}}$ denotes those pertaining to TP m for each $m \in \mathcal{M}$. Then, let P_m denote the power-level at which the signal is transmitted along any beam used by the m^{th} TP and let $\mathcal{P}$ denote the finite set of all possible power levels. Finally, let $\mathbf{P} = [P_m]_{m\in\mathcal{M}}$ be the vector of chosen power levels.

Let $R_{u,m,b}(\mathbf{\Gamma}, \mathbf{P})$ denote an estimated achievable peak rate for the u^{th} user when it is served by the m^{th} TP using the b^{th} beam, given all the chosen duty cycles in $\mathbf{\Gamma}$ and power levels in $\mathbf{P}$. We allow for any suitable peak rate estimation rule (cf. [6]) which uses: (i) estimates of the slow fading parameters (e.g. path loss, shadowing and spatial correlations) seen by user u on the frame of interest (we assume these estimates are available) but not the instantaneous fast-fading ones and (ii) the best analog receive beamforming vector at user u among those in the given finite receive beamforming codebook. We are now ready to pose the problem of interest in (1). In (1) $1\{.\}$ denotes an indicator function which is one if the input argument is true and is zero otherwise. Noting that the objective in (1) is the generalized PF utility in which $w_u > 0$ is the weight assigned to the u^{th} user, we proceed to the explain the constraints imposed:

- The first set of constraints enforce that the sum of the non-negative fractions $\{\gamma_{u,m,b}\}$ for each TP and beam combination, which are normalized by the duty cycle chosen for that combination and henceforth are referred to as allocation fractions, over all the users should not exceed unity. Further, the sum of duty cycles across all activated beams in any TP m should not exceed the number of transmit RF chains S_{TX}.
- The total number of activated beams at any TP should not exceed L. Note that $S_{\mathrm{TX}} \leq L \leq N_{\mathrm{TX}}$. Choosing a larger value of L can improve performance but at the cost of an increased channel state information (CSI) acquisition overhead.
- The chosen set of duty cycles and the power level at any TP m must be compatible, i.e., must satisfy $f_m(\mathbf{\Gamma}_{(m)}, P_m) \leq 1$. For instance, the function $f_m(.)$ can check the sum power budget at TP m.
- Considering each user, we impose minimum rate constraints. By incorporating such minimum rate constraints, we have addressed the *most general PF utility that allows for different user priorities and Quality-of-Service (QoS).*
- We also impose the constraint that each user is allowed to be served by any one TP using any one beam. This constraint is meaningful for coarse time scale optimization in mmWave systems since each user will receive bulk of its data along one beam, which is typically LoS. Note however that the beamforming vectors used by a TP in the fine time-scale (subframe granularity) to serve its associated users will be constructed based on the knowledge of instantaneous fading and can be any vectors in the span of its chosen beam group. A schematic is shown in Fig. 1 where it can be seen that the configuration on the right is better optimized.

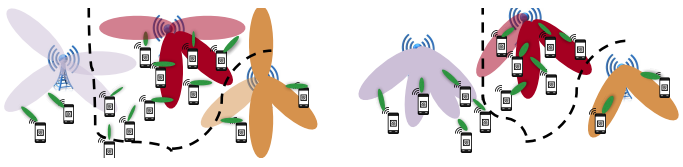

Fig. 1. Two system configurations each depicting user association (dashed line), transmit and receive analog beamforming, with the opacity of the beams being proportional to their power levels or duty cycles

III. An Alternating Optmization Framework

We adopt an alternating optimization framework to optimize (1). In particular, we optimize the user association and beam parameters (duty cycles and power levels) in an alternating manner. In each case we provide novel algortihms with certain optimality guarantees.

