Joint User and Antenna Selection in Massive-MIMO Systems With QoS-Constraints

Javed Akhtar, *Student Member, IEEE*, Ketan Rajawat, *Member, IEEE*, Vipul Gupta, and Ajit K. Chaturvedi, *Senior Member, IEEE*

***Abstract*—This article considers a downlink transmission in a multiuser massive multi-input–multi-output (massive-MIMO) system in both the large-scale and small-scale fading environments. While the base station has many antennas, the associated circuit costs require some antennas to be turned OFF, thus necessitating antenna/RF-chain selection. Furthermore, the users have quality-of-service (QoS) constraints and cannot be served within the specified power budget and also with arbitrarily low signal-to-interference-plus-noise ratios. Instead, the base stations seek to schedule only a subset of users for which the QoS-constraints can be met, while the remaining users must be scheduled in other time-frequency slots. The problem is considered under both long and short time-scales. The longer time-scale problem entails solving for the power, user allocation, and number of active antennas to be used over multiple coherence intervals. While the need to select the optimum set of users makes the problem combinatorial, a low-complexity algorithm is proposed that allows us to solve it in polynomial time. For the more computationally intensive per-coherence interval problem, we employ a convex relaxation-based approach in order to obtain an approximate solution. Detailed simulations are carried out to establish that the proposed algorithms outperform standard greedy approaches and are close to optimal for several settings.**

***Index Terms*—Antenna selection, block-coordinate descent (BCD), majorization–minimization (MM), massive multi-input multi-output (MIMO), quality-of-service (QoS), user scheduling.**

NOMENCLATURE

$\mathcal{A}^0$	Set of available antennas s.t. $\mid \mathcal{A}^0 \mid = M$.
$\mathcal{U}^0$	Set of available users s.t. $\mid \mathcal{U}^0 \mid = L$.
$\mathcal{U}$	Set of users scheduled s.t. $\mid \mathcal{U} \mid = K \leq L$.
$\mathcal{A}$	Set of transmit antennas s.t. $\mid \mathcal{A} \mid = N \leq M$.

I. INTRODUCTION

WITH the rapid proliferation of wireless devices and social media services, the data traffic has grown multifold. The next generation of cellular systems seek to meet the data challenge and provide data-centric services through the use of disruptive technologies such as massive multi-input–multi-output (massive-MIMO) [1], [2]. Massive-MIMO is an emerging 5th-generation technology where a base station (BS) equipped with a large antenna array communicates simultaneously with multiple users [3]. The large number of antennas installed at the BS can help to increase the system capacity, resulting in a large number of degrees of freedom [4]. The usefulness of the massive-MIMO system can be harnessed through favorable mode of propagation, where the interuser channels become asymptotically orthogonal to each other [5], [6]. This asymptotic behavior of the channel allows users to use the same time-frequency resource resulting in several fold increase of network spectral efficiency [7], [8].

While promising in theory, high device cost and energy consumption ultimately translate into practical constraints on the number of radio frequency (RF) chains (also known as the analog front-ends) that can be installed with an antenna [9]. Toward maximizing the sum-rate, it is desired to opportunistically select a subset of antennas with good channel conditions corresponding to the number of RF-chains installed [10], [11]. Apart from the implementation cost, a massive-MIMO system would require huge overheads (order of transmit antennas) to acquire channel state information (CSI), especially for multiuser scenario with frequency-division duplex systems [12]. However, the overhead requirements for a multiuser massive-MIMO systems can be reduced by invoking the hidden sparsity of the channel matrix but at a cost of channel estimation quality [13]. Moreover, if the channel matrix is rank deficient and equal power is allocated among the transmit antennas, then optimally selecting fewer transmit antennas was shown to increase the system capacity [14].

In cellular systems, supporting data services such as video chat, gaming, and other interactive applications, the quality-of-service (QoS) becomes a critical performance metric that shapes the user experience [15], [16]. Such QoS guarantees may be imposed as constraints on the minimum signal-to-interference-and-noise ratio (SINR) or minimum rate. Indeed, in resource-constrained multiuser systems, it becomes imperative for the BS to schedule only those users whose QoS-constraints can be met within the current coherence interval. The remaining users are either scheduled on a different band or in subsequent coherence intervals [17]. For instance, even if the power budget at the BS allows all users to be served, the data rate at some of the users may be too low to be of any use to them. Instead it may be prudent to divide the power budget among only the high-QoS users while scheduling the remaining users in later time slots. Observe that such an approach sacrifices fairness in favor of QoS. Indeed, the QoS-constraint must be met for all scheduled

Manuscript received September 22, 2019; revised March 16, 2020, May 30, 2020, and July 18, 2020; accepted July 22, 2020. Date of publication August 20, 2020; date of current version March 9, 2021. This work was supported by the Department of Science and Technology India under Grant EMR/2016/005959. *(Corresponding author: Javed Akhtar.)*

Javed Akhtar is with the Radisys India Pvt. Ltd., Bengaluru 560103, India (e-mail: javeda2309@gmail.com).

Ketan Rajawat and Ajit K. Chaturvedi are with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India (e-mail: ketan@iitk.ac.in; akc@iitk.ac.in).

Vipul Gupta is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA (e-mail: vipul_gupta@berkeley.edu).

Digital Object Identifier 10.1109/JSYST.2020.3014867

users, and consequently, it becomes necessary to the tradeoff performance and fairness for reliability. In other words, meeting the QoS-constraints will generally lead to lower sum-rate and unfair allocation of resources to the users.

This article considers the problem of joint user scheduling and antenna selection that yields the maximum downlink utility (sum-rate minus penalty) in massive-MIMO systems under a strict QoS-requirement. Efficient resource utilization is of particular importance in small/microcell settings that are of particular focus today. We attack the problem at two different time-scales with the choice of the solution depending on the availability of computational resources at the BS. The longer time-scale problem builds upon the antenna selection framework introduced in [18], and entails solving the problem only once every few coherence intervals. Although combinatorial in nature, a polynomial time algorithm is developed that yields the exact solution. Furthermore, it is shown that the approximate solution can be obtained using a dual scheme that is significantly less complex. It is shown that in general, having to schedule all the users is not only inefficient but also costly in terms of requiring more antennas.

Subsequently, the more complicated per-coherence interval problem of joint power allocation, antenna selection, and user selection is considered. The resulting problem is also combinatorial and difficult to solve. We put forth a convex relaxation-based approximation that is shown to outperform standard greedy approaches in the literature and performs near-optimally when the power budget is not too low. Detailed simulations are carried out to establish the near-optimality of the proposed algorithms. The key contributions of this article include the following: a) development of the QoS-constrained joint power allocation, antenna, and user selection problem; b) development of a polynomial time solution to the combinatorial long time-scale problem; and c) development of two convex relaxation-based approximate algorithms for the combinatorial short time-scale problem.

A. Related Works

This section briefly reviews the related works on antenna selection and/or user selection for large MIMO systems. For multiantenna systems, the problem of antenna selection is well-studied [19]–[23]. Of these, most transmit and receive side antenna selection schemes are greedy in nature [24]. Nongreedy approaches include [25] for nonprecoded MIMO downlink channel, where the simplification was obtained by first assuming uniform power allocation, obtaining the optimal set of antennas, and subsequently formulating a simplified convex optimization problem to obtain the optimal power allocation. A similar framework for a large distributed antenna system was considered in [26] under the large-scale fading scenario. Toward, achieving a minimum QoS at each user, authors in [18] have studied the problem of selecting the number of antennas at the base station so as to minimize the total transmit power. It was shown in [27] that for systems with nonnegligible circuit power consumption, antenna selection resulted in higher energy efficiency.

In the context of user scheduling, it was shown in [17], that a zero-forcing (ZF) beamforming MIMO system achieves the MIMO broadcast capacity in the limit of infinite number of users. As a consequence, a greedy user selection scheme based on selecting users with semiorthogonal channels was proposed. Extending [17], a greedy approach toward joint antenna selection and user scheduling was proposed for the downlink of ZF-based massive-MIMO systems in [28]. For the uplink case, a modified greedy scheme was proposed in [29] resulting in slight improvement. It is remarked that unlike this article, none of these schemes have considered QoS and circuit power constraints in tandem. An exception is the joint antenna selection and QoS-constrained user scheduling algorithm proposed in [30] that used a greedy scheme and will subsequently be compared with the proposed algorithm.

