Low-Complexity Beam Allocation for Switched-Beam Based Multiuser Massive MIMO Systems

Junyuan Wang, *Member, IEEE*, Huiling Zhu, *Member, IEEE*,
Lin Dai, *Senior Member, IEEE*, Nathan J. Gomes, *Senior Member, IEEE*, and
Jiangzhou Wang, *Senior Member, IEEE*

Abstract

This paper addresses the beam allocation problem in a switched-beam based massive multiple-input-multiple-output (MIMO) system working at the millimeter wave (mmWave) frequency band, with the target of maximizing the sum data rate. This beam allocation problem can be formulated as a combinatorial optimization problem under two constraints that each user uses at most one beam for its data transmission and each beam serves at most one user. The brute-force search is a straightforward method to solve this optimization problem. However, for a massive MIMO system with a large number of beams N, the brute-force search results in intractable complexity $O(N^K)$, where K is the number of users. In this paper, in order to solve the beam allocation problem with affordable complexity, a suboptimal low-complexity beam allocation (LBA) algorithm is developed based on submodular optimization theory, which has been shown to be a powerful tool for solving combinatorial optimization problems. Simulation results show that our proposed LBA algorithm achieves nearly optimal sum data

This paper was presented in part at the IEEE International Conference on Communications (ICC), London, June 2015.

J. Wang, H. Zhu, N. J. Gomes, and J. Wang are with the School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NT, UK (email: {jw712, h.zhu, n.j.gomes, j.z.wang}@kent.ac.uk).

L. Dai is with the Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China (email: lindai@cityu.edu.hk).

rate with complexity $O(K \log N)$. Furthermore, the average service ratio, i.e., the ratio of the number of users being served to the total number of users, is theoretically analyzed and derived as an explicit function of the ratio N/K.

Index Terms

Switched-beam based systems, beam allocation algorithm, sum data rate, submodular optimization, service ratio, massive multiple-input-multiple-output (MIMO)

I. Introduction

The rapid growth of smartphone users and high-data-rate applications such as online-gaming and streaming high-definition video has imposed ever-increasing high-data-rate requirements on the fifth-generation (5G) mobile communication systems. In order to meet this challenge, there has been a great interest in moving to millimeter wave (mmWave) spectrum where a wide range of bandwidth is available, which also enables deploying a massive number of antennas at the base-station (BS) [1]–[10].

With a massive number of antenna elements at the BS (popularly known as *massive multiple-input-multiple-output (MIMO)*), a few advantages can be harvested, e.g., the random channel vectors from users to the BS become pairwisely orthogonal, the effect of small-scale fading can be averaged out, and the transmit power can be reduced significantly [9], [10]. Thanks to the aforementioned benefits, massive MIMO has been also adopted in wireless sensor networks recently [11], [12]. Moreover, by employing massive MIMO in mmWave communication systems, narrow and high-gain beams can be formed to overcome the severe propagation loss of mmWave signals for establishing reliable links. Therefore, massive MIMO beamforming has been seen as a promising key technology for 5G cellular networks [1].

Generally, beamforming technologies can be divided into two categories: digital beamforming [13] and analog beamforming [14]. Digital beamforming is often employed in the conventional communication systems where beams are formed by digitally adjusting the amplitudes and phases of transmitted signals, which is flexible and programmable. However, performing digital beamforming requires individual radio frequency (RF) chain for each antenna element. For the massive MIMO systems, even though with very good performance, deploying a massive number of RF chains to enable digital beamforming is of high cost and with high power consumption at mmWave frequencies. By contrast, analog beamforming is implemented just by using a number of independent phase shifters on the BS antennas at the RF part, i.e., only one RF chain is needed for a single data stream, which is much simpler and cheaper. Therefore, more and more attention has been paid on the hybrid analog-digital beamforming [15]–[20] and pure analog beamforming [21]–[23].

For analog beamforming based systems, a switched-beam scheme along with beam selection turned to be a popular technique for data transmission through beams due to its simplicity [24]–[32]. In this scheme, a fixed number of beams are generated and pointed to different predetermined directions to cover the whole cell. The Butler method is a representative method to create fixed beams [25]. In the switched-beam based multiuser systems, each user uses only one beam to transmit and/or receive its data signal and switches its beam from one time slot to another according to the strength of the line-of-sight (LOS) signal to achieve efficient beam switching. As the switched-beam scheme requires LOS, it has been mainly investigated in satellite communication systems [26]–[28]. For cellular systems, the study was limited in wideband code division multiple access (WCDMA) mobile networks [29], [30]. Recently, due to the potential of high capacity by exploiting the mmWave bands, the switched-beam scheme has been also studied in the mmWave communication systems where LOS often exists [31], [32].

For such a switched-beam based multiuser system, a key challenge is how to assign multiple beams to multiple users to achieve a high sum data rate. This beam allocation problem shares similarities to that in the conventional random beamforming based systems [33]–[36]. Specifically, in a random beamforming based system, to exploit multiuser diversity gain, the number of users K is assumed to be much larger than the number of beams N, i.e., $K \gg N$. By assuming that all the N beams are used with equal power allocation, each user can calculate the received signal-to-interference-plus-noise ratios (SINRs) on the N beams and then feeds back the SINRs and the corresponding beam indices to the BS. After receiving feedback from all users on all beams, the BS assigns each beam to the best user with the highest SINR to maximize the sum data rate [33]–[35]. In this scenario, the beam allocation problem is reduced to N independent user selection problems, i.e., selecting the proper user for each beam to serve.

However, the beam allocation algorithms in conventional random beamforming based systems [33]–[35] cannot be directly adopted in switched-beam based massive MIMO systems. The main reason is that in massive MIMO systems, the number of beams can be much larger than the number of users, i.e., $N \gg K$, and thus some of the beams may not be used for data transmission, and the beams used for data transmission could vary when the channel condition changes. As a result, it is impossible for each user to calculate its received SINR on each beam due to the lack of information of active beams. Moreover, even under the condition $N \ll K$, it is shown in [36] that using all the beams for transmission is not always optimal due to strong inter-beam interference. Several suboptimal beam selection algorithms were further proposed to maximize the sum data rate [36], among which the simplest greedy algorithm still has the complexity $O(KN^2)$, which could be very high when the number of beams N is large.

This paper aims to develop a low-complexity beam allocation algorithm to maximize the sum data rate of a switched-beam based multiuser massive MIMO system where a massive number of N fixed beams are formed by using the Butler method with a uniform linear array of N antenna elements to serve K users, i.e., $N \gg K$. The beam allocation problem is formulated as a combinatorial optimization problem with two constraints including that each user can be served at most by one beam and each beam can serve at most one user. As the brute-force search leads to the complexity $O(N^K)$ to obtain the optimal solution, which is intractable when the number of beams N is large, in this paper, a suboptimal low-complexity beam allocation (LBA) algorithm is proposed based on submodular optimization theory, which is a powerful tool for solving combinatorial optimization problems [37]. Specifically, the original optimization problem is first reformulated as a non-monotone submodular maximization problem under two partition matroid constraints, which still has high computational complexity due to the non-monotone objective function. To reduce the complexity, the non-monotone submodular maximization problem is further decoupled into two sub-problems, including a beam-user association sub-problem and a beam allocation sub-problem which is a monotone submodular maximization problem subject to a single partition matroid constraint and can be efficiently solved by a greedy algorithm. The LBA algorithm is then proposed by combining the solutions of these two sub-problems. Simulation results show that compared with the optimal brute-force search, our LBA algorithm can achieve nearly the same sum data rate, but only with the complexity $O(K \log N)$.

Note that to maximize the sum data rate, some users might not be served, which causes delay for the unserved users. It is therefore of great importance to study how many users can be simultaneously served by the system. This performance is indicated in the paper by service ratio, which is defined as the ratio of the number of served users to the total number of users. An explicit expression of the average service ratio (i.e., the service ratio averaged over users' positions) is obtained and shown to be a monotonic increasing function of the ratio of the number of beams N to the number of users K. Simulation results verify that the analytical result serves as a good approximation of the average service ratio, which sheds important insights on the service delay.

The remainder of this paper is organized as follows. Section II introduces the system model and problem formulation. A low-complexity beam allocation algorithm is proposed in Section III, followed by the simulation results and discussions provided in Section IV. Concluding remarks are summarized in Section V.

Throughout this paper, $\mathbb{E}[\cdot]$ denotes the expectation operator. $x \sim \mathcal{CN}(u, \sigma^2)$ denotes a complex Gaussian random variable with mean u and variance σ^2. $|X|$ denotes the cardinality of set X. 2^X denotes the power set of set X. $X \cap Y$ and $X \cup Y$ denote the intersection and union of set X and set Y, respectively. $X \setminus Y$ denotes the relative compliment of set Y in set X. $\emptyset$ denotes the empty set. $\binom{n}{k}$ denotes a binomial coefficient, i.e., the number of ways to choose k elements from a set of n distinct elements. $\left\{ {n \atop k} \right\}$ denotes a Stirling number of the second kind, i.e., the number of ways to partition a set of n distinct elements into k non-empty subsets.