A. Optimizing user association and allocation fractions

In this section we jointly optimize the user association and allocation fractions for any given set of duty cycles and power levels $(\hat{\mathbf{\Gamma}}, \hat{\mathbf{P}})$. The problem of interest can be written as in (2), where we let $x_{u,m,b}$ denote an indicator variable that is one if user u can be served using beam b and TP m and zero otherwise. Notice that upon fixing any choice of $\{x_{u,m,b}\}$ the optimizination of allocation fractions in (2) decouples into $|\mathcal{M}||\mathcal{B}|$ optimization problems, one for each beam,TP combination. Define a ground set $\underline{\Omega} = \{(u,m,b), u \in \mathcal{U}, b \in \mathcal{B}, m \in \mathcal{M}\}$ where (u,m,b) conveys the association of user u with TP m and beam b (i.e., is equivalent to setting $x_{u,m,b} = 1$). Without loss of generality we suppose that only a tuple (u,m,b) for any $u \in \mathcal{U}$ & $b \in \mathcal{B}, m \in \mathcal{M}$ for which $R_{u,m,b}(\hat{\mathbf{\Gamma}}, \hat{\mathbf{P}}) \geq R_u^{\min}$ is included in $\underline{\Omega}$. This is because any tuple not satisfying this assumption will never be selected as its minimum rate cannot be met even when the assigned TP along the chosen beam fully allocates its resource to that user. Let $\underline{\Omega}^{(m',b')} = \{(u,m,b) \in \underline{\Omega} : b = b', m = m'\}$ denote all possible associations of users to the TP and beam combination m', b', where $b' \in \mathcal{B}$ & $m' \in \mathcal{M}$, and let $\underline{\Omega}_{(u')} = \{(u,m,b) \in \underline{\Omega} : u = u'\}$ denote all possible associations of a user $u' \in \mathcal{U}$. Define a family of sets $\underline{\mathcal{J}}$ as as the one which includes each subset of $\underline{\Omega}$ such that the tuples in that subset have mutually distinct users and the minimum rates of those users are feasible. Notice that any $\underline{\mathcal{G}} \in \underline{\mathcal{J}}$ specifies a particular choice of $\{x_{u,m,b}\}$ for (2) satisfying $\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}} x_{u,m,b} \leq 1, \ \forall \ u \in \mathcal{U}$ and if $\underline{\mathcal{G}}$ is maximal (i.e., $|\underline{\mathcal{G}}| = K$) then we have a valid user association satisfying $\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}} x_{u,m,b} = 1, \ \forall \ u \in \mathcal{U}$. Further, $\underline{\mathcal{J}}$ is a downward closed family, i.e., if $\underline{\mathcal{G}} \in \underline{\mathcal{J}}$ then each subset of $\underline{\mathcal{G}}$ is also a member of $\underline{\mathcal{J}}$. Next, we define a real-valued set function on $\underline{\mathcal{J}}$, $f_2 : \underline{\mathcal{J}} \rightarrow \mathbb{R}$ such that it is normalized $f(\phi) = 0$, where ϕ is the empty set, and for any non-empty set $\underline{\mathcal{G}} \in \underline{\mathcal{J}}$, we have

$$f(\underline{\mathcal{G}}) = \sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}} f_{m,b}(\underline{\mathcal{G}} \cap \underline{\Omega}^{(m,b)}). \tag{3}$$

Each $f_{m,b} : \underline{\mathcal{J}}^{(m,b)} \rightarrow \mathbb{R}$ in (3) is a normalized set function that is defined on the family $\underline{\mathcal{J}}^{(m,b)}$ comprising of each member of $\underline{\mathcal{J}}$ that is contained in $\underline{\Omega}^{(m,b)}$, as follows. For any set $\underline{\mathcal{A}} \in \underline{\mathcal{J}}^{(m,b)}$, we define $f_{m,b}(\underline{\mathcal{A}})$ to be the optimal objective value obtained by optimizing the allocation fractions of all users with $x_{u,m,b} = 1$ for the combination m, b under consideration. The optimization of the allocation fractions for any (m,b) and the associated users is detailed in Proposition 2 along with a a simple necessary and sufficient condition to determine feasibility of the minimum rates for the given choice of association. With these definitions in hand, can re-formulate the problem in (2) as the following constrained set function maximization problem.

$$\max_{\underline{\mathcal{G}}\in\underline{\mathcal{J}}:|\underline{\mathcal{G}}|=K} \{f(\underline{\mathcal{G}})\} \tag{4}$$

We offer our first main result that characterizes $f(.)$.

Theorem 1. *The set function $f(.)$ is a normalized submodular set function.*

Proof. The set function $f(.)$ in (3) defined on the family $\underline{\mathcal{J}}$ is normalized by construction. Then, to establish submodularity

$$\max_{\substack{\Gamma_{m,b}\in\mathcal{S},P_m\in\mathcal{P},\gamma_{u,m,b}\in[0,1]\\ \forall\, u\in\mathcal{U},b\in\mathcal{B},m\in\mathcal{M}}} \left\{\sum_{u\in\mathcal{U}} w_u \log\left(\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}} R_{u,m,b}(\mathbf{\Gamma},\mathbf{P})\gamma_{u,m,b}\right)\right\}$$
$$\text{s.t.}\sum_{b\in\mathcal{B}}\Gamma_{m,b}\le S_{\text{TX}}\ \&\ \sum_{u\in\mathcal{U}}\gamma_{u,m,b}\le 1,\ \forall\, b\in\mathcal{B}, m\in\mathcal{M};\ \sum_{b\in\mathcal{B}}\mathbf{1}\left\{\Gamma_{m,b}>0\right\}\le L\ \&\ f_m(\mathbf{\Gamma}_{(m)},P_m)\le 1,\ \forall\, m\in\mathcal{M};\quad (1)$$
$$\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}}\gamma_{u,m,b}R_{u,m,b}(\mathbf{\Gamma},\mathbf{P})\ge R_u^{\min},\ \sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}}\mathbf{1}\left\{\gamma_{u,m,b}>0\right\}=1,\ \forall\, u\in\mathcal{U};$$

