A Simple and Effective Approach for Transmit Antenna Selection in Multiuser Massive MIMO Leveraging Submodularity

Aritra Konar, *Member, IEEE*, and Nicholas D. Sidiropoulos, *Fellow, IEEE*

***Abstract*—Massive MIMO systems are expected to enable great improvements in spectral and energy efficiency. Realizing these benefits in practice, however, is hindered by the cost and complexity of implementing large-scale antenna systems. A potential solution is to use transmit antenna selection for reducing the number of radio-frequency (RF) chains at the base station. In this paper, we consider the NP-hard discrete optimization problem of performing transmit antenna selection in the downlink of a single cell, multiuser massive MIMO system by maximizing the downlink sum-rate capacity with fixed user power allocation subject to various RF switching constraints. Whereas prior work has focused on using convex relaxation based schemes, which lack theoretical performance guarantees and can be computationally demanding, we adopt a very different approach. We establish that the objective function of this antenna selection problem is monotone and satisfies an important property known as *submodularity*, while the RF switching constraints are expressible as the independent sets of a matroid. This implies that a simple greedy algorithm can be used to guarantee a constant-factor approximation for *all* problem instances. Simulations indicate that greedy selection yields a near-optimal solution in practice and captures a significant fraction of the total downlink channel capacity at substantially lower complexity relative to convex relaxation based approaches, even with very few RF chains. This paves the way for substantial reduction in hardware complexity of massive MIMO systems while using very simple algorithms.**

***Index Terms*—Multi-user massive MIMO, transmit antenna selection, sum-rate capacity, discrete optimization, submodularity, matroids, greedy algorithm.**

Manuscript received October 13, 2017; revised April 28, 2018; accepted July 23, 2018. Date of publication August 6, 2018; date of current version August 20, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tsung-Hui Chang. This work is supported in part by National Science Foundation under Grant CIF-1525194, and in part by the University of Minnesota through a Doctoral Dissertation Fellowship. This paper was presented in part at the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, November 2017. *(Corresponding author: Nicholas D. Sidiropoulos.)*

A. Konar was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. He is now with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: aritra@virginia.edu).

N. D. Sidiropoulos is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: nikos@virginia.edu).

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

Digital Object Identifier 10.1109/TSP.2018.2863654

I. Introduction

MASSIVE MIMO [2] is currently a leading physical-layer technology candidate for implementation in future 5G wireless cellular systems [3], [4]. The basic idea of massive MIMO involves equipping a cellular base station (BS) with a large number of transmit antennas for the purpose of serving few users sharing the same time-frequency resource. Theoretical studies have demonstrated that the spatial multiplexing and array gains offered by such large-scale antenna systems can result in dramatic improvements in performance with regard to data rate, spectral and energy efficiency and link reliability, relative to conventional MIMO systems [2], [5]–[7]. Additionally, these findings have been verified via several experiments conducted using measured propagation channels. However, in order to successfully reap these benefits in practice, one has to take into account the cost and hardware complexity associated with implementing large-scale antenna systems.

In conventional MIMO systems with few transmit antennas, each antenna element is provided with a dedicated RF chain. While antenna elements are fairly inexpensive and small, RF chains are costly, bulky, and power-consuming. Hence, assigning a dedicated RF chain to every antenna element is practically infeasible in massive MIMO systems, due to the large number of BS antennas. Consequently, the design of transmit precoders which use a limited number of RF chains to strike a favorable balance between performance and hardware efficiency is a well motivated problem for massive MIMO.

One possible approach is transmit antenna selection, where a network of RF switches is used to connect the RF chains with a subset of antennas. Transmit antenna selection in conventional MIMO systems (both single and multi-user ones) has been previously considered in [8]–[15]. For point-to-point massive MIMO systems, [16] proposed a pair of heuristic antenna selection algorithms with the goal of improving energy efficiency. Meanwhile, [17] considered SNR maximization as the selection criterion and developed a polynomial-time algorithm for optimally solving the problem, irrespective of the number of transmit antennas, as long as the number of receive antennas is limited to two. In [18], Gao *et al.* considered the multi-user case with a single receive antenna per user, and formulated the problem as maximizing the downlink sum-rate capacity with predetermined user power allocation subject to a fully-flexible (FF) switching constraint on the total number of selected antennas.

The authors adopted a convex relaxation based approach which first requires solving a relaxed convex optimization problem followed by a post-processing rounding step to determine the set of antennas. Thereafter, a zero-forcing beamformer (ZFB) was designed using the selected antennas for evaluating the practical performance of the selection scheme. Experiments conducted using measured massive MIMO channels revealed that the proposed two-step approach is capable of exploiting the significant correlation that exists across the power profiles of the antenna elements to attain a considerable fraction of the data rate achievable using all antennas. Similar results were reported in [19], [20], where the approach was extended to the case of partially connected (PC) switching constraints. Hence, antenna selection has much potential as a means of achieving a favorable performance-complexity trade-off in massive MIMO systems.

In this paper, we adopt, at a high level, the same two-step approach as [18]–[20]; i.e., we first consider the problem of antenna subset selection which maximizes the downlink sum-rate capacity of a multi-user massive MIMO system with fixed user power allocation subject to various switching constraints imposed on the network of RF switches, followed by designing a ZFB for the selected antennas. The major difference of our present work with the aforementioned references lies in the crucial first step: we refrain from employing a relaxation based selection algorithm. As mentioned previously, this approach has been empirically observed to work well in certain settings. However, solving a general convex programming problem for massive MIMO systems can be computationally demanding, even practically intractable. Furthermore, *apriori* one cannot theoretically guarantee that this heuristic convex relaxation approach will work well for an arbitrary problem instance. Since the antenna selection problem under consideration is known to be NP–hard in its general form, a natural question arises as to whether there even exists a polynomial-time algorithm which is both computationally efficient and theoretically sound for *all* instances.

Our key contribution is that we provide an affirmative answer to the above question. To be specific, we first establish that all instances of antenna selection based on the downlink capacity maximization criterion with fixed user power allocation can be equivalently expressed in the form of the maximum entropy sampling problem [21], whose objective function is monotone and additionally satisfies an important property known as *submodularity* [22]. Such a class of discrete set-functions is particularly notable for exhibiting a natural diminishing returns property. We also establish that a variety of RF switching constraints arising in both FF and PC RF switching networks can be explicitly expressed as the independent sets of a matroid [23], which are a very convenient and powerful mathematical abstraction for describing combinatorial structures. As a result, the antenna selection problem can be viewed as maximizing a monotone submodular function over the independent sets of a matroid. The upshot is that we are able to use a simple greedy algorithm to obtain solutions for antenna selection with minimal computational effort while also guaranteeing constant factor approximation for all instances by invoking a series of celebrated results [24], [25] regarding constrained submodular maximization. For the case of FF RF switching, the greedy algorithm guarantees a worst-case approximation factor of $(1 - 1/e)$,[1] which is furthermore un-improvable in polynomial-time; meanwhile, we obtain a $1/2$-factor approximation guarantee for PC RF switching. These results are independent of almost all choices of system parameters in massive MIMO systems ranging from the number of transmit antennas, users, receive antennas, RF chains, sub-carriers, type of channel model, and RF switching architecture employed, thereby making the greedy algorithm an extremely potent tool. We furthermore establish an equivalence between transmit antenna selection for maximizing downlink capacity in a MIMO broadcast channel (BC) and receive antenna selection for a design of experiments problem in the dual uplink MIMO multiple access channel (MAC), thereby revealing an interesting link between the two problems.

Finally, we point out that we are, to the best of our knowledge, the first to exploit the submodularity of the capacity based criterion for transmit antenna selection in multi-user massive MIMO systems to justify the use of the greedy algorithm as a principled approximation algorithm rather than a heuristic. That being said, such an approach has been previously considered in the context of *receive* antenna selection with FF switching for maximizing Shannon capacity in point-to-point MIMO systems [27] in order to provide theoretical performance guarantees for the greedy algorithm of [26]. However, the scenario the authors of [27] considered requires having more receive antennas than transmit antennas, which is not aligned with massive MIMO systems. Furthermore, in hindsight, while the submodularity of the objective function is readily apparent in this single user setting, this is not the case for our multi-user scenario, as it is necessary to utilize the duality of the capacity regions of the downlink MIMO BC and uplink MIMO MAC [28]–[30] followed by applying suitable algebraic manipulations to arrive at a formulation with a similar form.[2] Additionally, by utilizing the abstraction of matroids, we are able to describe a far broader class of switching architectures compared to the case of FF switching considered in [27]. This in turn, allows us to demonstrate that the greedy algorithm can yield provable approximation guarantees for this larger class of selection problems as well.

The efficacy of the greedy algorithm is experimentally evaluated in a variety of massive MIMO settings under different RF switching constraints. We use the convex relaxation approach as a performance benchmark and utilize a simple first-order method for improving its running time. Our experiments indicate that the worst-case performance guarantees for the greedy algorithm are very pessimistic; in practice near-optimal performance is obtained. Furthermore, the subset of antennas selected by the greedy algorithm can successfully capture a significant fraction of the total downlink sum-rate capacity using very few RF chains, even when selection is performed across multiple sub-carriers under restrictive switching constraints. Relative to the relaxation approach, the greedy algorithm also exhibits a far

[1] Here, e denotes Euler's number.

[2] In fact, after reformulation, the resulting problem (17) with FF switching in this paper and problem (2) of [27] are equivalent. Hence, in this case, the greedy algorithm yields the same performance guarantees for both problems.

superior performance-complexity trade-off, with running times comparable to the coherence interval of massive MIMO channels. Additionally, when a ZFB is designed for the reduced MIMO BC using the subset of antennas determined by the greedy algorithm, we observe that the resulting performance loss in terms of sum-rate is often minimal. Owing to its potent combination of simplicity, theoretical performance guarantees and impressive practical performance, we can reasonably claim that the proposed two-step, greedy antenna selection followed by ZFB approach now constitutes the state-of-the-art for transmit antenna selection in multi-user massive MIMO systems and paves the way for significant reduction in system complexity far beyond what was previously achievable by employing the approaches used in [18]–[20].

Relative to the conference version [1], which considered the special case of antenna selection in multi-user massive MIMO systems with a single receive antenna per user and FF RF switching constraints, the present journal version extends the results to the most general case with multiple receive antennas per user, PC RF switching, and multiple sub-carriers.

The rest of the paper is organized as follows. Section II outlines the system model while Section III describes the problem formulation. A brief overview of submodular functions and matroids is provided in Section IV, followed by establishing the submodularity of the antenna selection criterion in Section V. Section VI describes the greedy algorithm and its computational aspects. The case of antenna selection across multiple sub-carriers is discussed in Section VII, while in Section VIII we demonstrate how to improve the scalability of the convex relaxation based approach. A set of comprehensive experiments are carried out in Section IX, and conclusions are drawn in Section X.

Throughout the paper, we adopt the following notation. Capital boldface is reserved for matrices, vectors are denoted by small boldface, and scalars are represented in the normal face. We denote the $N-$dimensional complex Euclidean space by $\mathbb{C}^N$, while the $N \times N$ identity matrix is represented as $\mathbf{I}_N$. Superscript ^{T} is used to denote the transpose of a vector/matrix, while ^{H} denotes Hermitian transpose. Calligraphic font is used to denote finite sets. The notation $[n]$ is used as shorthand for the set $\{1, \ldots, n\}$, while the set of natural numbers is indicated by $\mathbb{N}$. For a differentiable function $f(.)$, its gradient is denoted by $\nabla f(.)$. Finally, the expectation operator associated with a random vector/variable is denoted as $\mathbb{E}\{.\}$.