Finally, this article is also related to the joint antenna selection and power allocation problem that arises in the context of distributed MIMO systems [31], [32]. While similar greedy approaches have been used in the distributed case, the specific challenges in the present setting are very different. For instance, the capacity limitation inherent to the backhaul communications in distributed MIMO systems makes the two problems quite different. For the short-time scale scenario, a rate maximization problem under the joint antenna and user selection problem was considered with the QoS-constraint in [33]. A maximal ratio transmitter based precoder with fixed number of antennas was considered, and the problem was solved using second-order cone programming. A conference version of this article has been presented at [34], where only the rate as an utility function under the long time-scale problem is considered. This article contains the more general and complete results in terms of the utility function under the short and long time-scale problem, primarily focusing on the small/microcell implementation in an urban scenario.

The rest of this article is organized as follows. In Section II, we describe the system model where we formulate the joint antenna selection and user scheduling (JASUS) problem under both the large-scale and small-scale fading. Section III details the algorithms toward obtaining the joint antenna-user pair under large-scale fading. Section IV details the joint optimal selection under small-scale fading for the cases when optimal antenna subset is fixed (i.e., number of RF-chains is known) and when the number of RF-chains is unknown. Finally the simulation results are presented in Section V and the conclusions are presented in Section VI.

Before proceeding to the system model and background, some notations are introduced. In general, scalar variables are denoted by lower case letters, bold lower case letters for denoting vectors while bold upper case letters for denoting matrices. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$. The trace, conjugate, transpose, Hermitian transpose, and pseudo-inverse of a matrix $\mathbf{A}$ are denoted by tr (A), $\mathbf{A}^\star$, $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathbf{A}^\dagger$, respectively. The space of all real and complex matrices of size $m \times n$ is denoted by $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$, respectively. (i, j)th entry of a matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{ij}$ and the diagonal matrix $\mathbf{A}$ with entries of $\mathbf{a} = [a_1, \ldots, a_L]^T$ on the main diagonal is denoted by $\text{diag}(a_1, \ldots, a_L)$ or $\text{diag}(\mathbf{a})$. $\|\mathbf{a}\|_0$ denotes the number of nonzero elements in the vector $\mathbf{a}$. Given a set of indices $\mathcal{X}$ and $\mathcal{Y}$, $\mathbf{A}_{\mathcal{X}\mathcal{Y}}$ denotes the matrix formed by the corresponding rows and columns of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Likewise, $\mathbf{A}_{\mathcal{X}}$ denotes the matrix formed by the rows corresponding to $\mathcal{X}$ but containing all the columns, while $\mathbf{A}_{\bullet\mathcal{Y}}$ denotes the matrix formed by the columns corresponding to $\mathcal{Y}$ but containing all the rows.

II. System Model

Consider a single cell (small/micro) system ([35, Ch. 3]) with M antennas at the BS, serving L single-antenna users [27], [36]. We consider a massive-MIMO setting, allowing M to be very

large [3], [8] and the effects of pilot contamination and interference from other cells are assumed negligible. Furthermore, the channel gains associated with the antennas are assumed to be uncorrelated. The system operates in time-division duplex mode, and each coherence interval comprises of both uplink and downlink data transmission phases. Let τ denote the length of the coherence interval, measured in number of samples. The uplink and downlink data transmission phases are of durations τ_u and τ_d symbols, respectively. Additionally, $\tau_p := (\tau - \tau_u - \tau_d)$ symbols are set aside for the BS to learn the channel gains. The training phase entails users transmitting their pilot sequences to the BS in their respective physical uplink resource blocks. Since, the pilot sequences are already known, the BS utilizes the received signals from all the users in order to estimate the required channel gains.

An urban environment with the small/microcell setting under a normal user-traffic scenario is considered wherein $L \ll M$, and, consequently, the BS seek to maximize its utility by multiplexing the downlink transmit signal across a subset $\mathcal{A} \subseteq \mathcal{A}^0 := \{1, \ldots, M\}$ of antennas such that $| \mathcal{A} | = N \leq M$. The intuition here is that while the sum-rate increases with N, the increased circuit power consumption (fixed RF power per antenna denoted by P_{RF}) may offset the gains when $N \gg L$ [28], [30]. Specifically, each active antenna consumes P_{RF} power and incurs an operating cost of c units. Furthermore, an upper limit on N may also be imposed if the number of RF-chains installed at the BS are fewer than M.

Let $\mathbf{H} \in \mathbb{C}^{L \times M}$ denote the small-scale fading matrix between the M antennas and the L users. The entries of $\mathbf{H}$ are independent, circularly symmetric complex Gaussian distributed, i.e., $h_{ij} := [\mathbf{H}]_{ij} \sim \mathcal{CN}(0, 1)$. The large-scale channel gain between the BS and the kth user is denoted by β_k. Defining $\mathbf{G} := \mathbf{H}_{\bullet \mathcal{A}}$ and letting $\mathbf{g}_k^T$ denote the kth row of $\mathbf{G}$ such that $\mathbf{G} = [\mathbf{g}_1\, \mathbf{g}_2, \ldots, \mathbf{g}_L]^T$, the signal received at the kth user is [37]

$$y_k = \sqrt{\beta_k}\mathbf{g}_k^T\mathbf{x} + n_k, \qquad k = 1, \ldots, L \tag{1}$$

where $\mathbf{x}$ is the transmitted signal vector, and $n_k \in \mathbb{C}$ is the complex additive white Gaussian noise with zero mean and variance σ^2. Without loss of generality and for the sake of brevity, we set $\sigma^2 = 1$ for the rest of this article.

When data are being transmitted to L users, let $u_k \in \mathbb{C}$ denote the unit magnitude data symbol for the kth user and let $\mathbf{u} \in \mathbb{C}^{L \times 1}$ be the vector that collects these symbols for all L users. The kth user is allocated the transmit power $p_k \in [0, P_{\max}]$, which is collected into the vector $\mathbf{p} \in \mathbb{R}^{L \times 1}$. The transmit signal $\mathbf{x}$ at the BS is the scaled and precoded version of $\mathbf{u}$, and is given by $\mathbf{x} := \mathbf{V}\sqrt{\text{diag}(\mathbf{p})}\mathbf{u}$, where $\mathbf{V} \in \mathbb{C}^{N \times L}$ is the precoding matrix. This article considers the case of ZF-based-precoding that is known to be near-optimal for $M \gg L$ and high SINR [35].

The users are QoS-constrained and cannot decode the message if the received signals have very low powers. The QoS-constraint at the k-th user can be written as an SINR constraint of the form $\text{SINR}(k) \geq \gamma$, where $\gamma > 0$ is the SINR threshold. Since the transmit power budget at the BS is limited to at most $P_{\max}$, only a subset of the users $\mathcal{U} \subseteq \mathcal{U}^0 := \{1, \ldots, L\}$ that meet the QoS-constraints are actually allocated nonzero power. In other words, $p_k = 0$ for all $k \notin \mathcal{U}$, and the total power allocated to the user data is given by $\mathbf{1}^T\mathbf{p} = \sum_{k=1}^{L} p_k = \sum_{k \in \mathcal{U}} p_k$. Since, the goal here is to maximize the utility, the QoS-constraint makes the resource allocation unfair to users with poor reception. Moreover, allocating small powers to users with poor reception is wasteful or suboptimal from the network throughput point of view, and the available power budget is better utilized toward boosting the SINR of users with high channel gains. In practice, the users not in $\mathcal{U}$ will be scheduled into a different channel (i.e., in a different time-frequency resource block) or will be associated with a different BS.