$$\max_{\substack{x_{u,m,b}\in\{0,1\},\gamma_{u,m,b}\in[0,1]\\ \forall\, u\in\mathcal{U},b\in\mathcal{B},m\in\mathcal{M}}} \left\{\sum_{u\in\mathcal{U}} w_u \log\left(\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}} R_{u,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\gamma_{u,m,b}x_{u,m,b}\right)\right\}$$
$$\text{s.t.}\sum_{u\in\mathcal{U}}\gamma_{u,m,b}x_{u,m,b}\le 1,\ \forall\, b\in\mathcal{B}, m\in\mathcal{M};\ \sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}}\gamma_{u,m,b}x_{u,m,b}R_{u,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\ge R_u^{\min},\ \forall\, u\in\mathcal{U};\quad (2)$$
$$\sum_{m\in\mathcal{M}}\sum_{b\in\mathcal{B}}x_{u,m,b}=1,\ \forall\, u\in\mathcal{U}.$$

of $f(.)$ on the family $\underline{\boldsymbol{\mathcal{J}}}$, it suffices to show that each $f_{m,b}(.)$ is submodular on the family $\underline{\boldsymbol{\mathcal{J}}}^{(m,b)}$. Without loss of generality, we consider any TP m with beam b and will prove that

$$f_{m,b}(\underline{\mathcal{E}}\cup(u_1,m,b))-f_{m,b}(\underline{\mathcal{E}})\ge f_{m,b}(\underline{\mathcal{F}}\cup(u_1,m,b))-f_{m,b}(\underline{\mathcal{F}}),\quad (5)$$

for all $\underline{\mathcal{E}}\subseteq\underline{\mathcal{F}}\in\underline{\boldsymbol{\mathcal{J}}}^{(m,b)}:|\underline{\mathcal{F}}|=|\underline{\mathcal{E}}|+1$ and any $(u_1,m,b)\in\underline{\Omega}\setminus\underline{\mathcal{F}}:\underline{\mathcal{F}}\cup(u_1,m,b)\in\underline{\boldsymbol{\mathcal{J}}}^{(m,b)}$. Towards this end, we expand $\underline{\mathcal{F}}=\underline{\mathcal{E}}\cup(u_2,m,b)$ where we must have $(u_2,m,b)\in\underline{\boldsymbol{\mathcal{J}}}^{(m,b)}$ with $u_2\neq u_1$. Then, we evaluate $f_{m,b}(\underline{\mathcal{F}}\cup(u_1,m,b))$ as described in Proposition 2 (using unit budget) and in the obtained optimal allocation fractions let the share of resource (allocation fraction) assigned to user u_1 in tuple (u_1,m,b) be δ_1 . Similarly, let the share of resource assigned to user u_2 in tuple (u_2,m,b) be δ_2. Define $\hat{\zeta}=1-\delta_1-\delta_2$. Thus, we have that

$$f_{m,b}(\underline{\mathcal{F}}\cup(u_1,m,b))=\hat{O}(\hat{\zeta})+w_{u_1}\log(R_{u_1,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\delta_1)+w_{u_2}\log(R_{u_2,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\delta_2),\quad (6)$$

where $\hat{O}(\hat{\zeta})$ is the objective value evaluated for the tuples in $\underline{\mathcal{E}}$ under the budget $\hat{\zeta}$, using Proposition 2. Further, it can be readily verified that

$$\begin{aligned} f_{m,b}(\underline{\mathcal{F}})&\ge \hat{O}(\hat{\zeta}+\delta_1)+w_{u_2}\log(R_{u_2,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\delta_2),\\ f_{m,b}(\underline{\mathcal{E}})&=\hat{O}(\hat{\zeta}+\delta_1+\delta_2)\\ f_{m,b}(\underline{\mathcal{E}}\cup(u_1,m,b))&\ge \hat{O}(\hat{\zeta}+\delta_2)+w_{u_1}\log(R_{u_1,m,b}(\hat{\mathbf{\Gamma}},\hat{\mathbf{P}})\delta_1).\end{aligned}\quad (7)$$

Using (6) and (7) in (5), it is now seen that a sufficient condition for (5) to hold is for (13) to be true. The latter is assured by Proposition 2 which yields our desired result. □

The significance of Theorem 1 is that (4) is a constrained submodular set function maximization problem which can be approximately maximized by leveraging existing distributed or centralized algorithms [19], which can guarantee optimality upto an additive constant.