II. System Model

Consider a downlink transmission scenario in a single-cell, multi-user (MU)-MIMO system where a BS equipped with M_T transmit antennas simultaneously serves K users, each of whom is equipped with $M_R \geq 1$ receive antennas. The massive MIMO paradigm envisions a setting where a small number of users, each with a few receive antennas, are served by a BS equipped with a very large antenna array; i.e., we have $M_T >> KM_R$. We make the assumption that the BS has already acquired perfect channel state information (CSI) for all users. Let N denote the number of available RF chains at the BS with $KM_R \leq N \leq M_T$. For a given subset $\mathcal{S} \subseteq [M_T]$ of $|\mathcal{S}| = N$ simultaneously active antennas, the received signal $\mathbf{y}_k \in \mathbb{C}^{M_R}$ of user k can be expressed as

$$\mathbf{y}_k = \sqrt{\rho}\mathbf{H}_k^{[\mathcal{S}]}\mathbf{x} + \mathbf{n}_k, \forall\, k \in [K] \tag{1}$$

where $\rho > 0$ is the transmit power budget, and $\mathbf{H}_k^{[\mathcal{S}]} \in \mathbb{C}^{M_R \times N}$ is the MIMO channel sub-matrix obtained by selecting a subset of N columns (indexed by $\mathcal{S}$) from the k^{th} user's full channel matrix $\mathbf{H}_k \in \mathbb{C}^{M_R \times M_T}$, whose entries correspond to quasi-static, flat-fading channel coefficients. Additionally, we let $\mathbf{H} := [\mathbf{H}_1^H, \ldots, \mathbf{H}_K^H]^H$ denote the concatenated channel matrix and let $\mathbf{H}^{[\mathcal{S}]} \in \mathbb{C}^{KM_R \times N}$ represent its sub-matrix with columns indexed by $\mathcal{S}$. Finally, $\mathbf{x} \in \mathbb{C}^N$ is the transmit signal vector applied across the N selected antennas satisfying the normalized transmit power constraint $\mathbb{E}\{\|\mathbf{x}\|_2^2\} \leq 1$, while $\{\mathbf{n}_k\}_{k=1}^K$ are complex, circularly symmetric, i.i.d. Gaussian noise vectors at the receivers, with $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M_R})$.

III. Problem Formulation

In the absence of antenna selection, the sum-rate capacity of the downlink MIMO BC with a normalized transmit power constraint is equal to the capacity of the dual uplink MIMO (MAC subject to a unit transmit sum-power constraint [28]–[30] and is given by

$$\begin{aligned} C(\mathbf{H}) = \max_{\{\mathbf{Q}_k\}_{k=1}^K} \ & \log_2 \det\left(\mathbf{I}_{M_T} + \rho\sum_{k=1}^K \mathbf{H}_k^H\mathbf{Q}_k\mathbf{H}_k\right) \\ \text{s.t.} \ & \mathbf{Q}_k \succeq \mathbf{0}, \forall\, k \in [K] \\ & \sum_{k=1}^K \text{Trace}(\mathbf{Q}_k) \leq 1 \end{aligned} \tag{2}$$

where $\{\mathbf{Q}_k\}_{k=1}^K \in \mathbb{C}^{M_R \times M_R}$ denote uplink covariance matrices to be determined. Note that (2) is a convex optimization problem, which can be optimally solved, for example, using the sum-power iterative waterfilling algorithm proposed in [31]. Thereafter, the optimal downlink covariance matrices can be determined from $\{\mathbf{Q}_k\}_{k=1}^K$ via the MAC to BC transformations described in [30, Equations (8)–(10)]. The non-linear dirty-paper coding (DPC) [32] technique is known to be the optimal capacity attaining transmission strategy in this case.

When only a subset $\mathcal{S}$ of $N < M_T$ transmit antennas are active at the BS, the capacity expression becomes a function of the selected antennas and is given by

$$\begin{aligned} C(\mathbf{H}^{[\mathcal{S}]}) = \max_{\{\mathbf{Q}_k\}_{k=1}^K} \ & \log_2 \det\left(\mathbf{I}_N + \rho\sum_{k=1}^K (\mathbf{H}_k^{[\mathcal{S}]})^H\mathbf{Q}_k\mathbf{H}_k^{[\mathcal{S}]}\right) \\ \text{s.t.} \ & \mathbf{Q}_k \succeq \mathbf{0}, \forall\, k \in [K] \\ & \sum_{k=1}^K \text{Trace}(\mathbf{Q}_k) \leq 1 \end{aligned} \tag{3}$$

Determining the capacity with antenna selection is tantamount to solving the following optimization problem

$$C(\mathbf{H}) = \max_{\mathcal{S},\{\mathbf{Q}_k\}_{k=1}^K} \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^{K} (\mathbf{H}_k^{[\mathcal{S}]})^H \mathbf{Q}_k \mathbf{H}_k^{[\mathcal{S}]}\right)$$
$$\text{s.t. } \mathbf{Q}_k \succeq \mathbf{0}, \forall\, k \in [K]$$
$$\sum_{k=1}^{K} \text{Trace}(\mathbf{Q}_k) \leq 1,\ \mathcal{S} \in \mathcal{I} \quad (4)$$

where $\mathcal{I} \subseteq 2^{[M_T]}$ denotes the family of antenna subsets which can be selected, and is determined by the type of RF switching architecture employed for antenna selection. For example, when a FF RF switching matrix is utilized (i.e., each of the N RF switches can connect to any one of the M_T antennas), the feasible set is described by the collection of subsets $\mathcal{I} = \{\mathcal{S} \subseteq [M_T] : |\mathcal{S}| = N\}$. Alternatively, when a PC RF switching matrix is used, the restricted connectivity partitions the antenna array into a number of disjoint sub-arrays, with each sub-array being served by a block of switches. To formalize this notion, let us consider a PC switching architecture with N RF switches organized into B distinct switching blocks (with $1 \leq B \leq N$). Let N_b denote the number of switches assigned to switching block $b \in [B]$ such that $\sum_{b\in[B]} N_b = N$. Each switching block provides connections to a sub-array of antenna elements, which partitions the array as $[M_T] = \cup_{b\in[B]} \mathcal{M}_b$, where $\mathcal{M}_b$ denotes the antenna sub-array associated with the b^{th} switching block, $|\mathcal{M}_b| \geq N_b$, $\sum_{b\in[B]} |\mathcal{M}_b| = M_T$ and $\mathcal{M}_{b_1} \cap \mathcal{M}_{b_2} = \emptyset, \forall\, b_1 \neq b_2, b_1, b_2 \in [B]$. With N_b switches, each block b is required to select a subset of antennas $\mathcal{S}_b \subseteq \mathcal{M}_b$ from each sub-array such that $|\mathcal{S}_b| = N_b$. The overall selection set $\mathcal{S} \subseteq [M_T]$ obtained over B switching blocks can then be defined as $\mathcal{S} := \cup_{b\in[B]} \mathcal{S}_b$. Since the sub-arrays $\{\mathcal{M}_b\}_{b=1}^B$ are disjoint, it follows that the selected sets $\{\mathcal{S}_b\}_{b=1}^B$ are disjoint as well. Consequently, we can equivalently express $\mathcal{S}_b = \mathcal{S} \cap \mathcal{M}_b, \forall\, b \in [B]$. The feasible set in this case can be described by the family of subsets $\mathcal{I} = \{\mathcal{S} \subseteq [M_T] : |\mathcal{S} \cap \mathcal{M}_b| = N_b, \forall\, b \in [B]\}$. Note that the number of switching blocks, B, the choice of array partition $\{\mathcal{M}_b\}_{b=1}^B$ and the number of switches N_b assigned to each sub-array can be arbitrary. For a concrete example, consider the case $M_T = 2N$, $M_R = 1$, $B = N$, and $|\mathcal{M}_b| = M_T/B = 2, N_b = N/B = 1, \forall\, b \in [B]$. This reduces to the special case of binary switching described in [19], where the entire array is partitioned into N sub-arrays comprising 2 antennas each, and each of the N switches is connected to only one of 2 antennas in each sub-array.

In contrast to (2), the additional antenna selection constraints make (4) a challenging, mixed-integer optimization problem since we must jointly design the covariance matrices $\{\mathbf{Q}_k\}_{k=1}^K$ for all users *and* select the optimal subset of antennas, which requires enumeration over a combinatorial number of subsets of $[M_T]$ in the worst case. As proposed in [18] for the special case $M_R = 1$, problem (4) can be partially simplified by opting to use a predetermined set of covariance matrices $\{\mathbf{Q}_k\}_{k=1}^K$ satisfying the transmit sum-power constraint $\sum_{k=1}^K \text{Trace}(\mathbf{Q}_k) \leq 1$ and then performing antenna selection. The resulting optimization problem can be expressed as

$$C(\mathbf{H}, \{\mathbf{Q}_k\}_{k=1}^K) = \max_{\mathcal{S}\in\mathcal{I}} \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^{K} (\mathbf{H}_k^{[\mathcal{S}]})^H \mathbf{Q}_k \mathbf{H}_k^{[\mathcal{S}]}\right) \quad (5)$$

One possible choice of uplink covariance matrices is to use the optimal solution of (2). In the high SNR regime, we can alternatively use the set of uniform power allocation matrices $\mathbf{Q}_k = \frac{1}{KM_R}\mathbf{I}_{M_R}, \forall\, k \in [K]$, as it corresponds to the asymptotically optimal choice (i.e., achieves sum-rate capacity) under the conditions $M_T \geq KM_R$ and the concatenated channel matrix $\mathbf{H} := [\mathbf{H}_1^H, \ldots, \mathbf{H}_K^H]^H$ being full row rank [28, Theorem 3]. In this case, (5) corresponds to selecting the subset of antennas which maximizes the fraction of the total downlink sum-rate capacity attainable with the full set of active antennas.

Unfortunately, even this "simplification" step does not make (5) any easier to solve. Note that for a given set of matrices $\{\mathbf{H}_k, \mathbf{Q}_k\}_{k=1}^K$, computing the optimal antenna subset

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\in\mathcal{I}} \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^{K} (\mathbf{H}_k^{[\mathcal{S}]})^H \mathbf{Q}_k \mathbf{H}_k^{[\mathcal{S}]}\right) \quad (6)$$

requires solving a subset selection problem. A simple combinatorial counting argument indicates that for the case of FF switching, a total of $|\mathcal{I}| = \binom{M_T}{N}$ enumerations may possibly be required, while the maximum number of such potential enumerations in the PC switching case is $|\mathcal{I}| = \prod_{b\in[B]} \binom{|\mathcal{M}_b|}{N_b}$. These intuitive arguments indicate that (5) is potentially NP–hard in the worst case. As formally established in [33], this is indeed the case.

From the previous discussion, it is clear that enumeration over all possible feasible subsets becomes intractable even for modest values of M_T and N, let alone for massive MIMO where hundreds of antennas can be potentially deployed at the BS. In order to deal with the combinatorial nature of the constraints, prior work [18]–[20] has relied upon using convex relaxations of the feasible set, which results in a convex programming problem that is optimally solvable in polynomial-time [34]. Since the solution of the relaxed problem is not guaranteed to be feasible for (5) in general, a post-processing fractional rounding step is performed to obtain a sub-optimal solution. While this approach has been empirically observed to work well in certain settings, overall, it suffers from the following drawbacks in general

1) Using general-purpose convex programming solvers to solve the relaxed problem incurs worst-case complexity $O(M_T^{3.5})$. While this is a polynomial-time result in M_T, for massive MIMO, where M_T could potentially be in the order of hundreds, it is clear that this approach can be very computationally very expensive. Ideally, we desire an algorithmic approach whose complexity scales *linearly* in M_T.
2) More importantly, the sub-optimality of the solution obtained by convex relaxation followed by rounding, relative to the globally optimal value of (5) (which is NP–hard to determine), cannot be theoretically quantified. Hence, it is not possible to guarantee that this approach will yield

a high quality sub-optimal solution of (5) for *all* possible combinations of choices of matrices $\{\mathbf{H}_k, \mathbf{Q}_k\}_{k=1}^K$, number of transmit and receive antennas M_T and M_R respectively, RF chains N and switching architectures $\mathcal{I}$.

In the forthcoming sections, we demonstrate that by exploiting the fact that (6) corresponds to a monotone submodular maximization problem subject to matroid constraints, we can use a simple greedy algorithm to obtain very powerful performance guarantees for all instances of this NP–hard problem. First, we provide a brief primer on submodular functions and matroids.

IV. Overview of Submodularity and Matroids

Given a ground set of n objects $\mathcal{V} := \{v_1, \ldots, v_n\}$, consider the set-function $f : 2^{\mathcal{V}} \to \mathbb{R}$ which assigns a real value to any subset $\mathcal{S} \subseteq \mathcal{V}$.