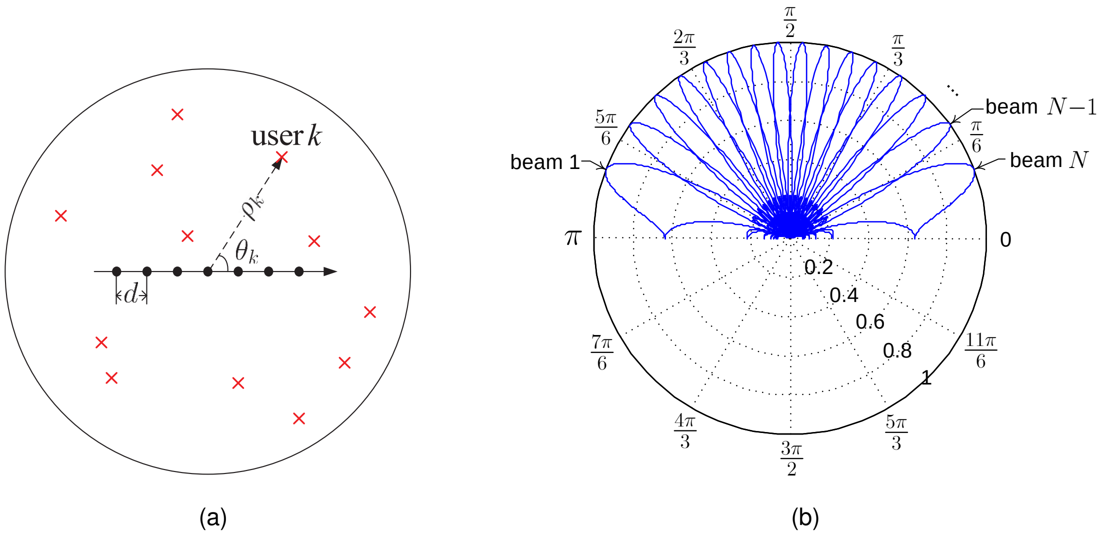


Fig. 1. Beamforming model. (a) Illustration of massive MIMO system. d denotes the uniform BS antenna elements spacing. (ρ_k, θ_k) is the polar coordinate of user k. "x" represents a user and "•" represents a BS antenna element. (b) Array pattern generated by using the Butler method. $N = 16$.

II. System Model and Problem Formulation

As shown in Fig. 1(a), the downlink transmission is considered for a multiuser switched-beam based system with K users and a BS with a linear array of N equally spaced identical isotropic antenna elements to form N fixed beams. It is assumed that K users are uniformly distributed within a circular cell with unit radius, and each user is equipped with a single antenna.[1] The BS is located at the center of the cell and all the BS antenna elements are equally spaced at distance $d = 0.5\lambda$, where λ is the propagation wavelength. (ρ_k, θ_k) denotes the location of user k.

By applying the Butler method to form the beams with $N = 2^i$ (where $i \geq 1$ is an integer), the normalized array factor of any beam n, $n = 1, 2, \cdots, N$, with respect to an angle of departure (AoD) θ of signal is given by [24]

$$A_n(\theta) = \frac{\sin(0.5N\pi\cos\theta - \beta_n)}{N\sin(0.5\pi\cos\theta - \frac{1}{N}\beta_n)}, \tag{1}$$

[1] Note that the following analysis can be also applied to the case where each user is equipped with multiple antennas by incorporating the receive beam gain into the received signal power model given by (3).

where

$$\beta_n = \left(-\frac{(N+1)}{2} + n \right) \pi. \tag{2}$$

Fig. 1(b) illustrates the array pattern generated according to (1) and (2) when $N = 16$, where the beam index increases from 1 to N from the left-hand side to the right-hand side.

Consider user k as the reference user. By assuming an LOS channel at the mmWave frequencies, the AoD of the received signal at user k is θ_k and the corresponding received power can be written as [38]

$$P_k = \sum_{n=1}^{N} c_{k,n} \cdot p_n \cdot D_n(\theta_k) \cdot \rho_k^{-\alpha}, \tag{3}$$

where p_n denotes the transmit power allocated on beam n. ρ_k is the distance from user k to the cell center and α is the path-loss exponent. $D_n(\theta)$ denotes the directivity[2] of beam n with regard to an AoD θ, given by [24]

$$D_n(\theta) = \frac{2\left[A_n(\theta)\right]^2}{\int_0^{\pi} \left[A_n(\psi)\right]^2 \sin(\psi) d\psi}. \tag{4}$$

Appendix A shows that (4) can be further reduced to (A.5), i.e.,

$$D_n(\theta) = N\left[A_n(\theta)\right]^2. \tag{5}$$

In (3), $c_{k,n} \in \{0, 1\}$ denotes the beam allocation indicator. If beam n is allocated to user k, $c_{k,n} = 1$; otherwise, $c_{k,n} = 0$. With beam switching, each user can only use one beam for its data transmission, i.e., $\sum_{n=1}^{N} c_{k,n} \leq 1$. Moreover, to avoid severe intra-beam interference, each beam can be used at most by one user, i.e., $\sum_{k=1}^{K} c_{k,n} \leq 1$. For a massive MIMO system where the number of beams N is much larger than the number of users K, only some of the beams will be used for data transmission, as illustrated in Fig. 2. Let N_s denote the number of allocated beams, given by

$$N_s = \sum_{k=1}^{K} \sum_{n=1}^{N} c_{k,n}, \tag{6}$$

[2] The directivity is a measure of how directive an individual antenna is relative to an isotropic antenna radiating the same total power. In other words, the directivity is the ratio of the power density of an anisotropic antenna relative to an isotropic antenna radiating the same total power [39].

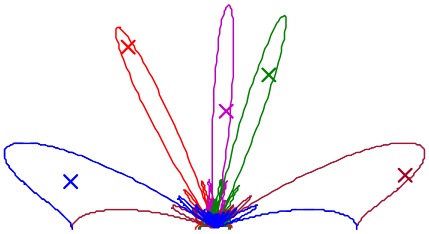


Fig. 2. Illustration of beam allocation in switched-beam based massive MIMO systems. “x” represents a user. A beam is allocated to the user in the same color. $N = 16$, $K = 5$.

which is no larger than K as each user can only select at most one beam for its data transmission. Assume that the total transmit power is fixed at P_t, and equally allocated to the beams selected for data transmission. The transmit power allocated on beam n is then given by

$$p_n = \begin{cases} \frac{P_t}{N_s}, & \text{if} \quad \sum_{k=1}^{K} c_{k,n} = 1, \\ 0, & \text{if} \quad \sum_{k=1}^{K} c_{k,n} = 0. \end{cases} \tag{7}$$

By assuming that the total system bandwidth is normalized to unity, the achievable data rate of user k can be written as

$$R_k = \log_2 \left(1 + \frac{P_k}{\sigma_0^2 + I_k} \right). \tag{8}$$

where σ_0^2 is the variance of the additive white Gaussian noise (AWGN), and I_k is the inter-beam interference power received at user k, which is given by

$$I_k = \sum_{j=1, j \neq k}^{K} \sum_{n=1}^{N} c_{j,n} \cdot p_n \cdot D_n(\theta_k) \cdot \rho_k^{-\alpha}. \tag{9}$$

In this paper, we aim at developing a beam allocation algorithm for maximizing the sum data rate of switched-beam based massive MIMO systems, which can be formulated as

$$\max_{\{c_{k,n}:\ k=1,2,\cdots,K;\ n=1,2,\cdots,N\}} \quad \sum_{k=1}^{K} R_k \tag{10a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_{k,n} \leq 1, \forall k \in \{1, 2, \cdots, K\}, \tag{10b}$$

$$\sum_{k=1}^{K} c_{k,n} \leq 1, \forall n \in \{1, 2, \cdots, N\}, \tag{10c}$$

$$c_{k,n} \in \{0, 1\}, \forall k \in \{1, 2, \cdots, K\}, \forall n \in \{1, 2, \cdots, N\}, \tag{10d}$$

where (10b) and (10c) follow the constraints that each user can select at most one beam to transmit, and each beam can be used by at most one user to avoid severe intra-beam interference, respectively.

For the combinatorial optimization problem given by (10a)–(10d), a brute-force search among all $(N+1)^K$ possible beam allocations leads to unaffordably high complexity in massive MIMO systems with a large number of beams N. In the literature [40]–[44], a widely adopted method for solving combinatorial optimization problems is to relax the indicator $c_{k,n}$ into a continuous variable between 0 and 1, and convert the objective function into a convex one, which could then be efficiently solved by the convex optimization algorithms. In our case, however, there are $N \times K$ indicators to be optimized, which still results in prohibitively high computational complexity for large N even with relaxation. As a result, in this paper, we resort to submodular optimization, which has been shown to be a powerful tool for solving combinatorial optimization problems [37]. In the following section, the beam allocation problem will be reformulated into a submodular maximization problem subject to matroid constraints.

III. Beam Allocation Design Based on Submodular Optimization

Before reformulating our beam allocation problem into a submodular optimization problem, let us first present the definitions of submodular functions and matroids given in [37] as follows.