Henceforth, the cardinality of the set $\mathcal{U}$ is denoted by $K := |\mathcal{U}|$. Note that when only K users are scheduled, the $N \times 1$ transmit signal is given by $\mathbf{x} = \mathbf{V}\sqrt{\text{diag}(\mathbf{p})}\mathbf{u}$, where $\mathbf{u} \in \mathbb{C}^{K \times 1}$, $\mathbf{p} \in \mathbb{R}^{K \times 1}$, and $\mathbf{V} \in \mathbb{C}^{N \times K}$. The list of parameters introduced thus far is provided in the Nomenclature.

The goal of this article is to solve the following utility maximization problem:

$$\max_{\mathbf{p}, \mathcal{A}, \mathcal{U}} U(\mathbf{p}, \mathcal{A}, \mathcal{U}) := \sum_{k \in \mathcal{U}} \log\left[1 + \text{SINR}(k)\right] - c|\mathcal{A}| \quad (\mathcal{P})$$

$$\text{s.t.} \sum_{k \in \mathcal{U}} p_k + NP_{\text{RF}} \leq P_T, \quad \text{SINR}(k) \geq \gamma \qquad \forall k \in \mathcal{U}$$

where c is the per antenna cost relative to the sum-rate (measured in bits/second/hertz) achieved, and P_T is the total power budget that includes the power consumed by the circuit as well as the transmission power [38]. Note that the utility function defined in $(\mathcal{P})$ is adimensional. The per antenna cost c quantifies the combined effects of circuit power consumption, cooling costs, and miscellaneous operating costs such as regular maintenance and energy management [39]. Additionally, if the BS is powered by renewable energy or batteries, the cost can be made to vary over time depending on the current energy availability, thereby affording further operating flexibility to the system [40]–[42]. The problem at hand is inherently multiobjective: the goal is to maximize the sum-rate while also minimizing the antenna operation costs. Additionally, the problem includes constraints on total power consumption and minimum received SINR. For such problems, the formulation in $(\mathcal{P})$ is not unique, and other related formulations can also be considered. For instance, the objective could be the maximization of energy efficiency while imposing restrictions on SINR and number of antennas [27]. Alternate goals include minimization of power consumed per antenna, worst case SINR maximization, antenna efficiency, etc, with appropriate constraints. While all of these formulations are related, they are generally not equivalent, but merely represent different approaches to selecting an operating point. The rate-minus-penalty objective function has been widely considered in the context of network utility maximization, and is selected here for ease of analysis. The SINR at the kth user also depends on $\mathbf{p}$, $\mathcal{A}$, and $\mathcal{U}$, and its exact form will be detailed in subsequent subsections. Observe, however, that regardless of the form of the SINR, solving $(\mathcal{P})$ is not straightforward as it involves set-valued variables $\mathcal{U}$ and $\mathcal{A}$. Note that the variables $\mathbf{p}$ and $\mathcal{U}$ are related; the nonzero entries of $\mathbf{p}$ correspond to the elements of $\mathcal{U}$. Both variables are, however, used for ease of exposition.

The problem $(\mathcal{P})$ can be solved on two different timescales depending on the computational power available at the BS. If the system is computationally limited, it is recommended to solve $(\mathcal{P})$ once every few coherence intervals. In each such "super-slot" or block, the optimum user and antenna sets are selected and an average power budget is assigned to all the selected users. At each coherence interval, the BS utilizes the precoder $\mathbf{V} = \sqrt{N - K}\mathbf{F}^H(\mathbf{F}\mathbf{F}^H)^{-1}$, where $\mathbf{F} := \mathbf{H}_{\mathcal{U}\mathcal{A}}$. The normalization ensures that columns of $\mathbf{V}$ are unit normed on an average, i.e., $\mathbb{E}_{\mathbf{H}}[||\mathbf{v}_k||_2^2] = 1$, where $\mathbf{v}_k$ denotes kth column of $\mathbf{V}$. Given sufficiently large number of coherence intervals per super-slot,

such a precoder ensures that the kth user is allocated an average power of p_k per coherence interval. The long timescale problem can be viewed as that of allocating resources over the entire block of channel gains.

On the other extreme, the full problem can also be solved at every coherence interval. That is, the optimum set of antennas and users is selected at every coherence interval, and optimal power is allocated to all the selected users. In this case, the BS transmits using the precoder $\mathbf{V} = \mathbf{F}^H(\mathbf{F}\mathbf{F}^H)^{-1}\mathbf{\Lambda}^{1/2}$, where the diagonal scaling matrix $\mathbf{\Lambda}$ ensures columns of $\mathbf{V}$ are unit normed for each coherence interval.

Furthermore, from Sec. G.2.2 of [3GPP TS 38.141], it can be seen that the coherence interval is between 2.5 and 200 ms, depending on the mobility of the user equipment (UE). Indeed, the UE mobility directly impacts the viability of the per-coherence time interval approach. On the other hand, allocation of resources every few coherence intervals is always possible regardless of the UE mobility. Hence, solving the instantaneous user and antenna selection problem is clearly more computationally intensive and requires perfect channel knowledge, but generally yields a higher overall utility. The more challenging case of imperfect CSI at the receiver is not considered here, and is left as an open problem. Subsequently, we detail each of the two settings and formulate the corresponding utility maximization problems.

A. Long Timescale: Block Resource Allocation

The system model for the long timescale problem builds upon the model from [18] where the large-scale fading coefficients $\{\beta_k\}_{k\in\mathcal{U}}$ are perfectly known while the small-scale fading matrix $\mathbf{F}$ is estimated imperfectly using τ_p pilots. Using the MMSE estimator for h_{ij}s along with the ZF-precoder, the effective SINR at the user k is given by [35]

$$\text{SINR}_L(k) = \frac{(N-K)\alpha_k p_k}{1+(\beta_k-\alpha_k)\sum_{k'\in\mathcal{U}} p_{k'}} \qquad k\in\mathcal{U} \tag{2}$$

where $K = |\mathcal{U}|$, $\alpha_k =: \frac{\tau_p\rho_{ul}\beta_k^2}{1+\tau_p\rho_{ul}\beta_k}$, and ρ_{ul} denote the uplink power used for pilot transmission. Note that SINR_L denotes the SINR under the long-timescale scenario. The expression in (2) is valid for the case when $N \gg K$. From the expression for the SINR in (2), we observe that (a) the SINR and consequently the problem $(\mathcal{P})$ depends only on $\{\beta_k\}$ and not on $\mathbf{H}$; and (b) the different antennas at the BS become statistically identical allowing us to solve for N instead of the set $\mathcal{A}$ [18]. Without loss of generality, we will take $\mathcal{A} = \mathcal{N} := \{1,\dots,N\}$ as the set of antennas selected in this case. In summary, the long timescale version of $(\mathcal{P})$ can be written as

$$\begin{aligned}
&\max_{\mathbf{p},N,\mathcal{U}} \sum_{k\in\mathcal{U}} \log\left[1+\text{SINR}_L(k)\right] - cN \quad (\mathcal{P}_L)\\
&\text{s.t.} \sum_{k\in\mathcal{U}} p_k + NP_{\text{RF}} \le P_T\\
&\qquad \text{SINR}_L(k) \ge \gamma \quad \forall k\in\mathcal{U}, \quad K\le N\le M.
\end{aligned}$$

Solving $(\mathcal{P}_L)$ is challenging due to the set-valued optimization variables $\mathcal{U}$. Indeed, without user selection, the problem is relatively easier and admits a closed-form solution [18]. Nevertheless, the algorithms from [18] cannot be directly applied to solve $(\mathcal{P}_L)$ due to the additional QoS-constraint.