Definition 1 (Submodularity): [22, p. 22] The set-function f is said to be *submodular* if and only if, for all subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$, it holds that

$$f(\mathcal{A}) + f(\mathcal{B}) \geq f(\mathcal{A} \cup \mathcal{B}) + f(\mathcal{A} \cap \mathcal{B}) \tag{7}$$

If we define $\Delta_f(v|\mathcal{A}) := f(\mathcal{A} \cup \{v\}) - f(\mathcal{A})$ as the marginal gain of the object $v \in \mathcal{V}$ with respect to (w.r.t.) $\mathcal{A} \subseteq \mathcal{V}$, then the above definition can be equivalently restated as:

Definition 2: If f is submodular, then, for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \setminus v$, it holds that

$$\Delta_f(v|\mathcal{A}) \geq \Delta_f(v|\mathcal{B}) \tag{8}$$

That is, given a subset of objects $\mathcal{A}$, the marginal gain derived by adding v to $\mathcal{A}$ does not increase when we add v to the superset $\mathcal{B}$. Hence, submodular functions exhibit a natural diminishing returns property. If f satisfies (7)–(8) with equality, it is said to be *modular*.

Definition 3 (Monotonicity): A set-function f is said to be *monotone* if $f(\mathcal{A}) \leq f(\mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$. Moreover, if f is also submodular, then monotonicity is equivalent to the condition $\Delta_f(v|\mathcal{V} \setminus v) \geq 0, \forall\, v \in \mathcal{V}$.

Additionally, every monotone submodular function can be characterized by a quantity known as *curvature*, which measures how “close” a submodular function is to being modular.

Definition 4 (Curvature): [35] The curvature $c_f \in [0, 1]$ of a monotone submodular function f is defined as

$$c_f = 1 - \min_{v \in \mathcal{V}} \frac{\Delta_f(v|\mathcal{V} \setminus v)}{\Delta_f(v|\emptyset)} \tag{9}$$

For a modular function, we have $c_f = 0$, whereas if $\Delta_f(v|\mathcal{V} \setminus v) = 0, \forall\, v \in \mathcal{V}$, then $c_f = 1$. The role played by curvature will become apparent later on when we discuss algorithmic techniques for obtaining approximate solutions for problems involving maximization of monotone submodular functions.

Finally, we point out that maximizing a monotone submodular function $f(\mathcal{A})$ (where $\mathcal{A} \subseteq \mathcal{V}$) *without* any constraints is a trivial problem whose optimal solution is the ground set $\mathcal{V}$. To make the problem interesting, maximization is usually performed subject to some constraints on $\mathcal{A}$, which can often be described via a matroid.

Definition 5 (Matroid): [23] A *matroid* is an ordered pair $(\mathcal{V}, \mathcal{I})$ consisting of a finite ground set $\mathcal{V}$ and a collection of subsets $\mathcal{I}$ of $\mathcal{V}$ called independent sets which satisfy the following axioms

A1) $\emptyset \in \mathcal{I}$

A2) If $\mathcal{B} \in \mathcal{I}$ and $\mathcal{A} \subseteq \mathcal{B}$, then $\mathcal{A} \in \mathcal{I}$

A3) If $\mathcal{A}, \mathcal{B} \in \mathcal{I}$ and $|\mathcal{A}| < |\mathcal{B}|$, then $\exists\, v \in \mathcal{B} \setminus \mathcal{A}$ such that $\mathcal{A} \cup \{v\} \in \mathcal{I}$

A matroid can be regarded as a generalization of the classical notion of linear independence established in linear algebra. We now provide two examples of matroids, which will be useful later on.

Example 1 (Uniform Matroid): A matroid with independent sets $\mathcal{I} = \{\mathcal{A} \subseteq \mathcal{V} : |\mathcal{A}| \leq k\}$ is called a uniform matroid and satisfies axioms (A1)–(A3).

Example 2 (Partition Matroid): Consider the partition $\mathcal{V} = \cup_{i=1}^m \mathcal{G}_i$ where $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset, \forall\, i \neq j, i, j \in [m]$. A matroid with independent sets $\mathcal{I} = \{\mathcal{A} \subseteq \mathcal{V} : |\mathcal{A} \cap \mathcal{G}_i| \leq k_i, \forall\, i \in [m]\}$ satisfies axioms (A1)–(A3).

V. Antenna Selection as Monotone Submodular Maximization Over Matroids

In this section, we show that the objective function of (6) is monotone submodular. First, for a given set of covariance matrices $\{\mathbf{Q}_k\}_{k=1}^K$ which satisfy the transmit sum-power constraint, we define the block diagonal matrix $\mathbf{Q} = \mathrm{blkdiag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_K) \in \mathbb{C}^{KM_R \times KM_R}$ and denote its matrix square-root factor as $\mathbf{Q}^{1/2}$. Given the concatenated channel matrix $\mathbf{H} \in \mathbb{C}^{KM_R \times M_T}$, let $\mathbf{H}^{[\mathcal{S}]} \in \mathbb{C}^{KM_R \times |\mathcal{S}|}$ represent the sub-matrix obtained by selecting the column subset $\mathcal{S} \subseteq [M_T]$. Next, we define the modified channel matrix

$$\tilde{\mathbf{H}} := \mathbf{Q}^{1/2}\mathbf{H} \tag{10}$$

from which it follows that we also have

$$\tilde{\mathbf{H}}^{[\mathcal{S}]} = \mathbf{Q}^{1/2}\mathbf{H}^{[\mathcal{S}]}, \forall\, \mathcal{S} \subseteq [M_T] \tag{11}$$

We also define the positive semidefinite Grammian matrix $\tilde{\mathbf{G}} := \tilde{\mathbf{H}}^H\tilde{\mathbf{H}}$ and let $\tilde{\mathbf{G}}[\mathcal{S}, \mathcal{S}]$ denote the $\mathbb{C}^{|\mathcal{S}| \times |\mathcal{S}|}$ sub-matrix of $\tilde{\mathbf{G}}$ with row and column indices in $\mathcal{S} \subseteq [M_T]$. The objective function of the antenna selection problem (6) can then be equivalently expressed as

$$f(\mathcal{S}) = \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^K (\mathbf{H}_k^{[\mathcal{S}]})^H \mathbf{Q}_k \mathbf{H}_k^{[\mathcal{S}]}\right) \tag{12a}$$

$$= \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho(\mathbf{H}^{[\mathcal{S}]})^H \mathbf{Q}\mathbf{H}^{[\mathcal{S}]}\right) \tag{12b}$$

$$= \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho(\tilde{\mathbf{H}}^{[\mathcal{S}]})^H \tilde{\mathbf{H}}^{[\mathcal{S}]}\right) \tag{12c}$$

$$= \log_2 \det\left(\mathbf{I}_{M_T}[\mathcal{S}, \mathcal{S}] + \rho\tilde{\mathbf{G}}[\mathcal{S}, \mathcal{S}]\right) \tag{12d}$$

where in the last step we have utilized the fact that we can write, without loss of generality (w.l.o.g.), $\mathbf{I}_{|\mathcal{S}|} = \mathbf{I}_{M_T}[\mathcal{S}, \mathcal{S}]$ for all subsets $\mathcal{S} \subseteq [M_T]$.

Proposition 1: The function $f(\mathcal{S})$ is submodular.

Proof: Note that $f(\mathcal{S})$ is the log determinant of the $\mathbb{C}^{|\mathcal{S}|\times|\mathcal{S}|}$ sub-matrix of the positive definite matrix $\mathbf{\Sigma} := \mathbf{I}_{M_T} + \rho\tilde{\mathbf{G}}$, which is known to be a submodular function [36]–[38]. The result essentially follows from the fact that differential entropy [39] as a function of a subset of random variables is submodular. Applying this fact to the special case of Gaussian random variables with distribution $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ yields the desired claim. ■

Proposition 2: The function $f(\mathcal{S})$ is monotone.

Proof: We are required to show

$$\begin{aligned}\Delta_f(v|\mathcal{V}\setminus v) &= f(\mathcal{V}) - f(\mathcal{V}\setminus v)\\ &= \log_2\det(\mathbf{\Sigma}) - \log_2\det(\mathbf{\Sigma}[\mathcal{V}\setminus v, \mathcal{V}\setminus v])\\ &\geq 0, \forall\, v \in \mathcal{V}\end{aligned}$$

As mentioned in [40], this holds as a simple consequence of Cauchy's theorem of interlacing eigen-values [41, Theorem 4.3.17]. ■

This concludes the proof of $f(\mathcal{S})$ being monotone submodular. We now turn our attention to the antenna selection constraints. Simple inspection reveals that in the FF switching case, the feasible subsets $\mathcal{I} = \{\mathcal{S} \subseteq [M_T] : |\mathcal{S}| \leq N\}$ correspond to a uniform matroid $([M_T], \mathcal{I})$, whereas in the PC case, the subsets $\mathcal{I} = \{\mathcal{S} \subseteq [M_T] : |\mathcal{S} \cap \mathcal{M}_b| \leq N_b, \forall\, b \in [B]\}$ define a partition matroid $([M_T], \mathcal{I})$.[3] Hence, the transmit antenna selection problem

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\in\mathcal{I}} f(\mathcal{S}) \tag{13}$$

corresponds to maximizing a monotone submodular function subject to matroid constraints.

An Alternative Interpretation via Duality

We now provide an alternative interpretation of the transmit antenna selection problem (13) in the downlink MIMO BC as a design of experiments problem in the dual uplink MIMO MAC. For simplicity of exposition, we consider the case of $\rho = 1$. In the uplink MAC, the received signal $\mathbf{v} \in \mathbb{C}^{M_T}$ at the BS is given by

$$\mathbf{v} = \sum_{k=1}^{K} \mathbf{H}_k^H \mathbf{u}_k + \mathbf{w} = \mathbf{H}^H \mathbf{u} + \mathbf{w} \tag{14}$$

where $\mathbf{u} = [\mathbf{u}_1^H, \ldots, \mathbf{u}_K^H]^H \in \mathbb{C}^{KM_R}$ is the vector of signals transmitted by the users satisfying $\sum_{k=1}^K \mathbb{E}[\|\mathbf{u}_k\|_2^2] \leq 1$, while $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M_T})$ is complex, circularly symmetric, Gaussian noise at the BS.

Assuming that the distribution of each transmitted vector is $\mathbf{u}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_k), \forall\, k \in [K]$ such that $\sum_{k=1}^K \mathrm{Trace}(\mathbf{Q}_k) \leq 1$, and the set of vectors $\{\mathbf{u}_k\}_{k=1}^K$ are mutually independent, it follows that $\mathbf{u} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q})$. We now consider the problem of estimating $\mathbf{u}$ from the vector of noisy, linear measurements $\mathbf{v}$ available at the BS. The maximum aposteriori (MAP) estimator of the transmitted signal $\mathbf{u}$ is given by

$$\hat{\mathbf{u}} = \left(\mathbf{Q}^{-1} + \mathbf{H}\mathbf{H}^H\right)^{-1} \mathbf{H}\mathbf{v} \tag{15}$$

If we are restricted to only choose a subset $\mathcal{S} \in \mathcal{I} \subseteq 2^{M_T}$ of measurements from $\mathbf{v}$ (i.e., receive antennas at the BS), where $([M_T], \mathcal{I})$ forms a matroid system, the MAP estimator as a function of $\mathcal{S}$ can be expressed as

$$\hat{\mathbf{u}}(\mathcal{S}) = \left(\mathbf{Q}^{-1} + \mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H\right)^{-1} \mathbf{H}^{[\mathcal{S}]}\mathbf{v}^{[\mathcal{S}]} \tag{16}$$

with estimation error covariance matrix

$$\mathbf{E}(\mathcal{S}) = \left(\mathbf{Q}^{-1} + \mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H\right)^{-1} \tag{17}$$

One approach for selecting a subset of measurements which yields a high quality estimator is to minimize the volume of the error covariance matrix $\mathbf{E}(\mathcal{S})$ subject to the matroid constraints $\mathcal{S} \in \mathcal{I}$ imposed on the receive antennas at the BS. This can be expressed in the form of the following problem

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\in\mathcal{I}} \log_2\det\left(\mathbf{Q}^{-1} + \mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H\right) \tag{18}$$

It has been established [33] that the objective function of (18) is monotone submodular. Additionally, problem (18) is actually (13) in disguise. This can be seen via the following series of equivalent transformations

$$\max_{\mathcal{S}\in\mathcal{I}} \log_2\det\left(\mathbf{Q}^{-1} + \mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H\right) \tag{19a}$$

$$\Leftrightarrow \max_{\mathcal{S}\in\mathcal{I}} \log_2\det\left(\mathbf{Q}^{-1}(\mathbf{I}_{KM_R} + \mathbf{Q}\mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H)\right) \tag{19b}$$

$$\Leftrightarrow \max_{\mathcal{S}\in\mathcal{I}} \log_2\det\left(\mathbf{I}_{KM_R} + \mathbf{Q}\mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H\right) \tag{19c}$$

$$\Leftrightarrow \max_{\mathcal{S}\in\mathcal{I}} \log_2\det\left(\mathbf{I}_{M_T} + (\mathbf{H}^{[\mathcal{S}]})^H\mathbf{Q}\mathbf{H}^{[\mathcal{S}]}\right) \tag{19d}$$

where in the last step we have made use of Sylvester's Determinant theorem [41], which states that

$$\det(\mathbf{I} + \mathbf{A}\mathbf{B}) = \det(\mathbf{I} + \mathbf{B}\mathbf{A}) \tag{20}$$

Hence, the experiment design problem (18) of choosing the optimal subset of receive antennas for minimizing the volume of the error covariance matrix of the MAP estimator of the transmitted signal in the uplink MAC is equivalent to the transmit antenna selection problem (13) of maximizing the sum-rate capacity in the downlink BC.