A. Basic Definitions

Definition 1. *Let U be a finite ground set, and 2^U be the power set of U (i.e., the set of all subsets of U, including the empty set and U itself). A set function $f(S)$ with the input $S \subseteq U$ (i.e., $S \in 2^U$) and a real value output, denoted by $f : 2^U \to \mathbb{R}$, is said to be submodular if*

$$f(S) + f(T) \geq f(S \cap T) + f(S \cup T), \tag{11}$$

for any $S, T \subseteq U$. An equivalent definition of a submodular function is that

$$f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T), \tag{12}$$

for any $S \subseteq T \subseteq U$ and $e \in U \setminus T$, i.e., the marginal gain of adding an extra element in the set decreases with the size of the set. Intuitively, if a set function is submodular, its marginal gain is diminishing when increasing the set size by adding more elements into it.

In particular, a set function $f(S)$ is monotone if

$$f(S) \leq f(T), \tag{13}$$

for any $S \subseteq T \subseteq U$.

Definition 2. *A matroid $\mathcal{M}$ is a pair $(U, \mathcal{I})$, denoted by $\mathcal{M} = (U, \mathcal{I})$, where U is a finite ground set and $\mathcal{I} \subseteq 2^U$ is a collection of subsets of U with the following properties:*
(1) $\mathcal{I}$ is nonempty, in particular, $\emptyset \in \mathcal{I}$;
(2) $\mathcal{I}$ is downward closed; i.e., for each $X \subseteq Y \in \mathcal{I}$, we have $X \in \mathcal{I}$;
(3) If $X, Y \in \mathcal{I}$ and $|X| > |Y|$, then there exists an element $e \in X \setminus Y$ such that $Y \cup \{e\} \in \mathcal{I}$.
Note that $\mathcal{I}$ is also called independent sets. Intuitively, for any set $X \in \mathcal{I}$, the elements in X are independent of each other.

Particularly, a partition matroid is a matroid where the ground set U is partitioned into some disjoint sets, $U_1, U_2, \cdots, U_l$, and

$$\mathcal{I} = \{X \subseteq U : |X \cap U_i| \leq w_i, \textit{ for all } i = 1, 2, \cdots, l\}, \tag{14}$$

for some given parameters $w_1, w_2, \cdots, w_l$.

Based on the above definitions, the original optimization problems given by (10a)–(10d) will be reformulated in the following subsection.

B. Problem Reformulation

Let us first define the ground set U as

$$U = \{u_{1,1}, u_{1,2}, \cdots, u_{1,N}; u_{2,1}, u_{2,2}, \cdots, u_{2,N}; \cdots; u_{K,1}, u_{K,2}, \cdots, u_{K,N}\}, \tag{15}$$

and the beam allocation set S as a subset of U such that $u_{k,n} \in S$ if beam n is allocated to user k, i.e., $c_{k,n} = 1, \forall k, n$; otherwise, $u_{k,n} \notin S$. For any beam allocation set $S \subseteq U$, the objective function of (10a) can then be written as

$$R_s(S) = \sum_{u_{k,n} \in S} \log_2 \left(1 + \frac{\frac{P_t}{|S|} D_n(\theta_k) \rho_k^{-\alpha}}{\sigma_0^2 + \sum_{u_{j,l} \in S, j \neq k} \frac{P_t}{|S|} D_l(\theta_k) \rho_k^{-\alpha}} \right), \tag{16}$$

according to (3) and (6)–(9).

The constraints can be written as an intersection of two partition matroids on the ground set U. Specifically, let us partition the ground set U into K disjoint subsets, $U_1^U, U_2^U, \cdots, U_K^U$, where $U_k^U = \{u_{k,1}, u_{k,2}, \cdots, u_{k,N}\}$ is the set containing all the possible beam allocations of user k, and the superscript denotes that the ground set is partitioned according to the user index. Since the beam allocation indicator $c_{k,n} = 1$ if $u_{k,n}$ belongs to the beam allocation set S, i.e., $u_{k,n} \in S$, the constraint given in (10b) can be written as $S \in \mathcal{I}_U$, where

$$\mathcal{I}_U = \{X \subseteq U : |X \cap U_k^U| \leq 1,\ k = 1, 2, \cdots, K\}. \tag{17}$$

According to the definition given by (14), it is clear from (17) that $\mathcal{M}_U = (U, \mathcal{I}_U)$ is a partition matroid. Similarly, by partitioning the ground set U into N disjoint subsets according to the beam index, i.e., $U_n^B = \{u_{1,n}, u_{2,n}, \cdots, u_{K,n}\}$, $n = 1, 2, \cdots, N$, constraint (10c) can also be written as a partition matroid constraint, $S \in \mathcal{I}_B$, with

$$\mathcal{I}_B = \{X \subseteq U : |X \cap U_n^B| \leq 1,\ n = 1, 2, \cdots, N\}. \tag{18}$$

$\mathcal{M}_B = (U, \mathcal{I}_B)$ denotes the corresponding partition matroid. The optimization problem formulated in (10a)–(10d) can then be rewritten as

$$\max_{S \subseteq U} \quad \sum_{u_{k,n} \in S} \log_2 \left(1 + \frac{\frac{P_t}{|S|} D_n(\theta_k) \rho_k^{-\alpha}}{\sigma_0^2 + \sum_{u_{j,l} \in S, j \neq k} \frac{P_t}{|S|} D_l(\theta_k) \rho_k^{-\alpha}} \right) \tag{19a}$$

$$\text{s.t.} \quad S \in \mathcal{I}_U, \tag{19b}$$

$$S \in \mathcal{I}_B. \tag{19c}$$

Appendix B shows that the objective function of (19a) is not submodular due to the existence of the inter-beam interference. By ignoring the inter-beam interference, it can be further relaxed into the following form with a submodular objective function:[3]

$$\max_{S \subseteq U} \quad \sum_{u_{k,n} \in S} \log_2 \left(\frac{P_t}{|S|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) \tag{20a}$$

$$\text{s.t.} \quad S \in \mathcal{I}_U, \tag{20b}$$

$$S \in \mathcal{I}_B. \tag{20c}$$

The proof of the submodularity of the objective function in (20a) can be found in Appendix B.

For maximization problems with a non-negative submodular objective function and multiple matroid constraints, a local-search based algorithm has been proposed in [45], which achieves at least $\frac{1}{p+2+\frac{1}{p}+\epsilon}$ of the optimal result, where $\epsilon > 0$ is a given parameter with small value and p is the number of matroid constraints. It is also shown in [45] that this algorithm requires at most $O\left(\frac{1}{\epsilon} p q^4 \log q\right)$ local operations, where q is the size of the ground set. In our case, with the size of the ground set $q = |U| = KN$ and the number of matroid constraints $p = 2$, the number of required local operations is $O\left(\frac{1}{\epsilon}(KN)^4 \log(KN)\right)$, which still results in high complexity when the number of beams N is large. In the following subsection, the beam allocation problem given by (20a)–(20c) will be decoupled into two sub-problems, based on which a low-complexity beam allocation algorithm will be proposed.

C. Low-Complexity Beam Allocation (LBA)

1) Problem Decomposition: In this subsection, the beam allocation optimization problem given by (20a)–(20c) will be decoupled into two sub-problems: (1) beam-user association for each user, and (2) beam allocation.

For the first sub-problem, let us define S_k^U as a subset of $U_k^U = \{u_{k,1}, u_{k,2}, \cdots, u_{k,N}\}$ of user k. For each user, we aim at maximizing its achievable data rate by properly choosing S_k^U. The

[3] It will be shown in Section IV-A that although the inter-beam interference is neglected in the objective function of (20a), the algorithm developed based on the relaxed objective function can still achieve nearly optimal sum data rate.

corresponding optimization problem can be written as

$$\max_{S_k^U \subseteq U_k^U} \quad \log_2\left(\frac{P_t}{|S|\sigma_0^2} D_n(\theta_k)\rho_k^{-\alpha}\right) \tag{21a}$$

$$\text{s.t.} \quad S_k^U \in \mathcal{I}_U, \tag{21b}$$

where (21b) follows the constraint given in (20b), i.e., each user can be only associated with one beam.

In the objective function of (21a), only the directivity $D_n(\theta_k)$ depends on S_k^U. Therefore, we have

$$S_k^{U*} = \{u_{k,\pi(k)}\}, \tag{22}$$

where the index of user k's associated beam, $\pi(k)$, is given by

$$\pi(k) = \arg \max_{n=1,2,\cdots,N} D_n(\theta_k). \tag{23}$$

That is, each user is associated with the beam with the largest directivity.

Based on the solution of the first beam-user association sub-problem, the ground set of possible beam allocations is reduced to

$$U' = \bigcup_{k=1,2,\cdots,K} S_k^{U*}. \tag{24}$$

Then the second beam allocation sub-problem can be written as

$$\max_{S \subseteq U'} \quad \sum_{u_{k,n} \in S} \log_2\left(\frac{P_t}{|S|\sigma_0^2} D_n(\theta_k)\rho_k^{-\alpha}\right) \tag{25a}$$

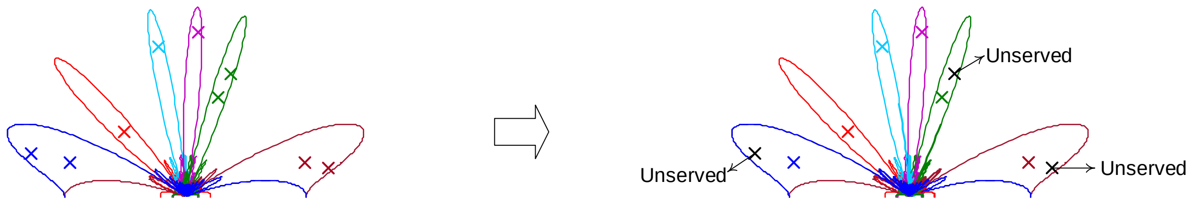
$$\text{s.t.} \quad S \in \mathcal{I}', \tag{25b}$$

where set $\mathcal{I}'$ of matroid $\mathcal{M}' = (U', \mathcal{I}')$ is given by

$$\mathcal{I}' = \{X \subseteq U' : |X \cap U_n'| \leq 1, n = 1, 2, \cdots, N\}, \tag{26}$$

with $U_n' = U' \cap U_n^B$.