B. Optimizing beam parameters and allocation fractions

We suppose that a user association, $\{x_{u,m,b}=\hat{x}_{u,m,b}\}$, has been given. We proceed to optimize the set of duty cycles and power levels based on the given association. In particular, for each TP $m\in\mathcal{M}$ we first determine $\mathcal{U}_m$ to be the set of users associated with TP m under the given association, i.e., $\mathcal{U}_m=\{u\in\mathcal{U}:\max_{b\in\mathcal{B}}\{\hat{x}_{u,m,b}\}=1\}$. Then, for each TP, we optimize the beam parameters as well as the intra-TP user association, i.e., we allow for each user in $\mathcal{U}_m$ to be served by TP m using any one beam in $\mathcal{B}$. This is important since otherwise we will be restricted to using only the beams that have at-least one associated user as per the given association and no beam alteration at any TP would be possible. Next, we define a state of any TP $m\in\mathcal{M}$ by its choice of duty cycles and power levels as $\psi_m=(\mathbf{\Gamma}_{(m)},P_m)$. The set of all feasible states that any TP can be is finite and is denoted by Ψ. Here by a feasible state for any TP m we mean the state that satisfies the constraints imposed on the sum of the chosen duty cycles and number of activated beams, as well as the compatibility between the chosen duty cycles and power level. Notice that $\Psi\subseteq(\otimes_{b\in\mathcal{B}}\mathcal{S})\otimes\mathcal{P}$. The system state is defined as the collection of states of all TPs, as $\boldsymbol{\psi}=\{\psi_m\}_{m\in\mathcal{M}}$. Finally, the set of all system states is denoted by $\mathbf{\Psi}=\otimes_{m\in\mathcal{M}}\Psi$. Notice that $\mathbf{\Psi}$ is also finite with cardinality at-most $|\mathcal{S}|^{|\mathcal{M}||\mathcal{B}|}|\mathcal{P}|^{|\mathcal{M}|}$. We write $R_{u,m,b}(\mathbf{\Gamma},\mathbf{P})$ as $R_{u,m,b}(\boldsymbol{\psi}),\ \forall\, b\in\mathcal{B},u\in\mathcal{U}_m$ and define the region of all feasible allocation fractions at any TP $m\in\mathcal{M}$ as,

$$\begin{aligned}\mathcal{F}_m(\boldsymbol{\psi})=\big\{&\gamma_{u,m,b}\in[0,1]\ \forall u\in\mathcal{U}_m, b\in\mathcal{B}:\\ &\sum_{u\in\mathcal{U}_m}\gamma_{u,m,b}\le 1,\ \forall\, b\in\mathcal{B};\\ &\sum_{b\in\mathcal{B}}\gamma_{u,m,b}R_{u,m,b}(\boldsymbol{\psi})\ge R_u^{\min},\ \forall\, u\in\mathcal{U}_m;\\ &\sum_{b\in\mathcal{B}}1\{\gamma_{u,m,b}>0\}\le 1,\ \forall\, u\in\mathcal{U}_m\big\}.\end{aligned}$$

We can now formulate the problem of interest in (8). Note here that for any (tentative) choice of system state $\boldsymbol{\psi}$, the inner maximization problem at each TP m in (8) is just the intra-TP user association and allocation fraction optimization problem, which we can express as a constrained submodular maximization problem (cf. Theorem 1). We adopt the natural greedy algorithm to sub-optimally solve this problem and let $h_m(\boldsymbol{\psi})$ denote the value obtained by TP m. We remind that each TP can adopt any centralized algorithm to solve its sub-problem and we set $h_m(\boldsymbol{\psi})=-\infty$ whenever the minimum rates of all users in $\mathcal{U}_m$ cannot be met using the chosen method. Let $h(\boldsymbol{\psi})$ denote the objective value in (8) obtained as $h(\boldsymbol{\psi})=\sum_{m\in\mathcal{M}}h_m(\boldsymbol{\psi})$.

Next, for each TP $m\in\mathcal{M}$, given its current state $\psi_m\in\Psi$, we define the set of actions as $\Theta(\psi_m,m)$. Here, $\Theta(\psi_m,m)\subset\Psi$ and denotes the set of other states TP m can transition

$$\max_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\left\{\sum_{m\in\mathcal{M}}\max_{\{\gamma_{u,m,b}\}\ \substack{u\in\mathcal{U}_m\\ b\in\mathcal{B}}\ \in\mathcal{F}_m(\boldsymbol{\psi})}\left\{\sum_{u\in\mathcal{U}_m} w_u\log\left(\sum_{b\in\mathcal{B}} R_{u,m,b}(\boldsymbol{\psi})\gamma_{u,m,b}\right)\right\}\right\} \tag{8}$$

to from its current one. As will be revealed by our analysis, limiting the size of $\Theta(\psi_m, m)$ reduces the complexity but at the cost of performance guarantee. To quantify the impact of any action $\vartheta \in \Theta(\psi_{m1}, m1)$ by TP $m1$ on itself as well as any other TP in $\mathcal{M}$, we let $g(\vartheta, m1, m2, \boldsymbol{\psi})$ denote the change in utility over users associated to TP $m2$ when TP $m1$ changes its state to ϑ and all other TPs retain their respective current states. Notice that each TP $m2$ can compute