B. Short Timescale: Instantaneous Resource Allocation

The goal here is to optimally allocate resources at every coherence interval. Without loss of generality, we merge the large-scale and small-scale fading gains into a consolidated channel gain matrix $\mathbf{H}\in\mathbb{C}^{L\times M}$, that is assumed to be perfectly known at the BS [27], [36]. At each coherence interval, the BS utilizes antennas belonging to the set $\mathcal{A}\subset\mathcal{A}_0$. Recalling, that $\mathbf{G} := \mathbf{H}_{\bullet\mathcal{A}}$ and $\mathbf{g}_k^T$ denotes the kth row of $\mathbf{G}$, the instantaneous SINR at the kth user is given by [43]

$$\text{SINR}_S(k) = \left(p_k|\mathbf{g}_k^H\mathbf{v}_k|^2\right) \Big/ \left(\sum_{i\neq k} p_i|\mathbf{g}_k^H\mathbf{v}_i|^2 + 1\right) \tag{3}$$

where $\mathbf{v}_k$ denotes the kth column vector of the ZF-precoder matrix $\mathbf{V}$. Note that SINR_S denotes the SINR under the short-timescale scenario. Since, the ZF-precoder is designed to cancel the interuser interference term appearing in the denominator of (3), the simplified expression for SINR becomes $\text{SINR}_S(k) = p_k|\mathbf{g}_k^H\mathbf{v}_k|^2$ [44]. Furthermore, the ZF-precoder is given by $\mathbf{V} = \mathbf{F}^H(\mathbf{F}\mathbf{F}^H)^{-1}\mathbf{\Lambda}^{1/2}$, where the diagonal normalization matrix $\mathbf{\Lambda}\in\mathbb{R}_{++}^{K\times K}$ ensures that the columns of $\mathbf{V}$ are unit normed and its (k,k)th entry is given by

$$[\mathbf{\Lambda}]_{kk} = \frac{1}{\left[(\mathbf{F}\mathbf{F}^H)^{-1}\right]_{kk}} =: \frac{1}{\left[(\mathbf{H}_{\mathcal{U}}\mathbf{\Delta}\mathbf{H}_{\mathcal{U}}^H)^{-1}\right]_{kk}} \tag{4}$$

for $k\in\mathcal{U}$ and $\mathbf{\Delta} := \text{diag}(\Delta_1,\dots,\Delta_M)$ is a diagonal antenna selection matrix with 0-1 entries. Specifically, $\Delta_i = 1$ if $i\in\mathcal{A}$ and zero otherwise. Therefore, the expression for the instantaneous SINR can be written as

$$\text{SINR}_S(k) = \frac{p_k}{\left[(\mathbf{H}_{\mathcal{U}}\mathbf{\Delta}\mathbf{H}_{\mathcal{U}}^H)^{-1}\right]_{kk}}, \qquad k\in\mathcal{U}. \tag{5}$$

In summary, the per-coherence interval resource allocation problem can be written as

$$\begin{aligned}
&\max_{\mathbf{p},\mathbf{\Delta},\mathcal{U}} \sum_{k\in\mathcal{U}} \log\left[1+\text{SINR}_S(k)\right] - c\,\text{tr}(\mathbf{\Delta}) \quad (\mathcal{P}_S)\\
&\text{s.t.} \sum_{k\in\mathcal{U}} p_k + \text{tr}(\mathbf{\Delta})P_{\text{RF}} \le P_T, \;\; \text{SINR}_S(k) \ge \gamma \quad \forall k\in\mathcal{U}\\
&\qquad K \le \text{tr}(\mathbf{\Delta}) \le M, \quad \Delta_i\in\{0,1\} \quad \forall i\in\mathcal{A}^0.
\end{aligned}$$

Observe that as compared to $(\mathcal{P}_L)$, the problem in $(\mathcal{P}_S)$ is further complicated due to the presence of the integer-valued optimization variable $\mathbf{\Delta}$.

III. Block Resource Allocation Problem

This section details the solution of block resource allocation problem. Despite being seemingly combinatorial, we will show that $(\mathcal{P}_L)$ can be exactly solved with complexity $\mathcal{O}(ML^2)$. Toward this end, we begin with first discussing the solution of the problem for a fixed given value of N.

A. Joint User Selection and Power Allocation for a Given N

For a fixed N, the second term in the objective function of $(\mathcal{P}_L)$ can be dropped, and the equivalent utility maximization problem becomes

$$\begin{aligned}
&\max_{\mathbf{p},\mathcal{U}} \sum_{k\in\mathcal{U}} \log\left[1+\text{SINR}_L(k)\right]\\
&\text{s.t.} \sum_{k\in\mathcal{U}} p_k \le P_{\max}, \quad \text{SINR}_L(k) \ge \gamma \quad \forall k\in\mathcal{U}
\end{aligned} \tag{6}$$

Algorithm 1: User Scheduling for Fixed N.

Input: $K_{\max},\ N,\ \gamma,\ \{\psi_k\}_{k=1}^{K_{\max}}$, and $P_{\max}$;
for $K = 1$ *to* $K_{\max}$ **do**
  Find $\breve{\lambda}$: $\sum_{k=1}^{K} \max\left\{\left(\frac{1}{\lambda} - \frac{1}{\psi_k(N-K)}\right), \frac{\gamma}{\psi_k(N-K)}\right\} = P_{\max}$;
  if *infeasible* **then**
    Break ;
  end
  Allocate $\breve{p}_k = \max\left\{\left(\frac{1}{\lambda} - \frac{1}{\psi_k(N-K)}\right), \frac{\gamma}{\psi_k(N-K)}\right\}$ for $1 \le k \le K$;
  Sum Rate$(K) := \sum_{k=1}^{K} \log\left[1 + (N-K)\psi_k \breve{p}_k\right]$;
end
Select $K^\star(N) := \arg\max_K$ *Sum Rate* ;
Output: User set $\mathcal{U}^\star(N) = \{1, \ldots, K^\star(N)\}$ and corresponding $\mathbf{p}^\star(N)$.

where $P_{\max} := P_T - NP_{\text{RF}}$. In order to avoid performing an exhaustive search over the user set $\mathcal{U}$, we establish the following result.

Lemma 1: The solution $(\mathbf{p}^\star, \mathcal{U}^\star)$ of (6) satisfies $\sum_{k\in\mathcal{U}^\star} p_k^\star = P_{\max}$.

Proof: Refer Appendix A. ■

It follows from *Lemma* 1 that the constraint $\sum_k p_k \le P_{\max}$ in (6) can be replaced with the constraint $\sum_k p_k = P_{\max}$ without loss of optimality. Further, for the optimal power allocation $\mathbf{p}^\star$, we have that $\text{SINR}_L(k) = \frac{(N-K)\alpha_k p_k^\star}{1+(\beta_k-\alpha_k)P_{\max}}$. Alternatively, defining $\psi_k := \frac{\alpha_k}{1+(\beta_k-\alpha_k)P_{\max}}$ for $1 \le k \le L$, the problem in (6) can equivalently be written as

$$\max_{\mathbf{p},\mathcal{U}} \sum_{k\in\mathcal{U}} \log\left[1 + (N-K)\psi_k p_k\right]$$
$$\text{s.t.} \sum_{k\in\mathcal{U}} p_k \le P_{\max}, \quad (N-K)\psi_k p_k \ge \gamma \qquad \forall k \in \mathcal{U}. \quad (7)$$

While the optimization variables in (7) are still set-valued, the following lemma paves the way for solving it efficiently. Observe that for a given set $\mathcal{U}$, the power allocation problem in (7) is convex and can be solved via waterfilling. Without loss of generality, let $\psi_1 > \psi_2 > \cdots > \psi_L$, and define $K_{\max} := \max\{K \mid \frac{\gamma}{N-K}\sum_{k=1}^{K}\frac{1}{\psi_k} \le P_{\max}\}$. Then, we have the following result.

Lemma 2: For a given $K \le K_{\max}$, set $\mathcal{U} = \{1, 2, \ldots, K\}$ yields the maximum sum-rate for (7).

Proof: Refer Appendix B. ■

The optimal user set $\mathcal{U}$ can be found by carrying out a line search on $K = 1, \ldots, K_{\max}$. Thanks to *Lemma* 2, for each value of K it is no longer required to perform an exhaustive search over $\binom{L}{K}$ possible user combinations. Instead, the problem in (7) is solved with $\mathcal{U} := \{1, \ldots, K\}$ for each $1 \le K \le K_{\max}$. As summarized in Algorithm 1, the optimal power allocation can be found via a waterfilling step that entails solving a nonlinear equation. For instance, the use of the bisection algorithm in Step 2 yields an ϵ-optimal value of $\breve{\lambda}$ in $\mathcal{O}(-\log(\epsilon))$ iterations. In other words, for a given tolerance, the overall runtime of Algorithm 1 is $\mathcal{O}(NL^2)$ in the worst case.