Appendix C shows that the objective function of (25a) is a monotone submodular function on the ground set U' when the transmit signal-to-noise ratio (SNR) $P_t/\sigma_0^2 > KN \sin^2 \frac{\pi}{2N} \left(\frac{K}{K-1}\right)^{K-1}$



Step 1: Each user is associated with the beam with the largest directivity.

Step 2: Each associated beam is allocated to the user with the highest received signal power.

Fig. 3. Illustration of the proposed LBA algorithm. “x” represents a user. $N = 16$, $K = 9$.

$\overset{\text{for large}K,N}{\approx}$ $6.707K/N$, where N is the number of beams and K is the number of users. For a massive MIMO system with $N \gg K$, the optimization problem given by (25a)–(25b) is a monotone submodular function maximization problem under a single matroid constraint, which can be effectively solved by a greedy algorithm that achieves at least $\frac{1}{2}$ of the optimal result [46]. Specifically, the algorithm starts with an empty beam allocation set S, and adds one element with the highest marginal gain at each step to set S while the updated S is still an element of $\mathcal{I}'$. Note that to maximize the objective function of (25a), elements in U' should be included in the beam allocation set S as far as possible while S still satisfies constraint (25b), i.e., each beam can only be used at most by one user. Therefore, we can conclude that the greedy algorithm is equivalent to allocating each associated beam to its best user with the largest received signal power.

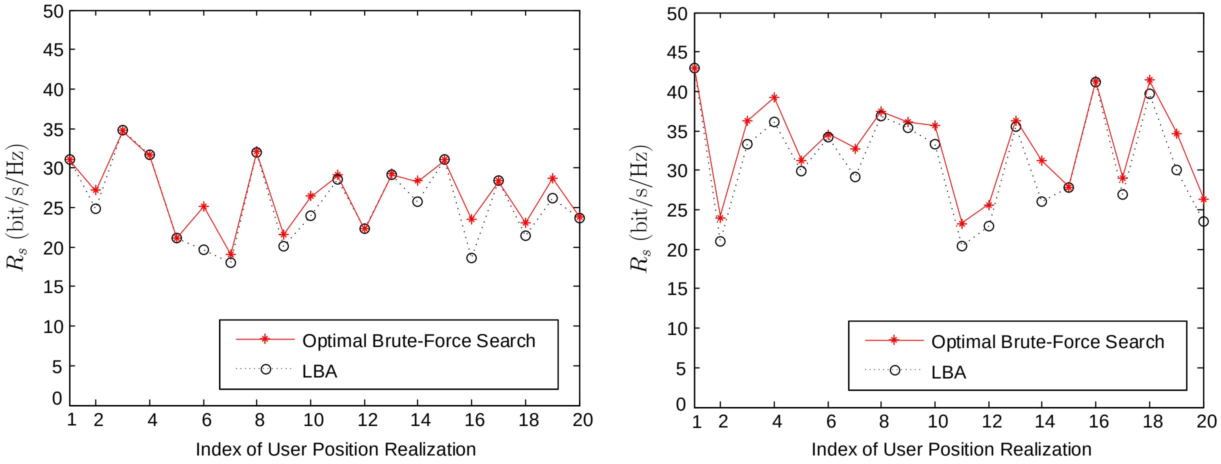
2) LBA Algorithm Design: By combining the solutions of the above two sub-problems, a two-step beam allocation algorithm is proposed: (1) Each user is associated with its best beam with the largest directivity; (2) Each associated beam is allocated to its best associated user with the largest received signal power. The detailed description of this two-step algorithm is presented in Algorithm 1. As Fig. 3 illustrates, in the first step, each user is associated with the beam with the highest directivity, which is highlighted in the same color. In the second step, if a beam is associated with more than one user (such as the blue beam), then it is allocated to the user with the highest received signal power.

It is clear from Algorithm 1 that in the first step of the proposed LBA algorithm, there are

Algorithm 1 Low-Complexity Beam Allocation (LBA)

```
**Initialization:** $S^* = \emptyset$. $\mathbf{v} = \mathbf{0}_{1\times N}$. $\mathbf{a} = \mathbf{0}_{1\times N}$.
**for** $k = 1$ to $K$ **do**
    $\pi(k) = \arg \max_{n=1,2,\cdots,N} D_n(\theta_k)$;
**end for**
**for** $k = 1$ to $K$ **do**
    **if** $v_{\pi(k)} = 0$ **then**
        $S^* = S^* \cup \{u_{k,\pi(k)}\}$;
        $v_{\pi(k)} = D_{\pi(k)}(\theta_k)\rho_k^{-\alpha}, \;\; a_{\pi(k)} = k$;
    **else**
        $k^{'} = a_{\pi(k)}$;
        **if** $D_{\pi(k)}(\theta_k)\rho_k^{-\alpha} > D_{\pi(k)}(\theta_{k'})\rho_{k'}^{-\alpha}$ **then**
            $S^* = S^* \setminus \{u_{k',\pi(k)}\} \cup \{u_{k,\pi(k)}\}$;
            $v_{\pi(k)} = D_{\pi(k)}(\theta_k)\rho_k^{-\alpha}, \;\; a_{\pi(k)} = k$;
        **end if**
    **end if**
**end for**
**Output:** $S^*$.
```

K iterations. In each iteration, the number of comparisons required to find the best beam of that user is $\log_2 N$ by adopting the binary search. In the second step, as there are K users in total, at most $K-1$ comparisons are needed when all the users are associated with the same beam. Therefore, the maximum number of comparisons required by our proposed LBA algorithm is $K \log_2 N + K - 1$. Compared with the optimal brute-force search, it can be clearly seen that by relaxing the objective function and decoupling the problem into two sub-problems, the computational complexity is significantly reduced from $O(N^K)$ to $O(K \log N)$.



(a) $K = 6$.

(b) $K = 10$.

Fig. 4. Achievable sum data rate R_s under 20 realizations of the users' positions with both the optimal brute-force search and the proposed LBA algorithm. $\alpha = 2.7$, $P_t/\sigma_0^2 = 20$dB, $N = 16$.

IV. Simulation Results and Discussions

In this section, simulation results will be presented to demonstrate the performance of the proposed LBA algorithm. As described in Section II, we assume that K users are uniformly distributed in a circular cell with unit radius, and a linear antenna array with N identical isotropic antenna elements is placed at the center of the cell.

A. Sum Data Rate

Fig. 4 presents the simulation results of the achievable sum data rate, $R_s \triangleq \sum_{k=1}^{K} R_k$, under 20 random realizations of the users' positions with the number of beams $N = 16$, and the number of users $K = 6$ and 10, respectively. It can be seen from Fig. 4 that our proposed LBA algorithm can achieve very close sum data rate to that of the optimal brute-force search under most realizations. Since the sum data rate closely depends on the locations of users $\{\mathbf{r}_k : k = 1, 2, \cdots, K\}$ as shown in Fig. 4, we further focus on the average sum data rate $\bar{R}_s \triangleq \mathbb{E}_{\{\mathbf{r}_k : k=1,2,\cdots,K\}} \left[\sum_{k=1}^{K} R_k\right]$, in the following discussion.

Specifically, Fig. 5 presents the simulation results of the average sum data rate $\bar{R}_s$ over 1000 realizations of users' positions for both the proposed LBA algorithm and the optimal brute-force