$$g(\vartheta, m1, m2, \boldsymbol{\psi}) = h_{m2}(\boldsymbol{\psi}') - h_{m2}(\boldsymbol{\psi}), \tag{9}$$

where $\boldsymbol{\psi}'$ includes the changed state of TP $m1$, ϑ, while $\boldsymbol{\psi}$ includes the original state of TP $m1$. Thus, the impact (or change in system utility) of TP $m1$ taking action ϑ can be quantified as $\sum_{m2\in\mathcal{M}} g(\vartheta, m1, m2, \boldsymbol{\psi})$.

With these definitions in hand, we offer Algorithm I and proceed to explain it. The time axis is divided into intervals of identical size, referred to as update intervals. In each interval, each TP $m1$ requests the impact of each of its allowed actions on the utility contribution of all other TPs. We allow for a *realistic scenario in which the backhaul connection between TP $m1$ and TP $m2 \in \mathcal{M} : m2 \neq m1$ is vulnerable to errors or jitters*. We model each backhaul channel as an erasure channel, where the erasures are independent across update intervals and backhaul channels. Accordingly, we suppose that the impact of all of its actions can be obtained by TP $m1$ with probability $q(\psi_{m1}, m1) \triangleq \prod_{m2\in\mathcal{M}:m2\neq m1} q(\psi_{m1}, m1, m2)$ where this probability is strictly positive and is allowed to depend on the TP and its state. Then, if no erasures occur, TP $m1$ determines its best action with respect to improvement in system utility (assuming all other TPs retain their respective states). Further, if the best such improvement is greater than a specified threshold ϵ, the corresponding action is accepted with a probability p_{m1}. The process continues till no TP can determine a state (in the absence of erasures) that can offer an improvement greater than ϵ. Notice that the sequence of system states seen across update intervals need not be monotonic (with respect to the system utility). This is because the algorithm is *distributed and multiple TPs can update their states in an interval*. Despite this we can guarantee convergence to a social equilibrium. Let us define an absorbing state (or social equilibrium) as one in which no single TP can improve the system utility more than ϵ, by changing its state to any one in the allowed set of actions. Specifically $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ is an absorbing state if for any other state $\boldsymbol{\psi}' \in \Psi$ such that $\boldsymbol{\psi}'$ differs from $\boldsymbol{\psi}$ only in the state of any one TP m and that differing state satisfies $\psi'_m \in \Theta(\psi_m, m)$, we have that $h(\boldsymbol{\psi}) \geq h(\boldsymbol{\psi}') - \epsilon$.

Proposition 1. *Algorithm I provably converges to an absorbing system state.*

Proof. To prove this claim, we note that an optimal system state (which yields the globally optimal system utility in (8) for the given association among all feasible states) exists and is also an absorbing state. Thus, the set of absorbing states is finite and non-empty. Further, given any non-absorbing system state it can be verified that we can construct a finite sequence of system states that begins at the given state and ends at an absorbing one, such that each transition from any state to the next one in that sequence involves an update by exactly one TP and yields a gain (in the system utility) better than ϵ. Moreover, given any two system states $\boldsymbol{\psi}, \boldsymbol{\psi}'$ we can deduce that a transition from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ is only possible if for each TP m either $\psi'_m = \psi_m$ or $\psi'_m \in \Theta(\psi_m, m)$. For each such possible transition we can explicitly determine the transition probability $p(\boldsymbol{\psi} \to \boldsymbol{\psi}')$ as in (11). Notice that this transition probability depends only on the system states $\boldsymbol{\psi}, \boldsymbol{\psi}'$. Consequently, the sequence of system states seen across the update intervals forms an *absorbing, time homogeneous Markov Chain*. Hence, convergence to an absorbing state is guaranteed. □

TABLE I
DISTRIBUTED BEAM GROUP MANAGEMENT

Initialize with a user association, a feasible system state $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, probabilities $\{p_m\}_{m\in\mathcal{M}}$ and a threshold $\epsilon \geq 0$.
Repeat
At each TP $m1 \in \mathcal{M}$:
For each TP $m2 \neq m1$ do
Request and obtain $\hat{g}(\vartheta, m1, m2, \boldsymbol{\psi}), \forall\ \vartheta \in \Theta(\psi_{m1}, m1)$
End For
If no erasure in feedback from any TP
Compute

$$\Delta(\vartheta, m1) = \sum_{m2\in\mathcal{M}} g(\vartheta, m1, m2, \boldsymbol{\psi}), \forall\ \vartheta \in \Theta(\psi_{m1}, m1) \tag{10}$$