It is remarked that in practice, if $K_{\max} < L$, the runtime of the algorithm may be much less. Furthermore, if the power allocation problem is infeasible for a specific value of K', it is not possible to add another user $k \ge K' + 1$ with power $p_k \ge \gamma/(N-K)\psi_k$.

1) Approximate User-Scheduling via Dual-Relaxation: Algorithm 1 necessitates a line search over K from 1 to $K_{\max}$. The complexity of such a search may be reduced by properly initializing K. Toward this end, observe that (7) can equivalently be written as the following problem in $\mathbf{p} \in \mathbb{R}^L$:

$$\max_{\mathbf{p}} \sum_{k=1}^{L} \log\left[1 + (N - \|\mathbf{p}\|_0)\psi_k p_k\right]$$
$$\text{s.t.} \sum_{k=1}^{L} p_k = P_{\max}$$
$$(N - \|\mathbf{p}\|_0)p_k \in \{0\} \cup [\gamma/\psi_k, P_{\max}] \qquad \forall 1 \le k \le L. \quad (8)$$

At this stage, we make some simplifying approximations. Assuming that $N \gg L$, we first replace $\|\mathbf{p}\|_0$ with L. Denoting, $w_k := \frac{\gamma}{\psi_k(N-L)}$ (minimum power required by user k) and $\mathcal{P}_k : \{0\} \cup [w_k, P_{\max}]$, the dual function for (8) becomes

$$\varrho(\nu) = \nu P_{\max} + \sum_{k=1}^{L} \max_{p_k\in\mathcal{P}_k} \log\left(1 + \frac{\gamma p_k}{w_k}\right) - \nu p_k. \quad (9)$$

Although $\mathcal{P}_k$ is a nonconvex set, the dual function can be readily found. Denoting $p_k(\nu) := \arg\max_{p_k} \log(1 + \frac{\gamma p_k}{w_k}) - \nu p_k$ for $k = 1, \ldots, L$, it can be seen that

$$p_k(\nu) = \begin{cases} \frac{1}{\nu} - \frac{w_k}{\gamma} & \nu = \frac{\gamma}{w_k(1+\gamma)} \\ w_k & \frac{\log(1+\gamma)}{w_k} > \nu > \frac{\gamma}{w_k(1+\gamma)} \\ 0 & \nu > \frac{\log(1+\gamma)}{w_k}. \end{cases} \quad (10)$$

Since, $p_k(\nu)$ is discontinuous function of ν, the dual function $\varrho(\nu)$ would also be a discontinuous nonincreasing function of ν. Furthermore, since the original problem is nonconvex, strong duality does not hold and consequently, the value of ν that minimizes $\varrho(\nu)$ would not necessarily correspond to a feasible allocation. Instead, we settle for a feasible power allocation that corresponds to the smallest ν satisfying $\sum_k p_k(\nu) \le P_{\max}$. Such a ν may be found by bisection, and for sufficiently large ν, $p_k = 0\ \forall k$. Observe further that users scheduled by the dual algorithm will also be in decreasing order of their channel gains ψ_k. That is, if the dual algorithm schedules K_d users, they would be $\mathcal{U} = \{1, \ldots, K_d\}$. Note, however, that the power allocation obtained from (10) is suboptimal as it corresponds to the dual of the approximate problem where $\|\mathbf{p}\|_0$ was replaced with L in (8).

In order to achieve the near-optimal performance, a slightly more expensive version is recommended. Let K_d be the output of the dual algorithm scheme when all the L users are considered in (10). The idea is to recalculate the powers for first K_d and $K_d + 1$ users, and select the solution corresponding to the higher sum rate. The approach is summarized in Algorithm 2 and entails carrying out only two more bisection searches as opposed to the $K_{\max}$ searches required in Algorithm 1. The worst case runtime of Algorithm 2 is therefore $\mathcal{O}(NL)$.

Finally it is also possible to use Algorithm 2 for initializing Algorithm 1. Specifically, a set of users $K_D^\star$ is obtained first using Algorithm 2. Subsequently, a line search is carried out as in Algorithm 1 starting at $K_D^\star$ and searching over $K_D^\star \pm 1$, $K_D^\star \pm 2$, and so on. We will show in Section V that such a hybrid approach may yield the optimum power allocation while also incurring less than $K_{\max}$ bisection searches.

Algorithm 2: Dual-Relaxation Approach.

Input: $L,\ \gamma,\ \{w_k\}_{k=1}^{L},\ \{\psi_k\}_{k=1}^{L}$, and $P_{\max}$;
Let, $\mathcal{U}_{d_1}$ be the set of users scheduled by solving (10) to obtain $\tilde{\mathbf{p}} = \tilde{\mathbf{p}}^1$ such that $|\ \mathcal{U}_{d_1}\ | = K_d$;
if $\sum_{k\in\mathcal{U}_{d_1}} \tilde{p}_k^1 = P_{\max}$ **then**
  $SR_1 = \sum_{k\in\mathcal{U}_{d_1}} \log\left[1+(N-K_d)\psi_k\tilde{p}_k^1\right]$. ;
else
  if $\left(\sum_{k\in\mathcal{U}_{d_1}} \tilde{p}_k^1 < P_{\max}\right)$ *and* $\left(P_{\max} \geq \sum_{k=1}^{K_d+1} w_k\right)$ **then**
    Allocate $\tilde{p}_{K_d+1}^2 = w_{K_d+1}$ and
    $\tilde{p}_\ell^2 = 0$ for users $\ell = (K_d+2),\ldots,L$;
    Solve (10) to obtain $\tilde{\mathbf{p}}^2$ for users $\ell = 1,\ldots,K_d$ with power budget $(P_{\max} - w_{K_d+1})$;
    Define $\mathcal{U}_{d_2} := \mathcal{U}_{d_1} \bigcup (K_d+1)$;
    $SR_2 = \sum_{k\in\mathcal{U}_{d_2}} \log\left[1+(N-K_d-1)\psi_k\tilde{p}_k^2\right]$.
  end
  if $\left(\sum_{k\in\mathcal{U}_{d_1}} \tilde{p}_k^1 < P_{\max}\right)$ *and* $\left(P_{\max} < \sum_{k=1}^{K_d+1} w_k\right)$ **then**
    Allocate $\tilde{p}_\ell^3 = 0$ for users $\ell = (K_d+1),\ldots,L$;
    Solve (10) to obtain $\tilde{\mathbf{p}}^3$ for users $\ell = 1,\ldots,K_d$ with power budget $P_{\max}$;
    Define $\mathcal{U}_{d_3} := \mathcal{U}_{d_1}$;
    $SR_3 = \sum_{k\in\mathcal{U}_{d_1}} \log\left[1+(N-K_d)\psi_k\tilde{p}_k^3\right]$.
  end
end
Select $l^\star := \max_i SR_i$;
Output: User set $\mathcal{U}^\star(N) = \mathcal{U}_{d_l^\star}$ and corresponding $\mathbf{p}^\star(N) = \tilde{\mathbf{p}}^{l^\star}$

Algorithm 3: Joint Optimal $(\mathbf{p}, N)$ for Long Timescale.