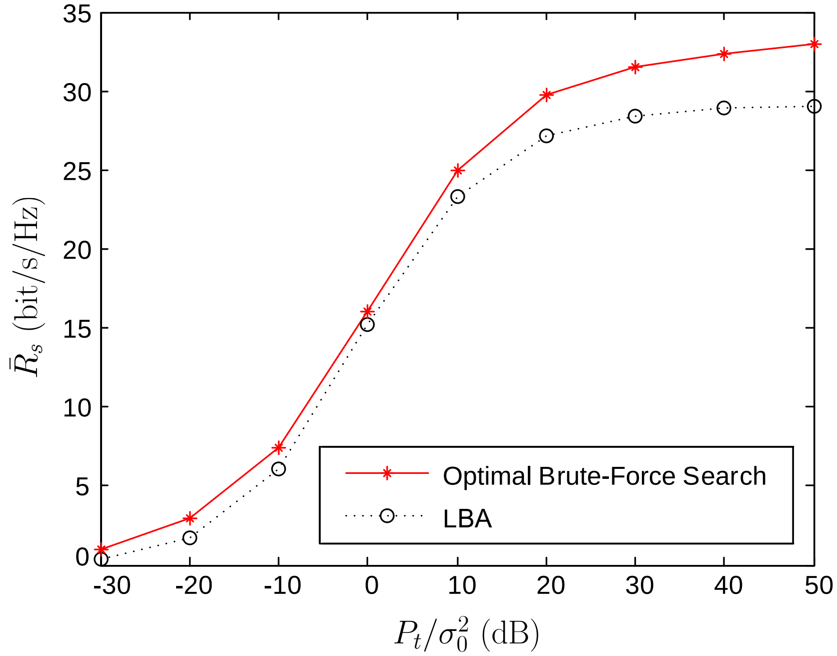
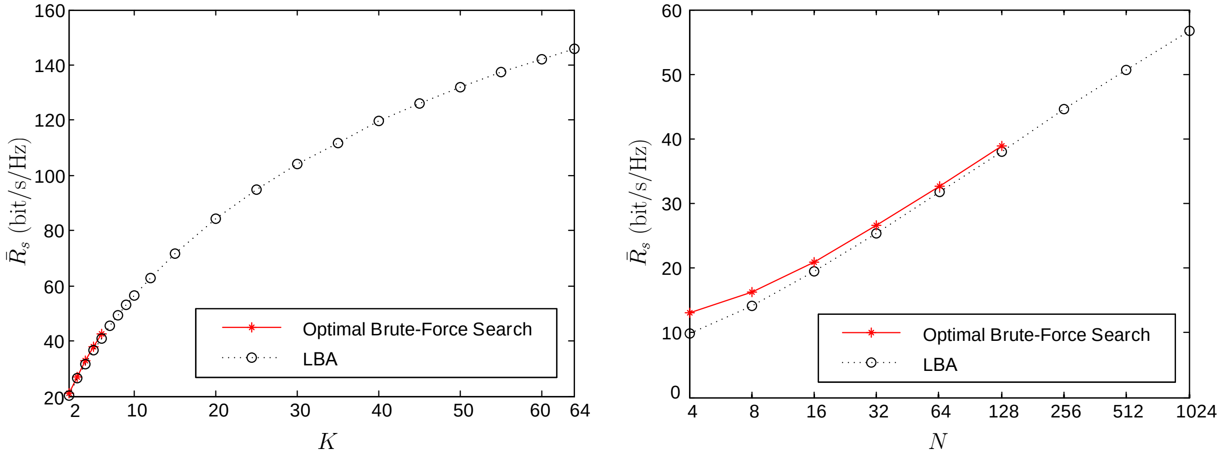


Fig. 5. Average sum data rate $\bar{R}_s$ with both the optimal brute-force search and the proposed LBA algorithm. $\alpha = 2.7$, $K = 8$, $N = 16$.

search under varying values of the transmit SNR P_t/σ_0^2. It can be clearly seen from this figure that the average sum data rate $\bar{R}_s$ first increases with P_t/σ_0^2, and then becomes saturated for large P_t/σ_0^2 as the system becomes interference-limited. A small gap of the average sum data rate between the optimal brute-force search and the proposed LBA algorithm can be observed for smaller values of P_t/σ_0^2, while the gap becomes noticeable when the transmit SNR P_t/σ_0^2 is higher than 20dB. In this case, as the system operates at the interference-limited region, the rate gap becomes significant due to ignoring the effect of inter-beam interference in our proposed LBA algorithm.

Fig. 6 further presents the simulation results of the average sum data rate $\bar{R}_s$ with both the optimal brute-force search and our proposed LBA algorithm under varying values of the number of users K and the number of beams N when the transmit SNR P_t/σ_0^2 is fixed at 20dB. Note that as the complexity of the brute-force search algorithm $O(N^K)$ sharply increases with K and N, we only present the optimal average sum data rate for $K \leq 6$ in Fig. 6(a), and for $N \leq 128$



(a) $N = 64$.

(b) $K = 4$.

Fig. 6. Average sum data rate $\bar{R}_s$ with both the optimal brute-force search and the proposed LBA algorithm. $\alpha = 2.7$, $P_t/\sigma_0^2 = 20$dB.

in Fig. 6(b). It can be seen from Fig. 6(a) that with the number of beams N fixed at 64, the average sum data rate $\bar{R}_s$ increases with the number of users K as more users can be served for a larger K. With the number of users K fixed at 4 in Fig. 6(b), the average sum data rate $\bar{R}_s$ logarithmically increases with the number of beams N thanks to the enhanced beam gain. It can be also clearly observed in both Fig. 6(a) and Fig. 6(b) that our proposed LBA algorithm achieves nearly the same average sum data rate as the optimal brute-force search. The gap is slightly enlarged when the number of beams N is small in Fig. 6(b) because with a small N, the beams are very wide and the overlap of two adjacent beams is large, which results in high inter-beam interference for each user.

B. Service Ratio

It should be noted that with the proposed LBA algorithm, as shown in Fig. 3, some users might not be served, since multiple users might be associated with the same beam, and only one of them can be served. In this section, how many users can be served simultaneously by using our LBA algorithm will be evaluated.

Specifically, let us first define the service ratio, P_s, as the ratio of the number of served users

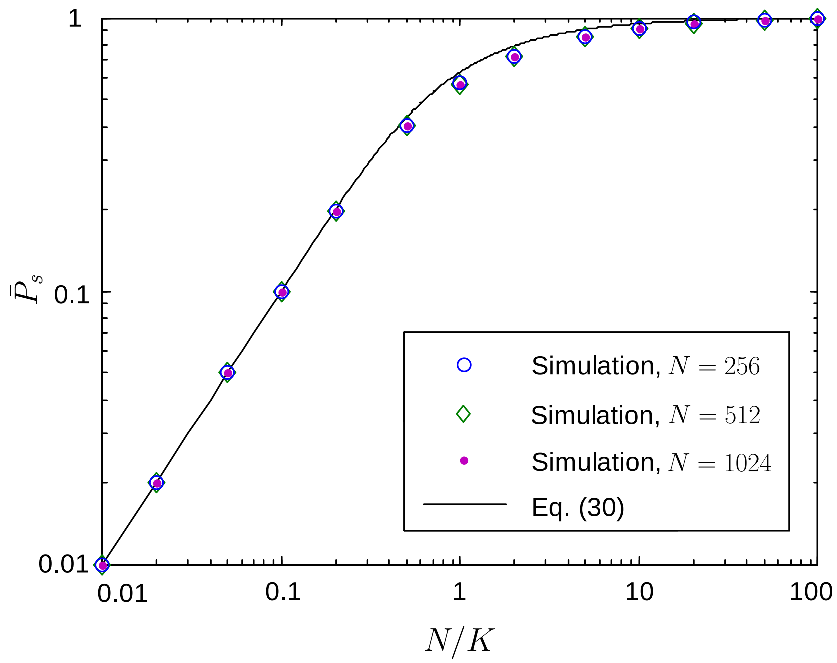


Fig. 7. Average service ratio $\bar{P}_s$ with the proposed LBA algorithm. $\alpha = 2.7$.

to the total number of users K, which is given by

$$P_s \triangleq \frac{N_s}{K}. \tag{27}$$

Here N_s is also the number of beams allocated for data transmission, which is given in (6). As the beam allocation result closely depends on the positions of users, similar to the average sum data rate, we further define the average service ratio as

$$\bar{P}_s \triangleq \mathbb{E}_{\{\mathbf{r}_k | k=1,2,\cdots,K\}} \left[P_s \right]. \tag{28}$$

Note that $\bar{P}_s$ sheds important light on the service delay performance as well. A larger $\bar{P}_s$ indicates that more users can be simultaneously served by the system, and thus shorter average service delay can be expected for each user.

Appendix D shows that by approximating the beam allocation problem as a ball-dropping problem, the average service ratio $\bar{P}_s$ can be obtained as

$$\bar{P}_s = \sum_{m=1}^{K} \frac{m}{K} \cdot \frac{\binom{N}{m}}{N^K} \sum_{j=1}^{m} (-1)^{m-j} \binom{m}{j} m^K, \tag{29}$$

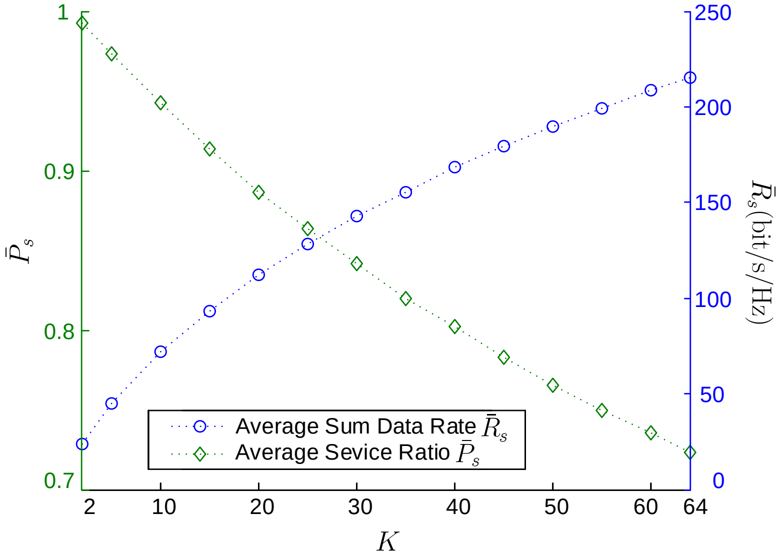


Fig. 8. Average sum data rate $\bar{R}_s$ and average service ratio $\bar{P}_s$ with the proposed LBA algorithm. $\alpha = 2.7$, $P_t/\sigma_0^2 = 20\text{dB}$, $N = 64$.

and for a large number of beams N, we have

$$\bar{P}_s \overset{\text{for large } N}{\approx} \frac{N}{K}\left(1 - e^{-\frac{K}{N}}\right). \tag{30}$$

(30) indicates that the average service ratio $\bar{P}_s$ is solely determined by the ratio of the number of beams N to the number of users K. As Fig. 7 illustrates, the average service ratio $\bar{P}_s$ increases as the ratio N/K increases. Intuitively, with a larger N/K, the average number of beams that could be used by each user increases, leading to a higher $\bar{P}_s$. Specifically, for a small $N/K \ll 1$, it can be easily obtained from (30) that $\bar{P}_s \approx N/K$. As $N/K \to \infty$, the average service ratio $\bar{P}_s \to 1$. We can clearly see from Fig. 7 that the average service ratio is above 0.9 when N/K is larger than 5, which indicates that over 90% of the users can be simultaneously served. The analysis is verified by the simulation results in Fig. 7.