Determine $\hat{\vartheta}_{m1} = \arg\max_{\vartheta\in\Theta(\psi_{m1},m1)}\{\Delta(\vartheta, m1)\}$
Else
Set $\hat{\vartheta}_{m1} = \psi_{m1}$ and $\Delta(\hat{\vartheta}_{m1}, m1) = 0$.
End If
Set decision to update to be false
If $\Delta(\hat{\vartheta}_{m1}, m1) > \epsilon$ then
Flip decision to update to true with probability p_{m1}
EndIf
For each received request from any TP $m2 \neq m1$ for its action ϑ', compute
$g(\vartheta', m2, m1, \boldsymbol{\psi})$ and report to TP $m2$.
If decision to update is true then
Update ψ_{m1} to $\hat{\vartheta}_{m1}$ and convey updated state to all other TPs.
End If
Collect all updated states from all other TPs.
Until Termination conditions are satisfied.

Remark 1. *Note that if for any TP $m1 \in \mathcal{M}$, the probabilities $q(\psi_{m1}, m1) \forall\ \psi_{m1} \in \boldsymbol{\Psi}$ all lie in the open interval $(0, 1)$, then Algorithm I will almost surely converge to an absorbing state even without any randomized rule (i.e., with $p_{m1} = 1$) at TP $m1$. In other words, since the backhaul injects randomization, Algorithm I will almost surely converge even if TP $m1$ always (deterministically) chooses its best action whenever it can determine that action's gain to be greater than ϵ.*

IV. CONCLUSIONS

We presented novel analytical results and algorithms for user association and analog beam parameter optimization over multi-cell mmWave networks.

APPENDIX

Proposition 2. *Consider any budget $\zeta > 0$, any set of users $\mathcal{U}'$ along with their minimum rates $\{R_u^{\min}\}_{u\in\mathcal{U}'}$ and peak rates $\{R_u\}_{u\in\mathcal{U}'}$ such that $\sum_{u\in\mathcal{U}'} \frac{R_u^{\min}}{R_u} \leq \zeta$. Let $\hat{O}(\zeta)$ denote the optimal objective value of the following optimization problem.*

$$\max_{\gamma_u\in[0,1]\forall\ u\in\mathcal{U}'}\left\{\sum_{u\in\mathcal{U}'} w_u\log\left(R_u\gamma_u\right)\right\}$$
$$\text{s.t.}\ \sum_{u\in\mathcal{U}'}\gamma_u \leq \zeta;\ \ \gamma_u R_u \geq R_u^{\min},\ \forall\ u\in\mathcal{U}'; \tag{12}$$

$$\left(\prod_{\substack{m\in\mathcal{M}\\ \psi_m\neq\psi'_m}} q(\psi_m,m)1\{\psi'_m=\hat{\vartheta}_m \;\&\; \Delta(\hat{\vartheta}_m,m)>\epsilon\}p_m\right)\left(\prod_{\substack{m\in\mathcal{M}\\ \psi_m=\psi'_m}} (1\{\Delta(\hat{\vartheta}_m,m)\leq\epsilon\}+1\{\Delta(\hat{\vartheta}_m,m)>\epsilon\}(1-q(\psi_m,m)+q(\psi_m,m)(1-p_m)))\right). \quad (11)$$

Then, for any non-negative scalars $\delta,\tilde{\delta}$, we have that

$$\hat{O}(\zeta)-\hat{O}(\zeta+\delta)\leq\hat{O}(\zeta+\tilde{\delta})-\hat{O}(\zeta+\delta+\tilde{\delta}). \quad (13)$$

Proof. We suppose, without loss of generality, that the user set $\mathcal{U}'$ can be parsed as $\mathcal{U}'=\mathcal{U}_1\cup\mathcal{U}_2$ where $\mathcal{U}_1=\{1,\cdots,k\}$, such that $R_u^{\min}>0,\forall\ u\in\mathcal{U}_1$ and $R_u^{\min}=0,\forall\ u\in\mathcal{U}_2$. Further, suppose that $w_1a_1<w_2a_2<\cdots<w_ka_k$, where $a_u=\frac{R_u}{R_u^{\min}}, u\in\mathcal{U}_1$.[2] Then, letting $\tilde{\zeta}=\zeta-\sum_{u\in\mathcal{U}_1}\frac{R_u^{\min}}{R_u}$ denote the slack budget, i.e., the budget left after meeting the minimum rate requirements, we can re-write (12) as,

$$\eta+\max_{\substack{\tilde{\gamma}_u\in[0,1]\\ \forall\ u\in\mathcal{U}'}}\left\{\sum_{u\in\mathcal{U}_1}w_u\log\left(1+a_u\tilde{\gamma}_u\right)+\sum_{u\in\mathcal{U}_2}w_u\log\left(R_u\tilde{\gamma}_u\right)\right\}\quad\text{s.t.}\quad\sum_{u\in\mathcal{U}'}\tilde{\gamma}_u\leq\tilde{\zeta}; \quad (14)$$