Input: $M,\ K,\ \gamma,\ \{\psi_k\}_{k=1}^{K},\ P_T$, and P_{RF} ;
for $N = K$ *to* M **do**
  Obtain $\mathbf{p}^\star(N)$ using Algorithm 1 ;
  Evaluate
  $U_L\left(\mathbf{p}^\star(N), N\right) = \sum_{k=1}^{K} \log\left[1+(N-K)\psi_k p_k^\star(N)\right] - cN$;
end
Output: $N^\star = \arg\max_N U(\mathbf{p}^\star(N), N)$ and $\mathbf{p}^\star(N^\star)$

B. Joint User Scheduling, Power Allocation, and Antenna Selection

Having obtained the solution to the joint power allocation and user scheduling problem from either Algorithms 1 or 2, it remains to find the optimal number of antennas to be used. The optimal N may again be found via a line search over N from 1 to $\lfloor P_T/P_{\text{RF}} \rfloor$, beyond which the problem becomes infeasible. For each candidate value of N, the maximum achievable sum-rate may be obtained using Algorithms 1 or 2. After the search, the value of N corresponding to the largest utility function (sum-rate minus penalty) should be selected and designated $N^\star$, as summarized in Algorithm 3. The worst case complexity of Algorithm 3 is $\mathcal{O}(M^2\,L^2)$ but may be reduced to $\mathcal{O}(M^2\,L)$ through the use of the dual-relaxation approach.

IV. Instantaneous Resource Allocation Problem

With the user and antenna selection discussed under the block resource allocation problem, we now turn our attention to the instantaneous resource allocation in $(\mathcal{P}_S)$. Toward this end, we begin with first discussing the solution of the problem for a fixed given $(\mathcal{A}, N)$.

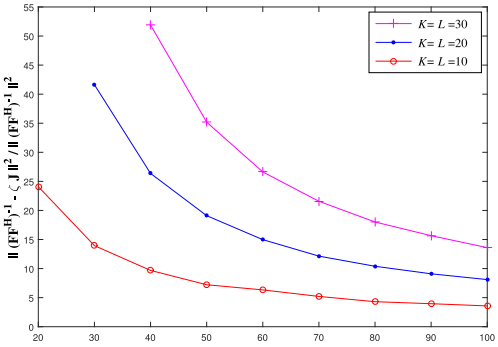

Fig. 1. Validity of approximation in (13) w.r.t. a subset of antenna N for $M = 100$.

A. Joint User Selection and Power Allocation for Given $\mathcal{A}$

Let, $\mathbf{\Delta}_{\mathcal{A}}$ denote the antenna selection matrix corresponding to a known antenna set $\mathcal{A}$, and let $t_k^{\mathcal{U}} := [(\mathbf{F}\mathbf{F}^H)^{-1}]_{kk} = [(\mathbf{H}_{\mathcal{U}}\mathbf{\Delta}_{\mathcal{A}}\mathbf{H}_{\mathcal{U}}^H)^{-1}]_{kk}$ for $1 \leq k \leq K$, where recall that $\mathbf{F} := \mathbf{H}_{\mathcal{U}\mathcal{A}}$ and $N = |\ \mathcal{A}\ |$. Given N, the second term in the objective function of $(\mathcal{P}_S)$ can be dropped, and the equivalent utility maximization problem becomes

$$\max_{\mathbf{p},\mathcal{U}} \sum_{k\in\mathcal{U}} \log\left[1+\frac{p_k}{t_k^{\mathcal{U}}}\right]$$
$$\text{s.t.} \sum_{k\in\mathcal{U}} p_k \leq P_{\max}, \quad p_k \geq \gamma\, t_k^{\mathcal{U}} \qquad \forall k \in \mathcal{U} \qquad (11)$$

where $P_{\max} := P_T - NP_{\text{RF}}$. Different from (7), the gains $\widetilde{\psi}_k^{\mathcal{A}} := \frac{1}{t_k^{\mathcal{U}}}$ now depend on the set of scheduled users $\mathcal{U}$. As a result, scheduling a specific user k changes the gains of all other users $k \neq k'$. Consequently, *Lemma* 2 no longer applies and solving (11) would incur combinatorial complexity. It is remarked that the issue can be partially averted in MIMO ZF-beamforming systems without the QoS-constraints. For instance, [17] proposed a semi-orthogonal user selection (SUS) algorithm that uses a greedy approach for user selection.

We propose to solve (11) approximately using the fact that N for massive-MIMO systems is large. Indeed, the following property is well-known and has been widely used to simplify the design of massive-MIMO systems [6]

$$\lim_{M\to\infty} (1/M)\mathbf{H}\mathbf{H}^H \to \mathbf{D} \qquad (12)$$

where the matrix $\mathbf{D} := \text{diag}(\beta_1,\ldots,\beta_L)$ are the large-scale channel gains between the BS and the users $k \in \mathcal{U}^0$.

In other words, for large M, the rows of $\mathbf{H}/\sqrt{M}$ are almost orthonormal. While (12) cannot be directly applied to the present case, we infer a similar property for the matrix $\mathbf{F}$. Specifically, the matrix $\mathbf{F}\mathbf{F}^H$ is almost diagonal, and the ratio $\rho := \|\text{diag}(\mathbf{F}\mathbf{F}^H)\|_2/\|\mathbf{F}\mathbf{F}^H\|_F \to 1$ as $N \to \infty$. It can, therefore, be concluded that the following approximation holds for large N:

$$(\mathbf{F}\mathbf{F}^H)^{-1} \approx \binom{N}{K}\,\text{diag}\left(\frac{1}{\|\mathbf{f}_1\|^2},\ldots,\frac{1}{\|\mathbf{f}_K\|^2}\right) := \zeta\mathbf{J} \qquad (13)$$

where $\binom{N}{K} = \frac{N}{N-K}$, $\mathbf{f}_k$ is the kth row of $\mathbf{F}$, $\zeta := \frac{N}{N-K}$ and $\mathbf{J} := \text{diag}(\frac{1}{\|\mathbf{f}_1\|^2},\frac{1}{\|\mathbf{f}_2\|^2},\ldots,\frac{1}{\|\mathbf{f}_K\|^2})$.

The validity of the approximation in (13) can be seen from Fig. 1, where it is observed that for sufficiently large value of N, the percentage error in the approximation is less than 10% when

$L \leq 20$. We remark that studying the effect of approximation on the solution of (11) is not straightforward since solving the exact problem incurs combinatorial complexity. Therefore, we resort to examining the effect of the approximation in isolation only.

It is remarked that such an approximation reduces otherwise the highly nonconvex nature of the problem and also legitimizes the near-optimality of the ZF-precoder for massive-MIMO systems. Using (13), it can be seen that (11) becomes equivalent to (7) if we use the gain $\widetilde{\psi}_k^{\mathcal{A}} = \|\mathbf{g}_k\|^2(N-K)/N$ for all $1 \leq k \leq L$, where $\mathbf{g}_k^T$ is the kth row of $\mathbf{H}_{\bullet\mathcal{A}}$ and also $\mathbf{g}_k^T = \mathbf{f}_k$ if $k \in \mathcal{U}$. Subsequently, we can make use of Algorithms 1 or 2 for power allocation.

B. Joint Antenna Selection and Power Allocation Without QoS-Constraints

Different from the block resource allocation, antenna selection for instantaneous resource allocation entails determining the set $\mathcal{A}$. Solving the resulting set-valued problem is difficult and necessitates making approximations. In order to better understand the antenna and user selection problems, let us first consider the joint antenna selection and power allocation problem without the combinatorial user selection requirement. To this end, the QoS-constraints are temporarily dropped, allowing the use of successive convex approximation algorithms. The subsequent section will reintroduce the QoS-constraints as originally intended, and utilize the results obtained in Sections IV-A and IV-B in order to solve $(\mathcal{P}_S)$ in Section II-B. Henceforth, the results obtained in Sections IV-A and IV-B will subsequently be utilized to solve $(\mathcal{P}_S)$ in Section II-B.

In the literature, the antenna selection problem is often solved via greedy algorithms whose performance in many cases is far from optimal [28]. In the context of massive-MIMO systems, other heuristic algorithms have also been proposed that perform slightly better than greedy algorithms [30]. This article considers a more principled approach that will be shown to yield the near-optimal performance at low complexity. The approach consists of the following three key steps:

1) relaxing the integer constraint on $\Delta_i \in \{0,1\}$ to $\Delta_i \in [0,1]$;
2) solving the resulting (nonconvex) problem using majorization–minimization (MM) [45] or block-coordinate descent (BCD) [46] approach;
3) randomized rounding [47] to obtain a feasible antenna set $\mathcal{A}$.