Recall that it is shown in Fig. 6(b) that the average sum data rate $\bar{R}_s$ increases as the number of beams N increases. Therefore, with the number of users K fixed, both the sum data rate and the service ratio can be improved by increasing the number of beams N. By contrast, for a given number of beams N, although the average sum data rate $\bar{R}_s$ increases with the number of users K as shown in Fig. 6(a), the average service ratio $\bar{P}_s$ decreases as K increases. In this case, the number of users K determines a trade-off between the sum data rate and the service ratio, which can be clearly observed from Fig. 8.

V. Conclusion

In this paper, the beam allocation problem in switched-beam based massive MIMO systems has been studied to maximize the sum data rate. Based on submodular optimization theory, a low-complexity suboptimal beam allocation algorithm has been proposed, which achieves nearly optimal sum data rate with complexity $O(K\log N)$ with respect to the number of users K and the number of beams N. As some users might not be served for the sake of maximizing the sum data rate, the average service ratio, i.e., the average percentage of users that can be served simultaneously, is further evaluated. The analysis shows that the average service ratio is solely determined by the ratio of the number of beams N to the number of users K. Simulation results corroborate that the average service ratio can be greatly improved by increasing the ratio N/K, which is above 90% when N/K exceeds 5.

Note that in this paper, as we aim at maximizing the sum data rate from the system's perspective, not all the users can be simultaneously served. For practical systems, however, it is equally important to serve as many users as possible and consider the fairness issue among users. To balance the sum data rate and service ratio, and ensure fairness among users with varying locations, some service ratio constraint and individual data rate requirements for the users can be added into the beam allocation problem, which deserves much attention and should be carefully investigated in the future work.

Appendix A
Derivation of (5)

From (1), the denominator of (4) can be written as

$$\int_0^{\pi} \frac{\sin^2(0.5N\pi\cos\psi - \beta_n)}{N^2\sin^2\left(0.5\pi\cos\psi - \frac{\beta_n}{N}\right)}\sin\psi d\psi \overset{\phi=0.5\pi\cos\psi-\frac{\beta_n}{N}}{=} \frac{2}{\pi}\int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \frac{\sin^2(N\phi)}{N^2\sin^2\phi}d\phi. \tag{A.1}$$

As N is a power of 2 and $\sin(2x) = 2\sin x \cos x$, we have

$$\int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \frac{\sin^2(N\phi)}{N^2\sin^2\phi} d\phi = \int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \prod_{k=0}^{\log_2 N-1} \cos^2(2^k\phi) d\phi, \tag{A.2}$$

which can be further expanded by substituting $\cos^2\left(2^k\phi\right) = \frac{\cos\left(2^{k+1}\phi\right)+1}{2}$ into it as follows

$$\begin{aligned}\int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \frac{\sin^2(N\phi)}{N^2\sin^2\phi} d\phi &= \int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \frac{1}{N} \prod_{k=1}^{\log_2 N} \left[\cos(2^k\phi)+1\right] d\phi \\ &= \frac{1}{N} \int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \left(1 + \sum_{k=1}^{\log_2 N-1} \cos\left(2^k\phi\right) \cdot \prod_{i=k+1}^{\log_2 N} \left[1+\cos\left(2^i\phi\right)\right]\right) d\phi \\ &= \frac{1}{N}\left(\pi + \sum_{k=1}^{\log_2 N-1} \int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \cos\left(2^k\phi\right) \cdot \prod_{i=k+1}^{\log_2 N} \left[1+\cos\left(2^i\phi\right)\right] d\phi\right). \end{aligned} \tag{A.3}$$

As $\int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \cos\left(2^k\phi\right) \cdot \prod_{i=k+1}^{\log_2 N} \left[1+\cos\left(2^i\phi\right)\right] d\phi = 0$ for any $k \geq 1$, it is easily obtained from (A.3) that

$$\int_{-\frac{\pi}{2}-\frac{\beta_n}{N}}^{\frac{\pi}{2}-\frac{\beta_n}{N}} \frac{\sin^2(N\phi)}{N^2\sin^2\phi} d\phi = \frac{\pi}{N}. \tag{A.4}$$

By combining (4), (A.1) and (A.4), we have

$$D_n(\theta) = N\left[A_n(\theta)\right]^2. \tag{A.5}$$

APPENDIX B

PROOF OF NON-SUBMODULARITY OF OBJECTIVE FUNCTION IN (19A) AND SUBMODULARITY OF OBJECTIVE FUNCTION IN (20A) ON GROUND SET U

A. Proof of Non-submodularity of Objective Function in (19a) on Ground Set U

Proof: Recall that the objective function of (19a) is given by

$$R_s(S) = \sum_{u_{k,n}\in S} \log_2\left(1 + \frac{\frac{P_t}{|S|}D_n(\theta_k)\rho_k^{-\alpha}}{\sigma_0^2 + \sum_{u_{j,l}\in S, j\neq k} \frac{P_t}{|S|}D_l(\theta_k)\rho_k^{-\alpha}}\right). \tag{B.1}$$

To prove that $R_s(S)$ is not submodular, we only need to find a counter example where (11) does not hold.

Let us consider a special case: for any two sets $S, T \subset U$ with $|S| \geq 2$ and $|T| \geq 2$, their intersection $S \cap T = \{u_{k^*,n^*}\}$, and the transmit SNR $P_t/\sigma_0^2 \to \infty$. In this case, the inter-beam interference suffered by each user would be much larger than the noise for both set S and set T. Then the achievable sum data rate with respect to S and T can be obtained as

$$R_s(S) \overset{P_t/\sigma_0^2 \to \infty}{\rightarrow} \sum_{u_{k,n} \in S} \log_2 \left(1 + \frac{D_n(\theta_k)}{\sum_{u_{j,l} \in S, j \neq k} D_l(\theta_k)} \right), \tag{B.2}$$

and

$$R_s(T) \overset{P_t/\sigma_0^2 \to \infty}{\rightarrow} \sum_{u_{k,n} \in T} \log_2 \left(1 + \frac{D_n(\theta_k)}{\sum_{u_{j,l} \in T, j \neq k} D_l(\theta_k)} \right), \tag{B.3}$$

respectively, both of which have finite values. In contrast, for the intersection $S \cap T = \{u_{k^*,n^*}\}$, as only beam n^* is used by user k^* in the whole system, there is no inter-beam interference. Therefore, we have

$$R_s(S \cap T) = \log_2 \left(1 + \frac{P_t}{\sigma_0^2} D_{n^*}(\theta_{k^*}) \rho_{k^*}^{-\alpha} \right) \overset{P_t/\sigma_0^2 \to \infty}{\rightarrow} \infty. \tag{B.4}$$

Since the sum of $R_s(S)$ and $R_s(T)$ is finite, it is obvious that

$$R_s(S) + R_s(T) < R_s(S \cap T) \leq R_s(S \cap T) + R_s(S \cup T). \tag{B.5}$$

According to (11), we can conclude that the objective function of (19a) is not a submodular function on the ground set U. ■

B. Proof of Submodularity of Objective Function in (20a) on Ground Set U

Proof: Let $R_s^{ap}(S)$ denote the objective function of (20a) as

$$R_s^{ap}(S) = \sum_{u_{k,n} \in S} \log_2 \left(\frac{P_t}{|S|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right). \tag{B.6}$$

Then for any $S \subseteq T \subseteq U$, and $u_{j,l} \in U \setminus T$, we have

$$\begin{aligned} R_s^{ap}(S \cup \{u_{j,l}\}) - R_s^{ap}(S) = & \sum_{u_{k,n} \in S} \log_2 \left(\frac{P_t}{(|S|+1)\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) + \log_2 \left(\frac{P_t}{(|S|+1)\sigma_0^2} D_l(\theta_j) \rho_j^{-\alpha} \right) \\ & - \sum_{u_{k,n} \in S} \log_2 \left(\frac{P_t}{|S|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) \\ = & \log_2 \left(\left(\frac{|S|}{|S|+1} \right)^{|S|} \cdot \frac{P_t}{(|S|+1)\sigma_0^2} D_l(\theta_j) \rho_j^{-\alpha} \right). \end{aligned} \tag{B.7}$$