where $\eta=\sum_{u\in\mathcal{U}_1}w_u\log\left(R_u^{\min}\right)$. Next, since (14) is a convex optimization problem which is feasible for the given budget, we can employ the necessary and sufficient K.K.T. conditions to deduce the following. For any given slack budget $\tilde{\zeta}$, we will have a partition of $\mathcal{U}_1$ as $\mathcal{U}_1=\{1,\cdots,m-1\}\cup\tilde{\mathcal{U}}_1(\tilde{\zeta})$, where $\tilde{\mathcal{U}}_1(\tilde{\zeta})=\{m,\cdots,k\}$ depends on $\tilde{\zeta}$, such that users 1 to $m-1$ will be assigned exactly their minimum rates. Further, the assigned allocation fractions must satisfy $\gamma_u=\frac{w_u}{\lambda},\forall\ u\in\mathcal{U}_2$, where $\lambda>0$ is the Lagrangian variable, and all users in $\tilde{\mathcal{U}}_1(\tilde{\zeta})$ are allocated resources in excess of their minimum rate requirements as $\tilde{\gamma}_u=\frac{w_u}{\lambda}-\frac{1}{a_u}, u\in\tilde{\mathcal{U}}_1(\tilde{\zeta})$. Next, since the available slack budget $\tilde{\zeta}$ must be fully used, we get that $\lambda=\frac{\sum_{u\in\mathcal{U}_2}w_u+\sum_{\tilde{\mathcal{U}}_1(\tilde{\zeta})}w_u}{\tilde{\zeta}+\sum_{u\in\tilde{\mathcal{U}}_1(\tilde{\zeta})}1/a_u}$. We define $a_0=w_0=0$ & $a_{k+1}=w_{k+1}=\infty$ and partition $\mathbb{R}_+$ as $\mathcal{I}_{k+1}\cup\mathcal{I}_k\cup\cdots\cup\mathcal{I}_1$, where $\mathcal{I}_t=\left[\frac{A_t}{w_ta_t}-B_t,\frac{A_t}{w_{t-1}a_{t-1}}-B_t\right)$, $A_t=\sum_{u\in\mathcal{U}_2}w_u+\sum_{j=t}^{k}w_j$ and $B_t=\sum_{j=t}^{k}1/a_j$ for $t=1,\cdots,k+1$. Depending on the interval in which the slack budget lies in the above partition, we can detail the optimal objective value. In particular, suppose that $\tilde{\zeta}\in\mathcal{I}_t$. Then, $\tilde{\mathcal{U}}_1(\tilde{\zeta})=\{t,\cdots,k\}$ and

$$\hat{O}(\tilde{\zeta})=\sum_{u\in\mathcal{U}_1\setminus\tilde{\mathcal{U}}_1(\tilde{\zeta})}w_u\log(R_u^{\min})+\sum_{u\in\mathcal{U}_2\cup\tilde{\mathcal{U}}_1(\tilde{\zeta})}w_u\log(w_uR_u)+\sum_{u\in\mathcal{U}_2\cup\tilde{\mathcal{U}}_1(\tilde{\zeta})}w_u\log\left(\frac{\tilde{\zeta}+\sum_{u\in\tilde{\mathcal{U}}_1(\tilde{\zeta})}1/a_u}{\sum_{u\in\mathcal{U}_2}w_u+\sum_{u\in\tilde{\mathcal{U}}_1(\tilde{\zeta})}w_u}\right). \quad (15)$$

We can now verify that whenever the slack budget lies in the open interval $\tilde{\zeta}\in\left(\frac{B_t}{w_ta_t}-A_t,\frac{A_t}{w_{t-1}a_{t-1}}-B_t\right)$ for some $t=1,\cdots,k+1$ we can compute the derivative of $\hat{O}(\tilde{\zeta})$ with respect to $\tilde{\zeta}$ by treating $\tilde{\mathcal{U}}_1(\tilde{\zeta})$ as constant, as $\frac{\partial\hat{O}(\tilde{\zeta})}{\partial\tilde{\zeta}}=A_t/(\tilde{\zeta}+B_t)$. Notice now that within each open interval $\frac{\partial\hat{O}(\tilde{\zeta})}{\partial\tilde{\zeta}}$ decreases as $\tilde{\zeta}$ increases. Next, we can define and obtain the right derivative at the left boundary of the t^{th} interval $\tilde{\zeta}=\frac{A_t}{w_ta_t}-B_t$, as

$$\lim_{\delta\to0^+}\frac{\hat{O}(\tilde{\zeta}+\delta)-\hat{O}(\tilde{\zeta})}{\delta}=\frac{A_t}{\tilde{\zeta}+B_t}=w_ta_t, \quad (16)$$

for $t=1,\cdots,k$. Similarly, we can and obtain the left derivative at the right boundary of the $(t+1)^{th}$ interval, $\tilde{\zeta}=\frac{A_{t+1}}{w_ta_t}-B_{t+1}=\frac{A_t}{w_ta_t}-B_t$, as

$$\lim_{\delta\to0^-}\frac{\hat{O}(\tilde{\zeta}+\delta)-\hat{O}(\tilde{\zeta})}{\delta}=\frac{A_{t+1}}{\tilde{\zeta}+B_{t+1}}=w_ta_t. \quad (17)$$