Recall that $\text{SINR}_S(k) = \frac{p_k}{[(\mathbf{H}\boldsymbol{\Delta}\mathbf{H}^H)^{-1}]_{kk}}$, implying that the objective function of $(\mathcal{P}_S)$ is a decreasing function of the term $[(\mathbf{H}\boldsymbol{\Delta}\mathbf{H}^H)^{-1}]_{kk}$. Applying the epigraph trick [48] by introducing variables t_k, the relaxed version of $(\mathcal{P}_S)$ can be written as

$$\max_{\mathbf{p},\boldsymbol{\Delta},\mathbf{t}} \sum_{k=1}^{L} \log\left(1+\frac{p_k}{t_k}\right) - c\,\text{tr}(\boldsymbol{\Delta}) \qquad (\mathcal{P}_{S_1})$$

$$\text{s.t.} \quad \sum_{k=1}^{L} p_k + \text{tr}(\boldsymbol{\Delta})P_{\text{RF}} \leq P_T \qquad (14a)$$

$$[(\mathbf{H}\boldsymbol{\Delta}\mathbf{H}^H)^{-1}]_{kk} \leq t_k \qquad \forall k \in \mathcal{U}^0 \qquad (14b)$$

$$0 \leq \Delta_i \leq 1 \qquad \forall i \in \mathcal{A}^0 \qquad (14c)$$

$$L \leq \text{tr}(\boldsymbol{\Delta}) \leq M \qquad (14d)$$

where $\mathbf{t} \in \mathbb{R}_{++}^L$ collects the temporary variables $\{t_k\}_{k=1}^L$.

The relaxed problem in $(\mathcal{P}_{S_1})$ is still nonconvex due to the nonconvex objective function. Instead, an approximate solution to $(\mathcal{P}_{S_1})$ will be obtained by exploiting the special structure of the objective function.

1) MM Algorithm: The MM-algorithm can be applied to nonconvex problems that admit a simple majorizer. The idea here is to construct successive convex approximations of the objective function in $(\mathcal{P}_{S_1})$ that adhere to certain regularity conditions. Specifically, it is required that the convex surrogate function is a tight upper bound to the original function. For simplicity, let us convert $(\mathcal{P}_{S_1})$ into a minimization problem whose objective function may be written as

$$f(\mathbf{p},\mathbf{t},\boldsymbol{\Delta}) = c\,\mathbf{tr}(\boldsymbol{\Delta}) - \sum_{k=1}^{L}\log(p_k+t_k) + \sum_{k=1}^{L}\log(t_k) \qquad (15)$$

where the first two terms are linear and convex, respectively, while the last term $\ell(\mathbf{t}) := \sum_k \log(t_k)$ is concave. Starting with an arbitrary initial $\mathbf{t}^0$ (further detailed in Section V-B), we make use of the following majorizer at $(\upsilon+1)$th iteration:

$$\ell(\mathbf{t} \mid \mathbf{t}^\upsilon) = \sum_{k=1}^{L}\left(\log(t_k^\upsilon) - 1 + \frac{t_k}{t_k^\upsilon} + \frac{\eta}{2}(t_k - t_k^\upsilon)^2\right) \qquad (16)$$

where $\mathbf{t}^\upsilon$ is the iterate from the previous iteration and η is a regularization parameter that makes $\ell(\mathbf{t} \mid \mathbf{t}^\upsilon)$ a strongly convex function of $\mathbf{t}$. Therefore, the convex surrogate of the objective function becomes

$$f(\mathbf{p},\mathbf{t},\boldsymbol{\Delta} \mid \mathbf{t}^\upsilon) = c\,\mathbf{tr}(\boldsymbol{\Delta}) - \sum_{k=1}^{L}\log(p_k+t_k) + \ell(\mathbf{t} \mid \mathbf{t}^\upsilon) \qquad (17)$$

and the next iterate is obtained by solving

$$\mathbf{t}^{\upsilon+1} := \arg\min_{\mathbf{t}} \min_{\mathbf{p},\boldsymbol{\Delta}} f(\mathbf{p},\mathbf{t},\boldsymbol{\Delta} \mid \mathbf{t}^\upsilon) \qquad (18)$$

$$\text{s.t.} \ \ (14a) - (14d). \qquad (19)$$

Observe here that the objective function $\min_{\mathbf{p},\boldsymbol{\Delta}} f(\mathbf{p},\mathbf{t},\boldsymbol{\Delta} \mid \mathbf{t}^\upsilon)$ is not smooth in $\mathbf{t}$, and therefore, standard results pertaining to the convergence analysis of the MM algorithm do not apply (see, e.g., [45]). Nevertheless, the MM algorithm was found to exhibit convergent behavior empirically.

2) BCD Algorithm: Next, we consider the BCD-algorithm where the optimization variables are partitioned into several mutually exclusive subsets. The algorithm comprises of solving the optimization problem over each subset of variables in an iterative fashion. In the present case, we begin with arbitrary values of $\boldsymbol{\Delta}^0$ and $\mathbf{p}^0$ and solve the following problems at the υth iteration.

Block 1: Given $\boldsymbol{\Delta}^\upsilon$ and $\mathbf{p}^\upsilon$, the optimal value of $\mathbf{t}$ is obtained by solving the following problem:

$$\mathbf{t}^{\upsilon+1} = \arg\max_{\mathbf{t}} \sum_{k=1}^{L}\log\left(1+\frac{p_k^\upsilon}{t_k}\right)$$
$$\text{s.t.} \quad [(\mathbf{H}\boldsymbol{\Delta}^\upsilon\mathbf{H}^H)^{-1}]_{kk} \leq t_k \quad \forall k = 1,\ldots,L. \qquad (20)$$

Interestingly, the objective function as well as the constraints are separable in each t_k, yielding the closed form solution to (21) as $t_k^{\upsilon+1} = [(\mathbf{H}\boldsymbol{\Delta}^\upsilon\mathbf{H}^H)^{-1}]_{kk}$ for all $1 \leq k \leq L$.

Block 2: Having $\mathbf{t}^{v+1}$ and $\mathbf{\Delta}^{v}$, the power allocation problem can be solved as

$$\mathbf{p}^{v+1} = \arg\max_{\mathbf{p}} \sum_{k=1}^{L} \log\left(1+\frac{p_k}{t_k^{v+1}}\right)$$
$$\text{s.t.} \sum_{k=1}^{L} p_k \leq P_T - \text{tr}(\mathbf{\Delta}^{v}) P_{\text{RF}} \tag{21}$$

which can be solved using the water filling algorithm.

Block 3: Having $\mathbf{p}^{v+1}$ and $\mathbf{t}^{v+1}$, the optimal value of $\mathbf{\Delta}$ can be obtained by solving $(\mathcal{P}_{S_1})$ with respect to $\mathbf{\Delta}$. It can be seen that the resulting problem is convex and can be cast as the following semidefinite program:

$$\max_{\mathbf{\Delta}} \sum_{k=1}^{L} \log\left(1+\frac{p_k^{v+1}}{t_k^{v+1}}\right) - c\,\text{tr}(\mathbf{\Delta}) \tag{22}$$
$$\text{s.t.} \begin{bmatrix} \mathbf{H}\mathbf{\Delta}\mathbf{H}^H & \mathbf{e}_k \\ \mathbf{e}_k^T & t_k^{v+1} \end{bmatrix} \succeq 0 \qquad \forall k \in \mathcal{U}^0$$
$$0 \leq \Delta_i \leq 1 \qquad \forall i \in \mathcal{A}^0 \tag{23}$$
$$L \leq \text{tr}(\mathbf{\Delta}) \leq \min\left\{M, \frac{P_T - \sum_{k=1}^{L} p_k^{v+1}}{P_{\text{RF}}}\right\} \tag{24}$$

where $\mathbf{e}_k$ is the vector that has 1 at the kth location and zeros elsewhere.