VI. Greed is Good

We now describe a simple greedy algorithm for obtaining high quality approximate solutions of (13) for both FF and PC switching. In the case of FF switching (i.e., submodular maximization over a uniform matroid), the algorithm starts with the empty set $\mathcal{S}_0 = \emptyset$ and incrementally constructs a solution in the following fashion. At iteration $i \in [N]$, the element $v \notin \mathcal{S}_{i-1}$ maximizing

[3]One may naturally ask why the inequalities in both definitions of $\mathcal{I}$ have been replaced with equalities. In this particular case, the equalities can be relaxed to inequalities w.l.o.g. since $f(\mathcal{S})$ is a monotone function.

the marginal gain w.r.t. the set $\mathcal{S}_{i-1}$ is added according to the following update rule (with ties broken arbitrarily).

$$\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{\operatorname*{argmax}_{v \notin \mathcal{S}_{i-1}} \Delta_f(v|\mathcal{S}_{i-1})\}, \forall\, i \in [N] \tag{21}$$

Note that the greedy algorithm requires $O(NM_T)$ evaluations of the objective function $f(.)$. Regarding the quality of the solution set $\mathcal{S}_{\text{gr}}$ determined by the greedy algorithm, Nemhauser *et al.* [24] proved the following celebrated result.

Theorem 1: [24] Maximizing a monotone submodular function over a uniform matroid via the greedy algorithm guarantees a $(1 - 1/e)$-factor approximation for all instances (13). That is, if $\mathcal{S}^*$ denotes the optimal set, then $f(\mathcal{S}^*) \geq f(\mathcal{S}_{\text{gr}}) \geq (1 - 1/e) f(\mathcal{S}^*)$.

Note that these results are independent of the choice of channels $\mathbf{H}$, the number of antennas M_T, M_R, users K, and RF chains N. The advantage of exploiting submodularity is evident, as it allows us to obtain a constant-factor approximation for *all* possible instances of the NP–hard problem (13) for the case of the FF switching by only using a simple greedy algorithm. Clearly, this approach is more principled compared to convex relaxation. Moreover, it was shown in [42] that obtaining a better worst-case approximation guarantee would require one to evaluate f on an exponential number of subsets. Hence, in this case, the greedy algorithm is also provably an optimal polynomial-time approximation algorithm for the given problem.

On the other hand, for the PC switching case where the feasible subsets constitute a partition matroid, the greedy algorithm again starts from the empty set $\mathcal{S}_0 = \emptyset$ and proceeds as follows

$$\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{\operatorname*{argmax}_{\substack{v \notin \mathcal{S}_{i-1},\\ \mathcal{S}_{i-1} \cup \{v\} \in \mathcal{I}}} \Delta_f(v|\mathcal{S}_{i-1})\}, \forall\, i \in [N] \tag{22}$$

until we cannot add an element v for which $\mathcal{S}_{i-1} \cup \{v\} \in \mathcal{I}$. Regarding the performance of the greedy algorithm in this case, the following result holds.

Theorem 2: [25] For monotone submodular maximization over a partition matroid, the greedy algorithm yields a $1/2$-factor approximation for all instances of (13).[4] That is, $f(\mathcal{S}^*) \geq f(\mathcal{S}_{\text{gr}}) \geq 0.5 f(\mathcal{S}^*)$.

Again, the result is independent of the choice of the parameters $\mathbf{H}, M_T, M_R, K, N$ while additionally also being independent of the type of PC switching architecture used. Hence, by using the greedy algorithm, we are able to obtain a constant-factor approximation guarantee for all PC switching architectures, which is a trait not shared by the convex relaxation approach. We point out that the greedy algorithm *is not* the optimal polynomial-time approximation algorithm in the present case of maximization over a non-uniform matroid. In [43], Calinescu *et al.* proposed a polynomial-time algorithm for the given problem which guarantees a $(1 - 1/e)$-factor approximation. However, the improved performance of the algorithm comes at the expense of being unsuited for practical implementation,[5] for which reason, we choose to use the greedy algorithm instead. As we demonstrate later on in this section, in addition to its favorable theoretical approximation guarantees, the greedy algorithm also enjoys low run-time complexity.

We further remark that the curvature c_f of f can be used to refine the performance analysis of the greedy algorithm and obtain improved approximation guarantees. When maximization is performed subject to a uniform matroid constraint, it can be shown [35] that the greedy algorithm provides a $\frac{1}{c_f}(1 - e^{-c_f})$-factor approximation guarantee. Note that for $c_f = 1$, we obtain the worst-case $(1 - 1/e)$-factor, whereas for $c_f \to 0$, we have $\lim_{c_f \to 0} \frac{1}{c_f}(1 - e^{-c_f}) = 1$, which is the optimal factor (i.e., the greedy algorithm returns the optimal solution in this case). On the other hand, for maximization over a partition matroid, the approximation factor for the greedy algorithm can be improved to $\frac{1}{1+c_f}$ [35]. Again, the performance of the greedy algorithm in terms of curvature is sub-optimal amongst the class of polynomial-time algorithms. In [40], Sviridenko *et al.* introduce a pair of polynomial-time algorithms which guarantee a $(1 - c_f/e)$-factor approximation for submodular maximization subject to an arbitrary matroid constraint, which is an improvement over the curvature based guarantees of the greedy algorithm $\forall\, c \in (0, 1)$. The drawback of these algorithms is that they are far more computationally involved compared to the greedy algorithm. Furthermore, our experiments indicate that the greedy algorithm returns a near-optimal solution in most instances, for which reason it remains the algorithm of choice in this paper.

We now discuss the computational complexity of the greedy algorithm. Note that at each iteration $i \in [N]$, we are required to compute $O(M_T)$ determinants of a $\mathbb{C}^{|\mathcal{S}_i| \times |\mathcal{S}_i|}$ matrix, which requires $O(|\mathcal{S}_i|^3)$ computations in general. This results in overall complexity $O(N^4 M_T)$, which can be very unfavorable since $N = O(M_T)$ potentially. However, this can be avoided via a simple algebraic manipulation. Let us define the set

$$\mathcal{S}_i^{(v)} := \mathcal{S}_{i-1} \cup \{v\}, \forall\, v \notin \mathcal{S}_{i-1}, \forall\, i \in [N] \tag{23}$$

Then, as a consequence of Sylvester's Determinant theorem, we obtain

$$f(\mathcal{S}_i^{(v)}) = \log_2 \det\left(\mathbf{I}_{|\mathcal{S}_i|} + \rho(\tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]})^H \tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]}\right) \tag{24a}$$

$$= \log_2 \det\left(\mathbf{I}_{KM_R} + \rho\tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]}(\tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]})^H\right) \tag{24b}$$

Note that in this case, the complexity incurred in evaluating the determinant is $O(M_R^3 K^3)$, which results in a substantially improved complexity bound of $O(NM_T M_R^3 K^3)$ since $KM_R \leq N$ and $KM_R = o(M_T)$ in massive MIMO. Furthermore, as pointed out in [44], the incremental nature of the greedy algorithm can be exploited to further reduce complexity in the following manner.

$$\tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]}(\tilde{\mathbf{H}}^{[\mathcal{S}_i^{(v)}]})^H = \tilde{\mathbf{H}}^{[\mathcal{S}_{i-1}]}(\tilde{\mathbf{H}}^{[\mathcal{S}_{i-1}]})^H + \tilde{\mathbf{h}}^{[v]}(\tilde{\mathbf{h}}^{[v]})^H \tag{25}$$

[4]In fact, this result holds for maximization over *any* non-uniform matroid.

[5]More specifically, the algorithm requires solving a continuous nonlinear relaxation of problem (17). This necessitates computing a continuous extension of the objective function and its gradient at each iteration, which may require evaluating an exponential sum of subsets of $[M_T]$ in the worst case. While random sampling can be used to accurately approximate this sum using a polynomial number of subsets, the sample size may still be too large for practical implementation.

where $\tilde{\mathbf{h}}^{[v]}$ is the column of $\tilde{\mathbf{H}}$ indexed by v. Define

$$\mathbf{M}^{[\mathcal{S}_{i-1}]} := \mathbf{I}_{KM_R} + \rho \tilde{\mathbf{H}}^{[\mathcal{S}_{i-1}]}(\tilde{\mathbf{H}}^{[\mathcal{S}_{i-1}]})^H \tag{26}$$

Note that $\mathbf{M}^{[\mathcal{S}_{i-1}]}$ is always invertible and is available at the start of iteration i. We can now write (24b) as

$$f(\mathcal{S}_i^{(v)}) = \log_2 \det\left(\mathbf{M}^{[\mathcal{S}_{i-1}]} + \rho \tilde{\mathbf{h}}^{[v]}(\tilde{\mathbf{h}}^{[v]})^H\right) \tag{27a}$$

$$= \log_2(1 + \rho(\tilde{\mathbf{h}}^{[v]})^H(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}\tilde{\mathbf{h}}^{[v]}) \tag{27b}$$

$$+ \log_2 \det(\mathbf{M}^{[\mathcal{S}_{i-1}]}) \tag{27c}$$

where we have used the Matrix Determinant Lemma [45] to arrive at the second expression. Hence, computing the marginal gain of $v \notin \mathcal{S}_{i-1}$ w.r.t. $\mathcal{S}_{i-1}$ requires us to simply evaluate the quantity $\tilde{\mathbf{h}}^{[v]})^H(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}\tilde{\mathbf{h}}^{[v]}$. Assuming $(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}$ is available at the start of iteration $i \in [N]$,[6] evaluating this quadratic form requires $O(M_R^2 K^2)$ operations, thus resulting in $O(M_T M_R^2 K^2)$ operations to identify the optimal element v^*. Computation of $(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}$ for use in the next iteration can be accomplished effectively via use of the Sherman-Morrison-Woodbury Formula [41]

$$(\mathbf{M}^{[\mathcal{S}_i]})^{-1} = (\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1} - \frac{\rho(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}\tilde{\mathbf{h}}^{[v^*]}(\tilde{\mathbf{h}}^{[v^*]})^H(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}}{1 + \rho(\tilde{\mathbf{h}}^{[v^*]})^H(\mathbf{M}^{[\mathcal{S}_{i-1}]})^{-1}\tilde{\mathbf{h}}^{[v^*]}} \tag{28}$$

which only entails $O(M_R^2 K^2)$ operations relative to standard inversion which requires $O(M_R^3 K^3)$ operations. Ultimately, this translates into an overall runtime of $O(M_T N M_R^2 K^2)$ for the greedy algorithm, which is linear in M_T (with the other parameters fixed). Note that the running-time can still be $O(M_T^2)$ if the number of RF chains $N = O(M_T)$.