Similarly,

$$R_s^{ap}(T \cup \{u_{j,l}\}) - R_s^{ap}(T) = \log_2 \left(\left(\frac{|T|}{|T|+1} \right)^{|T|} \cdot \frac{P_t}{(|T|+1)\sigma_0^2} D_l(\theta_j) \rho_j^{-\alpha} \right). \tag{B.8}$$

To show that the objective function of (20a), i.e., $R_s^{ap}(S)$, is submodular, according to (12), we only need to show that (B.7) is larger than (B.8), i.e,

$$\frac{|S|^{|S|}}{(|S|+1)^{|S|+1}} \geq \frac{|T|^{|T|}}{(|T|+1)^{|T|+1}}, \tag{B.9}$$

which holds as $|S| \leq |T|$. ■

APPENDIX C

PROOF OF MONOTONE SUBMODULARITY OF OBJECTIVE FUNCTION IN (25A) ON GROUND SET $U^{'}$

Proposition 1. *The objective function of (25a) is a monotone submodular function on Ground Set $U^{'}$ if*

$$\frac{P_t}{\sigma_0^2} > KN \sin^2 \frac{\pi}{2N} \left(\frac{K}{K-1} \right)^{K-1}. \tag{C.1}$$

Proof: Denote the objective function of (25a) as $R_s^{ap}(S)$. Since $R_s^{ap}(S)$ is shown to be a submodular function on the ground set U in Appendix B and $U^{'}$ is a subset of U, $R_s^{ap}(S)$ is also submodular on the ground set U'.

Moreover, for any $S \subset T \subseteq U^{'}$, we have

$$\begin{aligned} R_s^{ap}(S) - R_s^{ap}(T) &= \sum_{u_{k,n} \in S} \log_2 \left(\frac{P}{|S|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) - \sum_{u_{k,n} \in T} \log_2 \left(\frac{P}{|T|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) \\ &= |S| \log_2 \left(\frac{|T|}{|S|} \right) - \sum_{u_{k,n} \in T \setminus S} \log_2 \left(\frac{P}{|T|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right). \end{aligned} \tag{C.2}$$

Let $u_{k',n'} = \arg\min_{u_{k,n} \in T \setminus S} \log_2 \left(\frac{P}{|T|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right)$, we can obtain that

$$\sum_{u_{k,n} \in T \setminus S} \log_2 \left(\frac{P}{|T|\sigma_0^2} D_n(\theta_k) \rho_k^{-\alpha} \right) \geq (|T| - |S|) \log_2 \left(\frac{P}{|T|\sigma_0^2} D_{n'}(\theta_{k'}) \rho_{k'}^{-\alpha} \right). \tag{C.3}$$

By substituting (C.3) to (C.2), we have

$$R_s^{ap}(S) - R_s^{ap}(T) \leq |S| \log_2 \left(\frac{|T|}{|S|} \right) - (|T| - |S|) \log_2 \left(\frac{P}{|T|\sigma_0^2} D_{n'}(\theta_{k'}) \rho_{k'}^{-\alpha} \right). \tag{C.4}$$

For $u_{k^{'},n^{'}} \in U^{'}$, beam $n^{'}$ is the best beam with the largest directivity of user $k^{'}$, i.e., $D_{n^{'}}(\theta_{k^{'}}) \geq D_n(\theta_{k^{'}}), \forall n \neq n^{'}, n = 1, 2, \cdots, N$. It is clear from Fig. 1(b) that the minimum value of $D_{n^{'}}(\theta_{k^{'}})$ is achieved at the crosspoint of two adjacent beams. Let l and $l+1$ denote the indices of any two adjacent beams, and ϕ denote the AoD measured at the crosspoint. Then we have

$$D_{n^{'}}(\theta_{k^{'}}) \geq D_l(\phi), \tag{C.5}$$

and

$$D_l(\phi) = D_{l+1}(\phi). \tag{C.6}$$

According to (1), (2), and (5), $D_l(\phi)$ can be obtained as

$$D_l(\phi) = \frac{\sin^2\left(0.5N\pi\cos\phi - \left(-\frac{N+1}{2} + l\right)\pi\right)}{N\sin^2\left(0.5\pi\cos\phi - \frac{1}{N}\left(-\frac{N+1}{2} + l\right)\pi\right)}. \tag{C.7}$$

It can be then obtained from (C.6) and (C.7) that

$$0.5\pi\cos\phi - \frac{1}{N}\left(-\frac{N+1}{2} + l\right)\pi + 0.5\pi\cos\phi - \frac{1}{N}\left(-\frac{N+1}{2} + l + 1\right)\pi = 0, \tag{C.8}$$

i.e.,

$$\cos\phi = \frac{2l - N}{N}. \tag{C.9}$$

By substituting (C.9) into (C.7), $D_l(\phi)$ is obtained as

$$D_l(\phi) = \frac{1}{N\sin^2\frac{\pi}{2N}}. \tag{C.10}$$

By combining (C.5) and (C.10), we further have

$$D_{n^{'}}(\theta_{k^{'}}) \geq \frac{1}{N\sin^2\frac{\pi}{2N}}. \tag{C.11}$$

With the distance from user $k^{'}$ to the cell center $\rho_{k^{'}} \leq 1$, by combining (C.1), (C.4) and (C.11), we have

$$\begin{aligned} R_s^{ap}(S) - R_s^{ap}(T) &< |S|\log_2\left(\frac{|T|}{|S|}\right) - (|T| - |S|)\log_2\left(\frac{K}{|T|}\left(\frac{K}{K-1}\right)^{K-1}\right) \\ &= (|T| - |S|)\log_2\left(\left(\frac{|T|}{|S|}\right)^{\frac{|S|}{|T|-|S|}} \cdot |T| \cdot \frac{1}{K}\left(\frac{K-1}{K}\right)^{K-1}\right). \end{aligned} \tag{C.12}$$

As $1 \leq |S| < |T| \leq K$, we can easily obtain that

$$\left(\frac{|T|}{|S|}\right)^{\frac{|S|}{|T|-|S|}} \cdot |T| \leq K \left(\frac{K}{K-1}\right)^{K-1}, \tag{C.13}$$

where "=" holds when $|S| = K-1$ and $|T| = K$. By substituting (C.13) into (C.12), it is clear that

$$R_s^{ap}(S) - R_s^{ap}(T) < 0. \tag{C.14}$$

According to the definition of a monotone set function presented in Section III-A, we can conclude that $R_s^{ap}(S)$, i.e., the objective function of (25a), is a monotone submodular function on the ground set U'. ■

Appendix D

Derivation of (29) and (30)

A. *Derivation of (29)*

For given total number of users K, the average service ratio $\bar{P}_s$ can be written as

$$\bar{P}_s = \sum_{m=1}^{K} \frac{m}{K} \Pr\{N_s = m\}. \tag{D.1}$$

To find the probability mass function of the number of allocated beams N_s, let us first consider the following ball-dropping problem: Assume that each ball is dropped into N distinct boxes with equal probability $1/N$. What is the probability that there are N_s non-empty boxes after dropping K balls? The ball-dropping problem shares similarities to our beam allocation problem because by regarding beams as boxes and users as balls, associating each user with a beam is equivalent to dropping each ball into a box. The probability that N_s beams are used is then the probability that N_s boxes are non-empty. Note that here an implicit assumption is that each user has equal probability $1/N$ to be associated with any beam, which is a good approximation only when the number of beams is large, i.e., each beam approximately has equal width.

For the ball-dropping problem, the total number of possible combinations by dropping K distinct balls into N distinct boxes is N^K. Note that a Stirling number of the second kind $\left\{ {K \atop N_s} \right\}$ denotes the number of ways to partition a set of K distinct objects into N_s non-empty subsets

[47]. Then the number of combinations that N_s boxes are non-empty is given by $\binom{N}{N_s} N_s! \left\{ {K \atop N_s} \right\}$. Therefore, the corresponding probability that there are N_s non-empty boxes after dropping K balls can be obtained as

$$\Pr\{N_s = m\} = \frac{\binom{N}{m} m! \left\{ {K \atop m} \right\}}{N^K}, \tag{D.2}$$

where the Stirling number of the second kind, $\left\{ {K \atop m} \right\}$, is given by [47]

$$\left\{ {K \atop m} \right\} = \frac{1}{m!} \sum_{j=1}^{m} (-1)^{m-j} \binom{m}{j} m^K. \tag{D.3}$$

(29) can be then obtained by combining (D.1)–(D.3).