It is seen that the left and right derivatives match. Thus we have shown that the derivative of $\hat{O}(\tilde{\zeta})$ exists and is decreasing in $\tilde{\zeta}$ for all $\tilde{\zeta}>0$ (or equivalently for all $\zeta>\sum_{u\in\mathcal{U}_1}\frac{R_u^{\min}}{R_u}$). This implies that $\hat{O}(\tilde{\zeta})$ is concave in $\tilde{\zeta}$ for all $\tilde{\zeta}>0$, which suffices to deduce that the result in (13) is indeed true. Finally, we note that (14) can be specialized to the waterfilling problem considered in [17] by setting $\mathcal{U}_1=\mathcal{U}', w_u=w,\ \forall\ u\in\mathcal{U}'$. □

[2]Note that for any given set of weights and minimum rates the scalars $\{w_ka_k\}$ will be distinct almost surely whenever the slow fading coefficients (on which the peak rates depend) are drawn from a continuous distribution.

REFERENCES

[1] Y. Lim, et.al., "Performance analysis of massive mimo for cell-boundary users," *IEEE Trans. Wireless. Comm.*, Dec. 2015.

[2] T. Rappaport, et.al., "Millimeter wave mobile communications for 5G cellular: It will work," *IEEE Access*, May 2013.

[3] A. Alkhateeb, et.al., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun. (revised)*, Mar. 2015.

[4] A. Adhikary, et. al., "Joint spatial division and multiplexing for mm-wave channels," *IEEE JSAC.*, Jun. 2014.

[5] A. Adhikary and G. Caire, "JSDM and multi-cell networks: Handling inter-cell interference through long-term antenna statistics," *IEEE Asilomar*, Nov. 2014.

[6] Z. Li, et.al., "Directional Training and Fast Sector-based Processing Schemes for mmWave Channels," *arXiv*, Nov. 2016.

[7] N. Prasad, et.al., "A two time scale approach for coordinated multi-point transmission and reception over practical backhaul," in *IEEE Comsnets (invited)*, jan 2014.

[8] J. Andrews, et.al., "An overview of load balancing in HetNets: Old myths and open problems," in *IEEE commun. mag.(submitted)*, July 2013

[9] G. Athanasiou, et.al., "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks," *IEEE Trans. Netw.*, vol. 23(3), 2015.

[10] H. Shokri-Ghadikolaei, et.al., "Millimeter wave cellular networks: A mac layer perspective," *IEEE Trans. Commun.*, Oct. 2015.

[11] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," in *IEEE Journal Sel. Areas. Commun.*, Jun. 2014.

[12] Q. Ye, et.al., "User association for load balancing in heterogeneous cellular networks," in *IEEE Trans. on Wireless Comm.*, June 2013.

[13] N. Prasad, et.al., "Exploiting cell dormancy and load balancing in LTE hetnets: Optimizing the proportional fairness utility," in *IEEE Trans. on Commun.*, Oct. 2014.

[14] D. Bethanabhotla, et.al., "Optimal user-cell association for massive mimo wireless networks," *v2, arXiv*, feb 2015.

[15] E. Altman, et.al., "Spatial sinr games of base station placement and mobile association," *IEEE Infocom.*, 2009.

[16] "3GPP Technical Report 36.842. Small cell enhancements for E-UTRA and E-UTRANHigher Layer Aspects", V1.0.0 (2013-11) www.3gpp.org

[17] K. Thekumparampil, et.al., "Combinatorial Resource Allocation Using Submodularity of Waterfilling," in *IEEE Trans. Wireless Comm.*, Jan. 2016.

[18] H. Zhang, et.al., "Weighted Sum-Rate Maximization in Multi-Cell Networks via Coordinated Scheduling and Discrete Power Control", in *IEEE JSAC*, Jun. 2011.

[19] V. Singh, et.al.,"Optimizing user association and activation fractions in heterogeneous wireless networks," in *IEEE WiOpt*, June. 2015.

[20] Y. Ghasempour, et.al., "Managing Analog Beams in mmWave Networks," *Extended Version*, https://www.dropbox.com/home/Papers-NEC?preview=BeamManagementISIT17L.pdf Jan. 2017