In order to further improve the run-time and efficiently tackle larger-scale problems (even when $N = O(M_T)$), we utilize an accelerated version of the standard greedy algorithm originally proposed in [46]. Note that at each iteration $i \in [N]$ of the standard greedy algorithm, determining the element v^* with the maximum marginal gain $\Delta_f(v^*|\mathcal{S}_{i-1})$ requires $O(M_T)$ evaluations of $f(\mathcal{S}_i^{(v)})$. However, as a simple consequence of submodularity, the marginal gains of any fixed element $v \in [M_T]$ are monotonically non-increasing across iterations of the greedy algorithm, i.e., $\Delta_f(v|\mathcal{S}_i) \geq \Delta_f(v|\mathcal{S}_j), \forall\, i \leq j \in [N]$. The accelerated greedy algorithm exploits this fact by maintaining a list $\{g(v)\}_{v \notin \mathcal{S}_{i-1}}$ of the marginal gains of the unselected elements at each iteration sorted in descending order. The initial list is obtained by performing one step of the standard greedy algorithm. At the beginning of every subsequent iteration, the algorithm extracts the maximal element from the list and performs the update $g(v) \leftarrow \Delta_f(v|\mathcal{S}_{i-1})$. After the update, if it so happens that $g(v) \geq g(v'), \forall\, v' \notin \mathcal{S}_{i-1} \cup \{v\}$, then by virtue of submodularity, it holds that $\Delta_f(v|\mathcal{S}_{i-1}) \geq \Delta_f(v'|\mathcal{S}_{i-1})$ for all such v'. Consequently, the greedy algorithm has identified v^* without having to evaluate $\Delta_f(v'|\mathcal{S}_{i-1})$ for a potentially large number of elements v'. If the above statement does not hold, the list is resorted and the maximal element is queried again, until v^* is identified. While the exact runtime of the accelerated greedy algorithm is currently unknown, it has been empirically observed to be much faster relative to the standard greedy algorithm.

[6] Note that $(\mathbf{M}^0)^{-1} = \mathbf{I}_{KM_R}$

VII. The Case of Multiple Sub-carriers

In this section, we extend the results of the previous section to the case of performing antenna selection at the BS across multiple sub-carriers. Given a single-cell MU-MIMO OFDM system with L downlink sub-carriers, K users, and a subset of active antennas $\mathcal{S} \subseteq [M_T]$ at the BS, the downlink channels can be described as

$$\mathbf{y}_k(\ell) = \sqrt{\rho}\mathbf{H}_k^{[\mathcal{S}]}(\ell)\mathbf{x}(\ell) + \mathbf{n}_k(\ell), \forall\,(\ell, k) \in [L] \times [K] \tag{29}$$

where for a given sub-carrier $\ell \in [L]$, $\mathbf{H}_k^{[\mathcal{S}]}(\ell), \mathbf{y}_k(\ell)$, $\mathbf{x}(\ell)$ and $\mathbf{n}_k(\ell)$ respectively denote the downlink MIMO channel submatrix for user k obtained by selecting the subset of transmit antennas $\mathcal{S}$, the MIMO channel output at the k^{th} user, the transmit signal applied across $\mathcal{S}$, and i.i.d. circularly symmetric Gaussian noise. In order to perform antenna selection across multiple sub-carriers, we employ a generalization of the following capacity based criterion proposed in [18]

$$\max_{\mathcal{S} \in \mathcal{I}} \sum_{\ell=1}^{L} \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^{K} (\mathbf{H}_k^{[\mathcal{S}]}(\ell))^H \mathbf{Q}_k(\ell) \mathbf{H}_k^{[\mathcal{S}]}(\ell)\right) \tag{30}$$

where the uplink covariance matrices $\{\mathbf{Q}_k(\ell)\}_{k=1,\ell=1}^{K,L}$ are again fixed *apriori*. Note that each function

$$f_\ell(\mathcal{S}) := \log_2 \det\left(\mathbf{I}_{|\mathcal{S}|} + \rho \sum_{k=1}^{K} (\mathbf{H}_k^{[\mathcal{S}]}(\ell))^H \mathbf{Q}_k(\ell) \mathbf{H}_k^{[\mathcal{S}]}(\ell)\right), \quad \forall\,\ell \in [L] \tag{31}$$

is monotone submodular by virtue of Propositions 1 and 2. Hence, problem (30) can be expressed as

$$\max_{\mathcal{S} \in \mathcal{I}} \left\{F(\mathcal{S}) := \sum_{\ell=1}^{L} f_\ell(\mathcal{S})\right\} \tag{32}$$

which corresponds to maximizing a sum of monotone submodular functions subject to matroid constraints. Since both submodularity and monotonicity are preserved under non-negative sums, problem (32) is equivalent to maximizing a monotone submodular function over the independent sets of a matroid $([M_T], \mathcal{I})$. Hence, the greedy algorithm can again be applied to (32) to obtain solutions with constant-factor approximation guarantees for all instances of (32) irrespective of the number of sub-carriers L and other aforementioned parameters; i.e., for the case of FF switching, we obtain a $(1 - 1/e)$-factor approximation guarantee whereas for PC switching, we obtain a $1/2$-factor guarantee.

VIII. An Efficient Algorithm for Convex Relaxation

In order to benchmark the performance of the greedy algorithms, we adopt the convex relaxation based heuristic originally proposed in [18] for obtaining approximate solutions of (32). The first step in this approach is to relax the combinatorial selection constraints to obtain the following optimization problem

$$\max_{\mathbf{s}\in\mathbb{R}^{M_T}} \left\{F(\mathbf{s}) := \sum_{\ell=1}^{L} \log_2 \det(\mathbf{I}_{KM_R} + \rho\tilde{\mathbf{H}}(\ell)\mathrm{diag}(\mathbf{s})\tilde{\mathbf{H}}(\ell)^H)\right\} \tag{33a}$$

$$\text{s.t.} \quad \mathbf{s} \in \mathcal{P} := \mathrm{conv}(\mathbb{1}_{\mathcal{S}} \,|\, \mathcal{S} \in \mathcal{I}) \tag{33b}$$

where in the case of FF switching, the feasible set is explicitly defined as

$$\mathcal{P} := \{\mathbf{s} \in \mathbb{R}^{M_T} \,|\, \mathbf{s} \in [0,1]^{M_T}, \mathbf{1}^T\mathbf{s} \leq N\} \tag{34}$$

while for PC switching, for a given partition of the antenna-array into B sub-arrays, we have

$$\mathcal{P} := \left\{\mathbf{s} \in \mathbb{R}^{M_T} \,|\, \mathbf{s} \in [0,1]^{M_T}, \sum_{v\in\mathcal{M}_b} \mathbf{s}(v) \leq N_b, \forall\, b \in [B]\right\} \tag{35}$$

Note that (33) corresponds to maximizing a concave function over a convex polytope, and hence, is a convex optimization problem which is optimally solvable in polynomial-time. However, instead of resorting to general-purpose convex programming solvers which incur $O(M_T^{3.5})$ complexity in the worst-case, we propose to use the Frank-Wolfe (FW) algorithm [47], [48], which is far more computationally efficient for large-scale instances. The algorithm is iterative in nature, with updates of the form

$$\mathbf{r}_k = \arg\max_{\mathbf{s}\in\mathcal{P}} \nabla F(\mathbf{s}_k)^T\mathbf{s} \tag{36a}$$

$$\mathbf{s}_{k+1} = \beta_k\mathbf{r}_k + (1-\beta_k)\mathbf{s}_k, \forall\, k \in \mathbb{N} \tag{36b}$$

where $\{\beta_k\}_{k\in\mathbb{N}}$ is a step-size sequence. By selecting the step-size rule $\beta_k = 2/(k+2), \forall\, k \in \mathbb{N}$, the algorithm requires $O(1/\epsilon)$ iterations to guarantee convergence to a solution which is ϵ-sub-optimal w.r.t. the optimal objective value [48].

At each step of the FW algorithm, we are required to compute the gradient $\mathbf{g}_k := \nabla F(\mathbf{s}_k)$ which can be evaluated using matrix calculus [45] as follows

$$\mathbf{g}_k(m) = \rho\sum_{\ell=1}^{L}\left(\tilde{\mathbf{h}}_m(\ell)^H(\mathbf{I}_{KM_R} + \rho\tilde{\mathbf{H}}(\ell)\mathrm{diag}(\mathbf{s}^{(k)})\tilde{\mathbf{H}}(\ell)^H)^{-1}\tilde{\mathbf{h}}_m(\ell)\right), \forall\, m \in [M_T] \tag{37}$$

where $\tilde{\mathbf{h}}_m(\ell)$ corresponds to the m^{th} column of the ℓ^{th} concatenated channel matrix $\tilde{\mathbf{H}}(\ell)$. Furthermore, we point out that the linear maximization problem (36a) over the polytope $\mathcal{P}$ always admits a closed form solution. For the case of FF switching, this is accomplished by first determining the index set $\mathcal{S}$ of the N largest components of $\mathbf{g}_k$ followed by setting $\mathbf{r}_k^{[\mathcal{S}]} = 1$ and the remaining elements to be zero. The same principle also applies to PC switching, where for each sub-array $\mathcal{M}_b$ we determine the index of the $|\mathcal{S}_b|$ largest components of $\mathbf{g}_k^{[\mathcal{M}_b]}$, and then set the corresponding elements of $\mathbf{r}_k^{[\mathcal{M}_b]}$ to one and the rest to zero. Note that in both cases, the solution of (36a) is an extreme point of $\mathcal{P}$, which corresponds to selecting a subset of antennas which satisfy the given selection constraints. The next iterate $\mathbf{s}_{k+1}$ is then obtained by performing a convex combination with the past iterate. The per-iteration complexity is dominated by the cost of forming $\mathbf{g}_k$, which is $O(LM_TK^2M_R^2(KM_R+1))$, while the cost of computing the maximization step (36a) is at most $O(M_TN)$. This results in overall complexity $O(\frac{M_T}{\epsilon}(LK^2M_R^2(KM_R+1)+N))$, which scales only linearly with the number of transmit antennas M_T. A post-processing rounding step is then performed on the solution returned by the algorithm to obtain the set of antennas $\mathcal{S}_{\mathrm{FW}}$.

We remark that there are no known theoretical guarantees regarding the sub-optimality of the set $\mathcal{S}_{\mathrm{FW}}$ in contrast to $\mathcal{S}_{\mathrm{gr}}$. However, the FW algorithm can be interpreted as a *continuous* greedy algorithm, as it selects a subset of antennas that maximizes the linear objective (36a) at each iteration, and hence, the solution computed by it can still serve as a reasonable baseline for comparison.

IX. Simulation Results

In this section, we present simulation results to demonstrate the viability of our outlined antenna selection approach. We performed all our experiments on a Linux desktop with Intel i7 cores and 16 GB of RAM and averaged our results over 500 Monte-Carlo channel realizations. First, we carried out a preliminary experiment where a BS equipped with $M_T = 20$ transmit antennas serves $K = 3$ single receive antenna users in a rich scattering environment where the downlink channels are modeled using Rayleigh fading. A FF RF switching architecture is used at the BS for selecting antennas with a transmit power budget of $\rho = -2$ dB. Due to the modest size of the problem, we can afford to run exhaustive search in this case to compute the optimal solution of (13) and use it as a performance benchmark. Note that in the presence of Rayleigh fading, the average power of the channel coefficients is the same across all transmit antennas and users, and thus, even random antenna selection is expected to work well in this setting. For a given realization of $\mathbf{H}$, we first compute the optimal uplink covariance matrices by solving (2). Since $M_R = 1$, this simply entails computing a vector of user power allocations, and can be efficiently computed using the FW algorithm. Although we do not explicitly outline the details, the algorithm proceeds in a manner very similar to the one described in Section VIII; the main difference being that the constraints are now described by the $K-$dimensional probability simplex. We initialize FW from the center of the simplex and use a relative tolerance of $\epsilon = 1e^{-4}$ w.r.t. the objective function for termination. Thereafter, we use the greedy algorithm, the convex relaxation approach using FW, and random selection to obtain approximate solutions for (13). We varied the number of available RF chains N from 3 to 15 and compared the performance of both approaches in terms of spectral efficiency against the optimal solution of (13) obtained via exhaustive search. Furthermore, for comparison, we also computed the optimal solution of (4); i.e., the problem of joint antenna selection and power allocation via exhaustive enumeration over all possible antenna

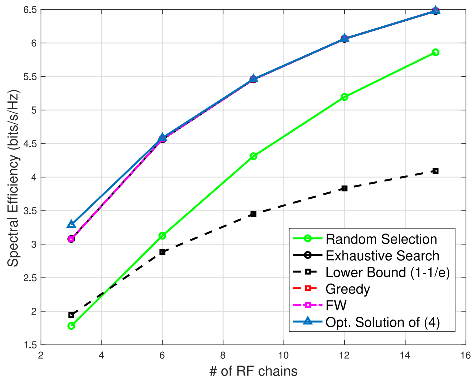


Fig. 1. Spectral Efficiency vs number of RF chains N for $M_T = 20$ BS antennas with fully flexible RF switching in presence of Rayleigh fading, $\rho = -2$ dB, $K = 3$ users with $M_R = 1$ receive antenna.