B. Derivation of (30)

Note that a Stirling number of the second kind $\left\{ {K \atop m} \right\}$ obeys the recurrence relation [47]

$$\left\{ {K+1 \atop m} \right\} = m \left\{ {K \atop m} \right\} + \left\{ {K \atop m-1} \right\}. \tag{D.4}$$

Then we have

$$\left\{ {K \atop m} \right\} = \frac{1}{m} \left(\left\{ {K+1 \atop m} \right\} - \left\{ {K \atop m-1} \right\} \right). \tag{D.5}$$

By combining (D.1), (D.2) and (D.5), the average service ratio $\bar{P}_s$ can be obtained as

$$\begin{aligned} \bar{P}_s &= \frac{1}{KN^K} \left(\sum_{m=1}^{K} \binom{N}{m} m! \left\{ {K+1 \atop m} \right\} - \sum_{m=1}^{K} \binom{N}{m} m! \left\{ {K \atop m-1} \right\} \right) \\ &= \frac{1}{KN^K} \left(\sum_{m=1}^{K+1} \binom{N}{m} m! \left\{ {K+1 \atop m} \right\} - \binom{N}{K+1} (K+1)! - N \sum_{m=1}^{K} \binom{N-1}{m-1} (m-1)! \left\{ {K \atop m-1} \right\} \right) \\ &= \frac{1}{KN^K} \left[\sum_{m=1}^{K+1} \binom{N}{m} m! \left\{ {K+1 \atop m} \right\} - \binom{N}{K+1} (K+1)! - N \left(\sum_{i=1}^{K} \binom{N-1}{i} i! \left\{ {K \atop i} \right\} - \binom{N-1}{K} K! \right) \right] \\ &= \frac{1}{KN^K} \left[\sum_{m=1}^{K+1} \binom{N}{m} m! \left\{ {K+1 \atop m} \right\} - N \sum_{i=1}^{K} \binom{N-1}{i} i! \left\{ {K \atop i} \right\} \right]. \end{aligned} \tag{D.6}$$

Note that for the probability mass function $\Pr\{N_s = m\}$, we have

$$\sum_{m=1}^{K} \Pr\{N_s = m\} = 1. \tag{D.7}$$

By substituting (D.2) into (D.7), it is clear that

$$\sum_{m=1}^{K} \binom{N}{m} m! \left\{ {K \atop m} \right\} = N^K. \tag{D.8}$$

The average service ratio $\bar{P}_s$ can be then obtained by combining (D.6) and (D.8) as

$$\begin{aligned} \bar{P}_s &= \frac{1}{KN^K}\left(N^{K+1} - N(N-1)^K\right) \\ &= \frac{N}{K} - \frac{N}{K}\left(1 - \frac{1}{N}\right)^K \\ &\overset{\text{for large } N}{\approx} \frac{N}{K} - \frac{N}{K}e^{-\frac{K}{N}}. \end{aligned} \tag{D.9}$$

Acknowledgment

This work is currently supported by the UK Engineering and Physical Sciences Research Council, project EP/L026031/1, NIRVANA. It was also partly carried out within the framework of the EU-Japan Horizon 2020 Research and Innovation Programme, grant agreement no. 643297 (RAPID).

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[2] A. Lozano and A. M. Tulino, “Capacity of multiple-transmit multiple-receive antenna architectures,” *IEEE Trans. Inf. Theory*, vol. 48, no. 12, pp. 3117–3128, Dec. 2002.

[3] T. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Sep. 2010.

[4] *Special Issue on Large-scale Multiple Antenna Wireless Systems*, *IEEE J. Select. Areas Commun.*, vol. 31, no. 2, Feb. 2013.

[5] X. You, D. Wang, B. Sheng, X. Gao, X. Zhao, and M. Chen, “Cooperative distributed antenna systems for mobile communications,” *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 35–43, June 2010.

[6] L. Dai, “A comparative study on uplink sum capacity with co-located and distributed antennas,” *IEEE J. Select. Areas Commun.*, vol. 29, no. 6, pp. 1200–1213, June 2011.

[7] H. Zhu, “Performance comparison between distributed antenna and microcellular systems,” *IEEE J. Select. Areas Commun.*, vol. 29, no. 6, pp. 1151–1163, June 2011.

[8] J. Wang, H. Zhu, and N. J. Gomes, “Distributed antenna systems for mobile communications in high speed trains,” *IEEE J. Select. Areas Commun.*, vol. 30, no. 4, pp. 675–683, May 2012.

[9] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[10] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser MIMO systems,” *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[11] D. Ciuonzo, P. Salvo Rossi, and S. Dey, “Massive MIMO channel-aware decision fusion,” *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 604–619, Feb. 2015.

[12] A. Shirazinia, S. Dey, D. Ciuonzo, and P. Salvo Rossi, “Massive MIMO for decentralized estimation of a correlated source,” *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2499–2512, May 2016.

[13] J. Litva and T. K. Lo, *Digital Beamforming in Wireless Communications*, Artech House, 1996.

[14] T. Ohira, “Analog smart antennas: an overview,” in *Proc. IEEE PIMRC*, pp. 1502–1506, Sep. 2002.

[15] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[16] A. Alkhateeb, O. E. Ayach, G. Leus, and R. Heath, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE J. Select. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[17] L. Liang, W. Xu, and X. Dong, “Low-complexity hybrid precoding in massive multiuser MIMO systems,” *IEEE Wireless Commun. Letters*, vol. 3, no. 6, pp. 653–656, Dec. 2014.

[18] F. Sohrabi and W. Yu, “Hybrid digital and analog beamforming design for large-scale MIMO systems,” in *Proc. IEEE ICASSP*, pp. 2929–2933, Apr. 2015.

[19] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, “On the number of RF chains and phase shifters, and scheduling design with hybrid analog-digital beamforming,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2016.

[20] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, “A comparison of MIMO techniques in downlink millimeter wave cellular networks with hybrid beamforming,” *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1952–1967, May 2016.

[21] V. Venkateswaran and A. Veen, “Analog beamforming in MIMO communications with phase shift networks and online channel estimation,” *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4131–4143, Aug. 2010.

[22] O. N. Alrabadi, E. Tsakalaki, H. Huang, and G. F. Pedersen, “Beamforming via large and dense antenna arrays above a clutter,” *IEEE J. Select. Areas Commun.*, vol. 31, no. 2, pp. 314–325, Feb. 2013.

[23] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, “Millimeter wave beamforming for wireless backhaul and access in small cell networks,” *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[24] R. C. Hansen, *Phased Array Antennas*, John Wiley & Sons, Inc., 2009.

[25] J. Butler and R. Lowe, “Beam-forming matrix simplifies design of electrically scanned antennas,” *Electronic Design,* Apr. 1962.

[26] J. P. Choi and V. W. S. Chan, “Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2983–2993, Nov. 2005.

[27] A. Destounis and A. D. Panagopoulos, “Dynamic power allocation for broadband multi-beam satellite communication networks,” *IEEE Commun. Letters*, vol. 15, no. 4, pp. 380–382, Apr. 2011.

[28] N. K. Srivastava and A. K. Chaturvedi, “Flexible and dynamic power allocation in broadband multi-beam satellites,” *IEEE Commun. Letters,* vol. 17, no. 9, pp. 1722–1725, Sep. 2013.

[29] A. Osseiran and M. Ericson, “System performance of multi-beam antennas for HS-DSCH WCDMA system,” in *Proc. IEEE PIMRC,* pp. 2241–2245, Sep. 2004.

[30] B. Allen and M. Beach, “On the analysis of switched-beam antennas for the W-CDMA downlink,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 569–578, May 2004.

[31] J. Brady and A. Sayeed, “Beamspace MU-MIMO for high-density gigabit small cell access at millimeter-wave frequencies,” in *Proc. IEEE SPAWC*, June 2014.

[32] S. K. Yong, M. E. Sahin, and Y. H. Kim, “On the effects of misalignment and angular spread on the beamforming performance,” in *Proc. IEEE CCCN*, Jan. 2007.

[33] W. Choi, A. Forenza, J. G. Andrews, and R. W. Heath. Jr., “Opportunistic space-division multiple access with beam selection,” *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2371–2380, Dec. 2007.

[34] J. Choi, “Opportunistic beamforming with single beamforming matrix for virtual antenna array,” *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 872–881, Mar. 2011.

[35] M. Xia, Y. Wu, and S. Aissa, “Non-orthogonal opportunistic beamforming: performance analysis and implementation,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1424–1433, Apr. 2012.

[36] J. L. Vicario, R. Bosisio, C. Anton-Haro, and U. Spagnolini, “Beam selection strategies for orthogonal random beamforming in sparse networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3385–3396, Sep. 2008.

[37] S. Fujishige, *Submodular Functions and Optimization*, Elsevier, 2005.

[38] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter wave channel modeling and cellular capacity evaluation,” *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[39] F. Gross, *Smart Antennas for Wireless Communications*, McGraw-Hill, 2005.

[40] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, “Multiuser OFDM with adaptive subcarrier, bit and power allocation,” *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[41] C. Y. Ng and C. W. Sung, “Low complexity subcarrier and power allocation for utility maximization in uplink OFDMA systems,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1667–1675, May 2008.

[42] C.-N. Hsu, H.-J. Su, and P.-H. Lin, “Joint subcarrier pairing and power allocation for OFDM transmission with decode-and-forward relaying,” *IEEE Trans. signal Process.*, vol. 59, no. 1, pp. 399–414, Jan. 2011.

[43] D. Fooladivanda and C. Rosenberg, “Joint resource allocation and user association for heterogeneous wireless cellular networks,” *IEEE Trans. Wireless Commn.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

[44] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, “Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services,” *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2366–2376, July 2014.

[45] J. Lee, V. S. Mirrokni, V. Nagarajan, and M. Sviridenko, “Maximizing nonmonotone submodular functions under matroid or knapsack constraints,” *SIAM J. Discrete Math.*, 23(4), pp. 2053–2078, Jan. 2010.

[46] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, “An analysis of approximations for maximizing submodular set functions-II,” *Mathematical Programming Study*, vol. 8, pp. 73–87, 1978.

[47] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley Publishing Company, 1994.