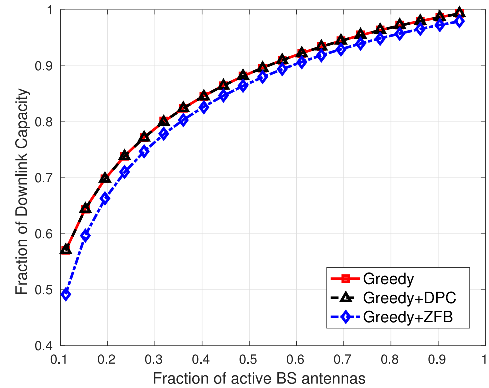


Fig. 2. Fraction of downlink sum-rate capacity vs fraction of active BS antennas for $M_T = 144$ BS antennas with FF RF switching, $\rho = 10$ dB, $K = 12$ users with $M_R = 1$ receive antenna.

TABLE I
WORST-CASE APPROXIMATION QUALITY OF OBTAINED SOLUTIONS (IN %)

N	3	6	9	12	15
Greedy	96.44	99.80	99.81	100	100
FW	96.20	99.54	99.86	99.93	100
Random	30.79	30.30	57.56	65.41	76.29

subsets and solving a power allocation problem of the form (3) for each subset of selected antennas. The results are depicted in Fig. 1 while the worst-case approximation quality of solutions obtained via random selection, the greedy and FW algorithms are presented in Table I. It is clear that the greedy and FW algorithms exhibit near-optimal performance relative to exhaustive search in all cases in contrast to random selection, which can perform very poorly with few RF chains. Note that the quality of the solution determined by the greedy algorithm is significantly better than the worst-case $(1 - 1/e)$ lower bound. We additionally computed the improved $\frac{1}{c_f}(1 - e^{-c_f})$ curvature-based approximation factor by using the curvature bound mentioned in [40] and by explicitly computing c_f via (9). However, the best approximation factor we obtained was only 0.685 (on average), which is still very pessimistic compared to what we observe in practice. It is also worth pointing out that in this particular case (with independent Rayleigh fading), the solutions of (13) obtained via exhaustive search, the greedy and FW algorithms turn out to be near-optimal for the joint selection and power allocation problem (4) as well.

Having demonstrated the ability of the greedy algorithm to compute very high quality solutions for the antenna selection problem, we now turn to the question of whether this translates into attaining a significant fraction of the total sum-rate capacity in the massive MIMO regime. We consider a setup with $M_T = 144$ antennas at the BS with FF switching, $\rho = 0$ dB, and $K = 12$ single receive antenna users. For determining the downlink channels $\{\mathbf{h}_k^H\}_{k=1}^K$, we use the model [49]

$$\mathbf{h}_k^H = \sqrt{\frac{M_T}{L_k}} \sum_{l=1}^{L_k} \alpha_k^{(l)} \mathbf{a}_t(\phi_k^{(l)})^H, \forall\, k \in [K] \quad (38)$$

where $L_k \sim \mathcal{U}\{5, 6, \ldots, 15\}$ is the number of scattering paths between the BS and the k^{th} user that is modeled as a uniform random variable with non-zero support on the set $\{5, \ldots, 15\}$, $\alpha_k^{(l)} \sim \mathcal{CN}(0, 1)$ is the complex gain of the l^{th} path between the BS and the k^{th} user, $\mathbf{a}_t(.)$ is the array response vector at the transmitter, and $\phi_k^{(l)} \sim \mathcal{U}[-\pi/2, \pi/2]$ denotes the azimuth angle of departure (AoD) of the l^{th} path for the k^{th} user. Assuming the BS is equipped with a uniform linear array (ULA), we have

$$\mathbf{a}_t(\phi) = \frac{1}{\sqrt{M_T}}[1, e^{jkd\sin(\phi)}, \ldots, e^{jkd(M-1)\sin(\phi)}]^T \quad (39)$$

where $k = 2\pi/\lambda$, λ is the carrier wavelength and $d = \lambda/2$ is the spacing between antenna elements. For each realization of $\mathbf{H}$, we again computed the optimal power allocation using FW as described in the previous paragraph. The result was used to determine the sum-rate downlink capacity using the full set of antennas. Next, we ran the greedy algorithm on (13) using a fixed number of RF chains and computed the resulting sum-rate with the selected subset of antennas. Then, we removed the inactive antennas and used one more step of FW to determine the optimal user power allocation for the reduced MIMO MAC channel. By duality of the MAC and BC, the resulting sum-rate is also what is theoretically achievable using DPC at the BS. Owing to the impracticality of DPC, we also used the selected subset of antennas to design a zero-forcing beamformer for the reduced MIMO BC as outlined in Appendix A. We repeated this over varying numbers of RF chains ranging from 16 to 136. The obtained results are shown in Fig. 2. It can be seen that even with only 11% of the antennas activated, the greedy algorithm captures almost 60% of the total downlink capacity on average. Furthermore, due to the monotonicity and submodularity of the objective function of (13), the property of diminishing returns is evident; i.e., the marginal gains in sum-rates decreases as more antennas are activated. Overall, this enables us to capture a significant fraction of the total downlink capacity using only a small subset of antennas determined by the greedy algorithm, which is very efficient in practice. We also note that the DPC rate obtained by performing user power re-allocation on the reduced MIMO BC is negligible in terms of improvement over the fraction of the DPC rate of the MIMO

BC with all antennas activated attained by the greedy algorithm. While we could combine the greedy antenna selection and power allocation steps within an alternating optimization formulation framework to obtain approximate solutions for (4), due to the tangible lack of improvement in the objective function, we do not pursue this approach further. Finally, it is evident that using ZFB on the subset of antennas selected by the greedy algorithm incurs minimal performance loss relative to using DPC, thus providing empirical justification for our usage of the two-step greedy submodular selection followed by ZFB approach. Indeed, ZFB with greedy antenna selection captures 70% of the total sum-rate with only a fourth of the total antennas activated.

We now extend our experiments to the case where $M_R > 1$ and a RF switching matrix with partial connectivity is employed at the BS. We employ a generalization of the downlink channel model in (38) to the multiple receive antenna case, where each channel matrix is described as

$$\mathbf{H}_k = \sqrt{\frac{M_T M_R}{L_k}} \sum_{l=1}^{L_k} \alpha_k^{(l)} \mathbf{a}_{r_k}(\psi_k^{(l)}) \mathbf{a}_t(\phi_k^{(l)})^H, \forall\, k \in [K] \tag{40}$$

where in addition to the previously defined quantities, $\mathbf{a}_{r_k}(.)$ denotes the array response vector at the k^{th} receiver, and $\psi_k^{(l)} \sim \mathcal{U}[-\pi/2, \pi/2]$ denotes the azimuth angle of arrival (AoA) of the l^{th} path for the k^{th} user. Here we assume that each user is equipped with a ULA as well. For our experiments, we set the uplink covariance matrices to be $\mathbf{Q}_k = \frac{1}{KM_R}\mathbf{I}_{KM_R}, \forall k \in [K]$, which, as mentioned previously, is a good choice in the high SNR regime; we avoid using the iterative waterfilling algorithm described in [31] as it is computationally cumbersome. In our first experiment, we consider a scenario with $M_T = 128$ BS antennas, $K = 8$ users with $M_R = 2$ antennas each, and $\rho = 20$ dB. We employ a PC switching network at the BS which partitions the antenna array into $B = 8$ disjoint sub-arrays of 16 antennas each. For a given number of RF chains, we assign an equal number to each sub-array and run the greedy algorithm to determine the subset of antennas. Thereafter, we apply ZFB on the selected antenna subset both with and without coordination amongst the receive antennas of each user. The total number of RF chains is increased from 24 to 112 in increments of 8 for each realization of $\mathbf{H}$, and the results averaged over 500 channel realizations are depicted in Fig. 3, where the fraction of total downlink capacity is plotted as a function of the number of active antennas per sub-array. It is evident that the greedy algorithm again produces high quality solutions, attaining 80% of the total capacity on average with only 4 active antennas per sub-array (i.e., 32 out of 128 total antennas). Additionally, both zero-forcing approaches perform very well with the selected antennas, with only modest performance losses. As expected, ZFB without receive antenna coordination (depicted in solid blue), is the slightly inferior of the two approaches. However, even this approach attains excess of 75% of the total capacity with only 5 active antennas per sub-array (or approximately 31% of the total antennas).

We also performed a similar experiment in a more challenging scenario with $M_T = 256$ BS antennas, $K = 16$ users with $M_R = 4$ antennas each, and $\rho = 20$ dB. The PC switching network at the BS partitions the antenna array into $B = 16$ disjoint sub-arrays of 16 antennas each. The averaged results are depicted in Fig. 4, with the greedy algorithm exhibiting similar performance as the previous experiment. However, the performance of ZFB without receive antenna coordination is decidedly more inferior in this case due to the larger number of receive antennas per user; i.e., the payoff in employing coordination is more significant here. For example, attaining 70% of the total capacity using ZFB with coordination requires activating 8 BS antennas per sub-array (or 50% of the total antennas) while achieving the same without coordination requires activating 75% of the total antennas.

We now consider the case of multiple sub-carriers with PC switching, which is arguably the most challenging case. Additionally, we also add the comparison with the convex relaxation heuristic based on the FW algorithm described in Section VIII. First, we consider a scenario with $M_T = 192$, $M_R = 1$, $K = 24$, $L = 64$ downlink sub-carriers and a PC switching architecture with $B = 24$ sub-arrays of 8 antennas each, $N = 48$ RF chains, with 2 active antennas per sub-array; i.e., only a fourth of the antennas can be activated in total and also per sub-array. The channel model (38) is used to generate the downlink

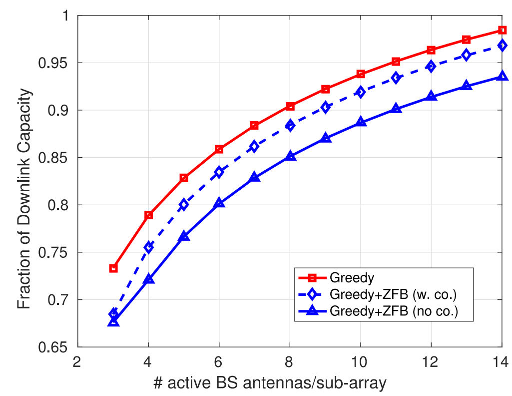


Fig. 3. Fraction of downlink sum-rate capacity vs number of active BS antennas per sub-array for $M_T = 128$ BS antennas with PC RF switching using $B = 8$ sub-arrays of equal size, $\rho = 20$ dB, $K = 8$ users with $M_R = 2$ receive antennas.

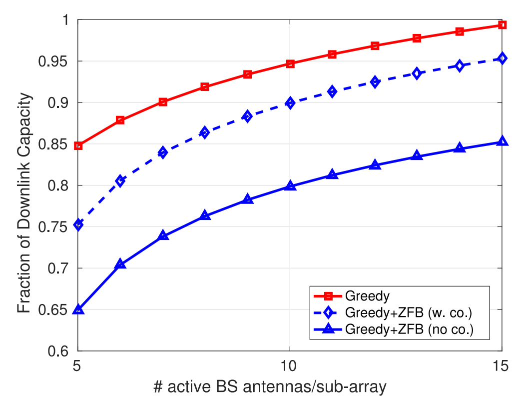


Fig. 4. Fraction of downlink sum-rate capacity vs number of active BS antennas per sub-array for $M_T = 256$ BS antennas with PC RF switching using $B = 16$ sub-arrays of equal size, $\rho = 20$ dB, $K = 16$ users with $M_R = 4$ receive antennas.

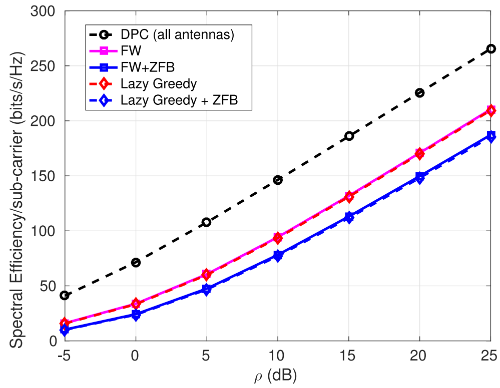


Fig. 5. Spectral efficiency/sub-carrier vs ρ for $M_T = 192$ BS antennas with PC RF switching using $B = 24$ sub-arrays with 8 antennas each, 2 active antennas/sub-array, $L = 64$ sub-carriers, $K = 24$ users with $M_R = 1$ receive antenna.

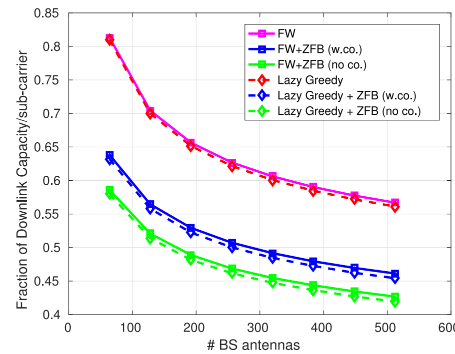
TABLE II
FRACTION OF TOTAL DOWNLINK CAPACITY ATTAINED WITH 3 ACTIVE ANTENNAS/SUB-ARRAY

ρ (dB)	-5	0	5	10	15	20	25
Lazy Gr.	0.53	0.60	0.68	0.75	0.80	0.83	0.86
Lazy Gr. + ZFB	0.40	0.51	0.61	0.69	0.74	0.79	0.82

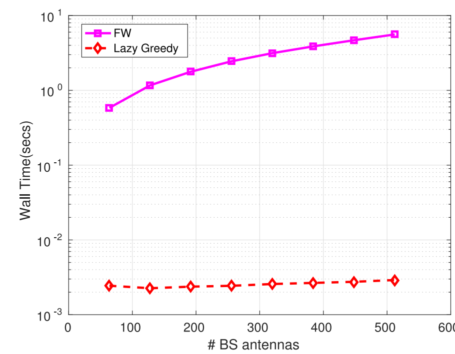
channels. We first compute the optimal user power allocation per sub-carrier. Next, antenna selection across all sub-carriers is performed using the FW based relaxation heuristic, which is run for a maximum of 1000 iterations or until it achieves convergence w.r.t. the objective function, as measured by a relative tolerance factor of $\epsilon = 1e^{-5}$. We also modify the greedy algorithm by utilizing lazy evaluations to improve its running time. Finally, ZFB is applied on the selected subset of antennas to compute the sum-rate per sub-carrier. We depict the average sum-rate per sub-carrier in Fig. 5 as a function of the transmit power budget ρ. Surprisingly, the FW based relaxation plus rounding heuristic works just as well as the greedy algorithm in terms of its performance relative to the sum-rate capacity achievable with all antennas activated. However, the greedy algorithm with lazy evaluations is significantly faster, with an average execution time of just 3 milli-secs compared to 2 secs for FW. Hence, we obtain a speed-up factor of almost 700% by using the greedy algorithm. We note that when the transmit power budget is very small, then the stringency of the switching constraints and the fact that selection is performed across sub-carriers results in greedy selection plus ZFB attaining only 25% of the total sum-rate capacity. Considering that only a fourth of the antennas are active anyways, even in this very adversarial setting, the performance is still satisfactory. By increasing the transmit power budget, the performance markedly improves. For $\rho = 10$ dB, we can attain excess of 50% of the total capacity on average, while this figure improves to 70% for $\rho = 25$ dB. We also performed additional experiments where we increased the number of antennas which can be activated per sub-array (i.e., we increased the number of RF chains) while keeping the other parameters fixed. The results are reported in Tables II and III, from which it is evident that the activating more antennas results in attaining higher sum rates at lower transmit power.

TABLE III
FRACTION OF TOTAL DOWNLINK CAPACITY ATTAINED WITH 4 ACTIVE ANTENNAS/SUB-ARRAY

ρ (dB)	-5	0	5	10	15	20	25
Lazy Gr.	0.64	0.71	0.78	0.83	0.86	0.88	0.90
Lazy Gr. + ZFB	0.54	0.64	0.72	0.78	0.82	0.85	0.87



(a) Fraction of total downlink capacity/sub-carrier vs M_T



(b) Wall Time vs M_T

Fig. 6. Fraction of downlink sum-rate capacity attained/sub-carrier and wall-times vs number of BS antennas using $N_{RF} = 32$ RF chains in a PC RF switching network with $B = 32$ sub-arrays of equal size, $\rho = 20$ dB, $L = 32$ sub-carriers, $K = 12$ users with $M_R = 2$ receive antennas.

In our last experiment, we test the performance of FW and lazy greedy in another challenging scenario where $K = 12$ users with $M_R = 2$ antennas are served by a BS with $N_{RF} = 32$ RF chains and $L = 32$ sub-carriers. We set $\rho = 20$ dB and vary the number of BS antennas from $M_T = 64$ to 512. A PC switching network is employed which partitions the entire array into $B = 32$ sub-arrays of equal size; i.e., only 1 antenna can be activated per sub-array. For $M_T = 64$, our setup exactly corresponds to the case of binary switching described in [19]. We depict the performance in terms of downlink capacity attained per sub-carrier and timing in Fig. 6(a) and 6(b) respectively. It can be observed that FW is marginally better than lazy greedy in terms of sum rate; however, in terms of running time, the greedy algorithm has a very significant advantage. The speed-up obtained varies from 250% for $M_T = 64$ and increases to 2000% for $M_T = 512$. Clearly, lazy greedy exhibits a vastly superior performance-complexity trade-off compared to FW. While greedy selection performed on the entire MIMO BC with DPC exhibits very favorable performance in terms of sum-rate attained versus fraction of active antennas, designing a ZFB with the selected antennas for the reduced MIMO BC results in a fairly substantial drop in sum-rate. Yet, even with only 6.25% of the total antennas active (i.e., $M_T = 512$), this scheme attains 45% (42%) of the total capacity with (without) receive antenna coordination, which is still impressive.

X. Conclusion

We demonstrated that the NP–hard problem of maximizing downlink sum-rate capacity in a multi-user massive MIMO system with fixed user power allocation subject to FF and PC RF switching constraints on the selected transmit antennas corresponds to a monotone submodular maximization problem over the independent sets of a matroid. In contrast to convex relaxation based approaches, the benefit of viewing the problem through the lens of submodularity is that a simple greedy algorithm can be used to guarantee constant-factor approximation for all problem instances, independent of a large number of system parameters in massive MIMO. Simulations revealed that not only does the greedy algorithm return a near-optimal solution in practice, but it can also capture a significant fraction of the total downlink sum-rate capacity at far lower complexity compared to convex relaxation. Furthermore, when a ZFB is designed using the antenna subset obtained via greedy selection, the resulting performance loss in terms of sum-rate was observed to be modest. This two step combination of greedy antenna selection followed by applying a ZFB was demonstrated to work well even when selection is performed across multiple sub-carriers under restrictive switching constraints. The overall approach is well aligned with the vision of massive MIMO, as it uses simple signal processing techniques at the BS to attain favorable trade-offs between performance and hardware efficiency. Hence, it has considerable potential for being employed in practice for reducing the complexity of implementing massive MIMO systems.

Appendix A
Zero-Forcing Beamformer Design

As the transmit antenna selection schemes outlined in the paper utilize the downlink sum-rate capacity of the MIMO BC channel as the selection criterion, achieving the fraction of the total sum-rate with the selected antennas requires employing DPC at the BS. This is an undesirable proposition, due to the high complexity of implementing DPC. Instead, as proposed in [18], we evaluate the performance of the antenna selection schemes using the much simpler zero-forcing beamforming (ZFB) strategy, which, while being sub-optimal relative to DPC, is known to demonstrate very effective performance in the massive MIMO regime. In this section, we provide a brief overview of the ZFB approach, as presented in [50].

We assume w.l.o.g. that $L = 1$, as the ZFB design problem decouples across subcarriers for a given set of antennas. First, we consider the case of $M_R = 1$. In a linear transmit beamforming strategy, each user stream is precoded by a different beamforming vector. Given a subset of N selected antennas $\mathcal{S}$ at the BS, we define the beamforming matrix $\mathbf{W} := [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{K \times N}$, power allocation matrix $\mathbf{P} := \mathrm{diag}(p_1, \ldots, p_K)$ and the transmit symbol vector $\mathbf{s} \in \mathbb{C}^K$. It then follows that $\mathbf{x} = \mathbf{WPs} = \sum_{k=1}^K s_k p_k \mathbf{w}_k$ and the received signal at each user can be expressed as

$$y_k = \mathbf{H}_k^{[\mathcal{S}]} p_k s_k \mathbf{w}_k + \sum_{j=1, j\neq k}^{K} \mathbf{H}_k^{[\mathcal{S}]} p_j s_j \mathbf{w}_j + \mathbf{n}_k, \forall\, k \in [K] \tag{41}$$

In ZFB, the matrix $\mathbf{W}$ is chosen to be the right inverse of the concatenated channel matrix $\mathbf{H}^{[\mathcal{S}]}$ in order to cancel inter-user interference. The sum-rate achievable with this scheme is

$$\begin{aligned} R_{\mathrm{ZFB}}(\mathcal{S}) = \max_{\mathbf{P}} \sum_{k=1}^{K} \log_2(1 + \rho p_k) \\ \text{s.t.} \sum_{k=1}^{K} \gamma_k^{-1} p_k \leq 1 \end{aligned} \tag{42}$$

where $\gamma_k^{-1} = \|\mathbf{w}_k\|_2^2 = ([\mathbf{H}^{[\mathcal{S}]}(\mathbf{H}^{[\mathcal{S}]})^H]^{-1})(k,k)$. The optimal power allocation in (42) can be computed via waterfilling as $p_k = \max\{\mu\gamma_k - 1/\rho, 0\}, \forall\, k \in [K]$, where μ is the solution of the nonlinear equation $\sum_{k=1}^K \max\{\mu\gamma_k - 1/\rho, 0\} = 1$.

For the case of $M_R > 1$, co-ordination amongst the receive antennas at each user can be utilized to improve the performance of ZFB via the following post-processing step at each user: denoting the thin SVD of $\mathbf{H}_k^{[S]}$ as $\mathbf{H}_k^{[\mathcal{S}]} = \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^H, \forall\, k \in [K]$, we perform the receive filtering step $\tilde{\mathbf{y}}_k = \mathbf{U}_k^H \mathbf{y}_k$. Thereafter, the collection of processed vectors $\{\tilde{\mathbf{y}}_k\}_{k=1}^K$ can be viewed as the output of a MIMO BC with KM_R single antenna users whose downlink channels are given by $\{s_{k,m_r} \mathbf{v}_{k,m_r}^H\}_{k=1,m_r=1}^{K,M_R}$, where $\mathbf{v}_{k,m_r}$ is the m_r^{th} column of the matrix $\mathbf{V}_k$. It then follows that the ZFB scheme described in the previous paragraph can be applied. Note that this strategy requires the BS to provide feedback to all users regarding the index of the transmit antennas selected. This can be avoided by opting to not perform receive antenna co-ordination and instead treating each antenna as belonging to a separate user, which allows a straightforward application of the ZFB technique described in the previous paragraph.

Finally, one may naturally question why the transmit antennas selected for DPC should perform well for ZFB, and whether it is more reasonable to base the selection criterion directly upon ZFB instead. To this, we point out that: i) for the ZFB selection problem, we are unaware of the existence of any polynomial-time algorithm which outputs a solution with provable approximation guarantees for all choices of problem parameters, and hence, one runs the potential risk of selecting a set of antennas which can exhibit very poor performance for an adversarial instance; and ii) our experiments indicate that the loss in sum-rate incurred by using ZFB for the antenna subset determined via submodular maximization is modest. For this reason, we adopt the two-step submodular selection followed by ZFB approach in this paper.

Appendix B
The Single User Case

In this section, we discuss the case of transmit antenna selection for a point-to-point MIMO channel equipped with M_T

transmit antennas and M_R receive antennas. For a given channel matrix $\mathbf{H} \in \mathbb{C}^{M_R \times M_T}$, we consider the following problem of selecting a subset $\mathcal{S}$ of N transmit antennas mentioned in [27]

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\in\mathcal{I}} \left\{ f(\mathcal{S}) := \log_2 \det\left(\mathbf{I}_{M_R} + \frac{\rho}{|\mathcal{S}|} \mathbf{H}^{[\mathcal{S}]} (\mathbf{H}^{[\mathcal{S}]})^H \right) \right\} \quad (43)$$

where $\mathbf{H}^{[\mathcal{S}]} \in \mathbb{C}^{M_R \times N}$ is the MIMO channel sub-matrix obtained by selecting a subset of N columns (indexed by $\mathcal{S}$) from $\mathbf{H}$ and $\mathcal{I} \subseteq [M_T]$ corresponds to the independent sets of a uniform or partition matroid, corresponding to FF or PC RF switching, respectively. Note that the transmit power budget ρ is split equally amongst all the selected antennas. According to the authors of [27], the objective function of (43) is not submodular, and the following counter-example is provided to establish the claim. A special case of the MIMO channel with $M_T = 2$ and $M_R = 1$ is considered, where the channel coefficient between the i^{th} transmit antenna and the receive antenna is denoted by $h_i, \forall\, i \in \{1, 2\}$. Then, we have

$$f(\{1\}) = \log_2(1 + \rho|h_1|^2),\ f(\{2\}) = \log_2(1 + \rho|h_2|^2),$$
$$f(\{1,2\}) = \log_2\left(1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)\right),\ f(\{\emptyset\}) = 0 \quad (44)$$

As correctly pointed out in [27], for this example, $f(\mathcal{S})$ is not guaranteed to be monotone for all possible $\rho, |h_1|^2, |h_2|^2$. The authors then go on to claim that this also implies that $f(\mathcal{S})$ is not submodular. However, monotonicity is not a necessary condition for a set-function to be submodular. For example, graph-cut is a well known set-function which is non-monotone, and yet, submodular [51]. We now establish that this is the case for the counter-example provided in [27] as well; i.e., $f(\mathcal{S})$ is actually submodular. For this purpose, from Definition 1 in Section IV, it suffices to show that

$$\begin{aligned} & f(\{1\}) + f(\{2\}) \geq f(\{1,2\}) + f(\{\emptyset\}) \\ \Rightarrow\ & \log_2(1 + \rho|h_1|^2) + \log_2(1 + \rho|h_2|^2) \\ & \qquad \geq \log_2\left(1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)\right) \end{aligned} \quad (45)$$

which is equivalent to the condition

$$\begin{aligned} & \log_2\left(\frac{(1 + \rho|h_1|^2)(1 + \rho|h_2|^2)}{1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)} \right) \geq 0 \\ \iff & \frac{1 + \rho|h_1|^2 + \rho|h_2|^2 + \rho^2|h_1|^2|h_2|^2}{1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)} \geq 1 \\ \iff & 1 + \frac{\frac{\rho}{2}(|h_1|^2 + |h_2|^2 + 2\rho|h_1|^2|h_2|^2)}{1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)} \geq 1 \\ \iff & \frac{|h_1|^2 + |h_2|^2 + 2\rho|h_1|^2|h_2|^2}{1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)} \geq 0 \end{aligned} \quad (46)$$

Simple inspection reveals that the last condition is indeed valid for all $\rho, |h_1|^2, |h_2|^2$, and, consequently, $f(\mathcal{S})$ is a submodular function. Unfortunately, this implies that the counter-example provided in [27] is incorrect, and thus, cannot be used to establish that the general case of $f(\mathcal{S})$ is non-submodular.

Instead, a valid counter-example can be constructed for $M_T = 3$ and $M_R = 1$, as we now show. Consider the case of $\rho = 1$ and channel coefficients with powers $|h_1|^2 = 5$, $|h_2|^2 = 15$ and $|h_3|^2 = 1$. Given the antenna subsets $\mathcal{A} := \{1\}$ and $\mathcal{B} := \{1, 3\}$ (note $\mathcal{B} \supset \mathcal{A}$), we compute the incremental gain of adding the antenna element $v := \{2\}$ to both sets. Doing so, we obtain

$$\begin{aligned} \Delta_f(v|\mathcal{A}) &= f(\{1,2\}) - f(\{1\}) = \log_2 11 - \log_2 6 = 0.8745 \\ \Delta_f(v|\mathcal{B}) &= f(\{1,2,3\}) - f(\{1,3\}) = \log_2 8 - \log_2 4 = 1 \end{aligned} \quad (47)$$

Clearly, $\Delta_f(v|\mathcal{A}) < \Delta_f(v|\mathcal{B})$, which is a violation of the diminishing returns property (Definition 2 in Section IV). As a result, $f(\mathcal{S})$ is not submodular in general.

Finally, we point out that if the power allocation is fixed to ρ/M_T beforehand (i.e., we have equal power splitting amongst *all* antennas), then the objective function of (43) is indeed submodular. This can be seen by first applying Sylvester's Determinant theorem to obtain $f(\mathcal{S}) := \log_2 \det(\mathbf{I}_{|\mathcal{S}|} + \frac{\rho}{M_T}(\mathbf{H}^{[\mathcal{S}]})^H \mathbf{H}^{[\mathcal{S}]})$, which can also be represented in the form of Equation (12) in Section V.

References

[1] A. Konar and N. D. Sidiropoulos, "Greed is good: Leveraging submodularity for antenna selection in massive MIMO," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2017, pp. 1522–1526.

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

[3] J. G. Andrews *et al.*, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large scale arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[6] E. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

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

[8] S. Sandhu, R. Nabar, D. Gore, and A. Paulraj, "Near-optimal selection of transmit antennas for a MIMO channel based on Shannon capacity," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2000, pp. 567–571.

[9] R. W. Heath Jr., S. Sandhu, and A. Paulraj, "Antenna selection for spatial multiplexing systems with linear receivers," *IEEE Commun. Lett.*, vol. 5, pp. 142–144, Apr. 2001.

[10] D. Gore, R. W. Heath Jr, and A. Paulraj, "Transmit selection in spatial multiplexing systems," *IEEE Commun. Lett.*, vol. 6, no. 11, pp. 491–493, Nov. 2002.

[11] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[12] R. Chen, J. Andrews, and R. W. Heath Jr., "Efficient transmit antenna selection for multiuser MIMO systems with block diagonalization," in *Proc. IEEE Global Telecommun.*, Washington, DC, USA, Nov. 2007, pp. 3499–3503.

[13] H. A. Seleh and W. Hamouda, "Transmit antenna selection for decision feedback detection in MIMO fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4440–4444, Sep. 2009.

[14] O. Mehanna, N. D. Sidiropoulos, and G. B. Giannakis, "Joint multicast beamforming and antenna selection," *IEEE Trans. Signal Process.*, vol. 61, pp. 2660–2674, May 2013.

[15] S. Khademi, S. Chepuri, G. Leus, and A. J. van der Veen, "Zero-forcing pre-equalization with transmit antenna selection in MIMO systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 5046–5050.
[16] H. Li, L. Song, and M. Debbah, "Energy efficiency of large-scale multiple antenna systems with transmit antenna selection," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 638–647, Feb. 2014.
[17] M. Gkizeli and G. Karystinos, "Maximum-SNR antenna selection among a large number of transmit antennas," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 891–901, Oct. 2014.
[18] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3917–3928, Nov. 2015.
[19] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Multi-switch for antenna selection in massive MIMO," in *Proc. IEEE Global Telecommun.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[20] A. G.-Rodriguez, C. Masouros, and P. Rulikowski, "Reduced switching connectivity for large scale antenna selection," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2250–2263, May 2017.
[21] J. Lee, "Maximum entropy sampling," in *Encyclopedia of Environmetrics*, vol. 3, A. H. El-Shaarawi, and W. W. Piegorsch, Eds. Chichester, U.K.: Wiley, 2001, pp. 1229–1234.
[22] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed. (Annals of Discrete Mathematics 58). Amsterdam, The Netherlands: Elsevier Science, 2005.
[23] J. Oxley, *Matroid Theory*. London, U.K.: Oxford Univ. Press, 2011.
[24] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Program.*, vol. 14, no. 1, pp. 265–294, Dec. 1978.
[25] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions—II," *Math. Program. Stud.*, vol. 8, pp. 73–87, 1978.
[26] A. Gorokhov, D. Gore, and A. Paulraj, "Receive antenna selection for MIMO flat-fading channels: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2687–2696, Oct. 2003.
[27] R. Vaze and H. Ganapathy, "Sub-modularity and antenna selection in MIMO systems," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1446–1449, Sep. 2012.
[28] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.
[29] P. Viswanath and D. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.
[30] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.
[31] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.
[32] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
[33] C. W. Ko, J. Lee, and M. Queyranne, "An exact algorithm for maximum entropy sampling," *Oper. Res.*, vol. 43, no. 4, pp. 684–691, Aug. 1995.
[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[35] M. Conforti and G. Cornuejols, "Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the Rado-Edmonds theorem," *Discrete Appl. Math.*, vol. 7, no. 3, pp. 251–274, Mar. 1984.
[36] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Inf. Control*, vol. 39, no. 1, pp. 55–72, Oct. 1978.
[37] A. K. Kelmans and B. N. Kimelfeld, "Multiplicative submodularity of a matrix's principal minor as a function of the set of its rows," *Discrete Math.*, vol. 44, no. 1, pp. 113–116, 1983.
[38] A. Krause, A. Singh, and C. Guestrin, "Near optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, pp. 235–284, Feb. 2008.
[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[40] M. Sviridenko, J. Vondrak, and J. Ward, "Optimal approximation for submodular and supermodular optimization with bounded curvature," in *Proc. ACM-SIAM SODA*, San Diego, CA, USA, Jan. 2015, pp. 1134–1148.
[41] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2013.
[42] G. L. Nemhauser and L. A. Wolsey, "Best algorithms for approximating the maximum of a submodular set function," *Math. Oper. Res.*, vol. 3, no. 3, pp. 177–188, Aug. 1978.
[43] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrak, "Maximizing a submodular set function subject to a matroid constraint," *SIAM J. Comput.*, vol. 40, no. 6, pp. 1740–1766, Dec. 2011.
[44] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *Proc. IEEE Conf. Decis. Control*, Atlanta, GA, USA, Dec. 2010, pp. 2572–2577.
[45] M. Brooks, The Matrix Reference Manual, 2011. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html.
[46] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," in *Optimization Techniques: Lecture Notes in Control and Information Sciences*, vol. 7, J. Stoer, Eds., Berlin, Heidelberg: Springer, 1978, pp. 234–243.
[47] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Res. Logist. Quart.*, vol. 3, no. 1/2, pp. 95–110, Mar. 1956.
[48] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. 427–435.
[49] A. Alkhateeb, O. El Ayach, G. Leus, and R. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[50] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.
[51] S. Jegelka and J. Bilmes, "Submodularity beyond submodular energies: Coupling edges in graph cuts," in *Proc. IEEE CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 1897–1904.

Aritra Konar (M'17) received the B.Tech. degree in electronics and communications engineering (ECE) from West Bengal University of Technology, Kolkata, India, and M.S. and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2011, 2014, and 2017, respectively.

He is currently a Postdoctoral Associate with the Department of ECE, University of Virginia, Charlottesville, VA, USA. His research interests include statistical signal processing, wireless communications, nonlinear optimization, and data analytics.

Nicholas D. Sidiropoulos (F'09) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, and M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1988, 1990, and 1992, respectively.

He was an Assistant Professor with the University of Virginia, an Associate Professor with the University of Minnesota, a Professor with TU Crete and the University of Minnesota, and is currently a Professor and Chair with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, USA. His research spans topics in signal processing theory and algorithms, optimization, communications, and factor analysis—with a long-term interest in tensor decomposition and its applications. His current research is focused on signal and tensor analytics for learning from big data.

Dr. Sidiropoulos was the recipient of the NSF/CAREER award in 1998, and the IEEE Signal Processing (SP) Society Best Paper Award in 2001, 2007, and 2011. He was an IEEE SP Society Distinguished Lecturer (2008–2009), and a Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007–2008). He was the recipient of the 2010 IEEE SP Society Meritorious Service Award, and the 2013 Distinguished Alumni Award from the Department of ECE, University of Maryland. He is a Fellow of EURASIP and is currently serving as Vice President-Membership of the IEEE SP Society.