Standard convergence guarantees for the BCD algorithm exist for the case of unconstrained minimization problems with differentiable objective functions. In the present case, however, the constraints are complicated and cannot be readily incorporated into the objective function without making it nondifferentiable. Consequently, it is difficult to show that the BCD algorithm in its current form is convergent. Nevertheless, as we shall see in Section V, no divergent or oscillatory behavior was observed.

As the final step, we apply the randomized rounding technique [47], where the ith antenna is selected with probability Δ_i while ensuring that the number of antennas is feasible. The resulting set of antennas serve as the input to the joint user selection and power allocation problem detailed in Section IV-A. The entire process is repeated multiple times and the antenna set that yields the lowest objective function is chosen as the solution.

It can be seen that the proposed algorithms rely on a key approximation, namely, that the antenna set selected while allocating power to all the users without QoS-constraints, is near optimal when fewer users are selected. Such an approximation is justified when the K is not too small as compared to L, e.g., for the case when QoS threshold γ is small. Conversely, when γ is high, the antenna set that is optimal for all users may be suboptimal for a small group of users.

C. Joint Antenna Selection and Power Allocation With QoS-Constraints

The instantaneous resource allocation problem in $(\mathcal{P}_S)$ is solved in three steps. In the first step, an optimal antenna set is obtained as discussed in the Section IV-B, i.e., without the QoS-constraints. In the second step, a feasible antenna set with integer elements $\in \{0,1\}$ are obtained using the randomization method detailed earlier. Finally, in the third step, the user scheduling algorithms provided in Section IV-A are applied. The three steps are repeated a fixed number of times and the resulting antenna and user sets are output. It is remarked the overall algorithm is still heuristic and no claims regarding its convergence can be made. To provide a more clear understanding for the readers, Fig. 2 presents the summarized approach followed in this article for the two resource allocation scenarios.

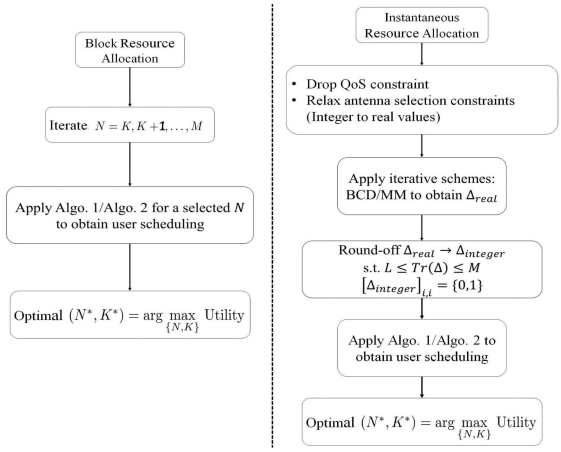

Fig. 2. Flowchart for the proposed approach in the two resource allocation scenarios.

TABLE I
PERFORMANCE COMPARISON FOR DIFFERENT USER SCHEDULING SCHEMES FOR LARGE-SCALE FADING WITH $M = N = 128,\ L = 50,\ P_{\text{RF}} = 0.1$ W, AND $P_{\max} = P_T$

P_T(dB)	Alg. 1 (Sum-Rate)	Alg. 2 (Sum-Rate)	$K_{\max}$	$K^\star$	Hybrid-approach (# Steps)
-20	61.15	54.28	17	14	4
-17	75.37	67.62	21	16	6
-15	86.30	78.09	23	18	7
-13	98.34	89.60	26	20	8
-10	118.89	109.68	31	23	10
-8	134.40	124.96	35	26	11
-5	160.63	151.46	41	30	13
-3	180.22	171.49	46	33	15
0	212.88	205.71	49	38	14
5	276.52	275.41	50	45	7
10	349.97	349.97	50	50	1

V. SIMULATION RESULTS

This section provides simulation results and comparisons between the various proposed schemes discussed under the block and instantaneous resource allocation scenarios. Simulation results are generated by averaging over 100 Monte Carlo runs. Unless otherwise specified, we set $M = 100, c = 0.1$ units per antenna, $P_{\text{RF}} = -10$ dB per antenna, and $\gamma = 2$. We begin with the discussion of the long timescale resource allocation in Section V-A and subsequently discuss the short timescale resource allocation problem in Section V-B.

A. Long Timescale: Block Resource Allocation

The large-scale fading components are set as $\beta_k(\text{dB}) = 130 + 37.6\log(d_k \times 10^{-3})$ [18], where d_k denotes the distance of the kth user from the BS. For simulation purposes, users were assumed to be uniformly distributed between $d_k = 50$ to 250 m.

We begin with the quantitative comparison of the sum-rate performance (in bits/second/hertz) of the proposed algorithms, i.e., Algorithms 1 and 2 with varying total power budget P_T for a large system as presented in Table I. Recall that $K_{\max} :=$

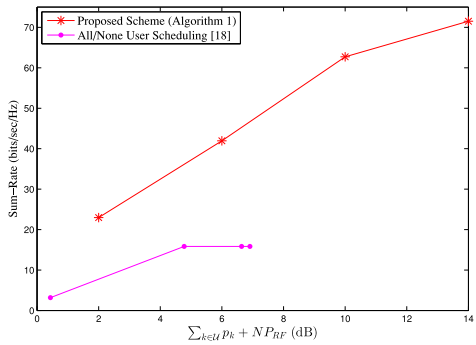

Fig. 3. Sum-rate performance w.r.t. the total power consumed (in dB) by the proposed scheme in Algorithm 1 versus Algorithm in [18] with $L = 10$ and $c = 0.01$.

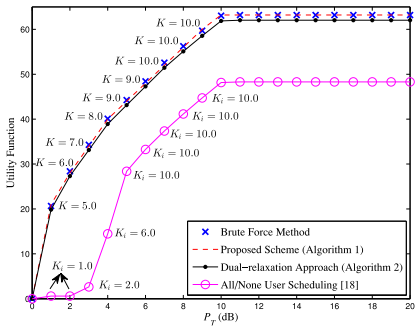

Fig. 4. Utility function $U_L(\mathbf{p}^\star(N), N)$ for varying power (P_T) with $L = 10$ and $c = 0.01$.

$\max\{K \mid \frac{\gamma}{N-K}\sum_{k=1}^{K}\frac{1}{\psi_k} \le P_{\max}\}$ and also that $K^\star \le K_{\max}$ is the optimal number of users scheduled. The last column in the table presents the number of bisection searches required to achieve the optimal performance using the hybrid approach as discussed in Section III-A under the dual relaxation scheme. Observe that the dual-relaxation scheme, i.e., Algorithm 2 is suboptimal for low power budget but achieves the near-optimal performance as the power budget increases.

Next, we show the key distinction between the proposed scheme in Algorithm 1 and the proposed scheme of the all/none algorithm in [18]. The objective problem in [18] is to minimize the total power allocated to all the users and to the RF-circuit associated with each antenna element under a given QoS-constraint. Thus, a closed-form feasible solution is obtained in [18] that allocates power to all the users if sufficient power is available or no users are allocated power. Thus, Fig. 3 presents the sum-rate perfromance of both the schemes for a total power budget of P_T (dB) $= [2,\ 6,\ 10,\ 14]$ and is subsequently presented against the actual total power consumed $((\sum_{k\in\mathcal{U}} p_k + NP_{\text{RF}}))$ on average by the two different schemes. It can be seen that our proposed scheme utilizes all the available power while the scheme proposed in [18] utilizes less power and in turn achieves a lower sum-rate. Thus, our proposed scheme has a clear advantage in a cellular system where the total power budget is fixed and known.

Fig. 4 shows the utility achieved for P_T ranging from 1 to 20 dB. The average number of users scheduled by the Algorithm 1 is denoted by K and the average number of users scheduled by all/none user scheduling scheme [18] is denoted by K_i. As expected, the output of Algorithm 1 matches that of the brute force search (BFS) algorithm that examines every possible subset of users for all values of N. For the purposes of comparison, Algorithm 2 and the minimum power allocation scheme of [18] are also presented, where the goal is simply minimize the power consumption as long as the QoS-constraints are met for all users. Note that the utility achieved by the method in [18] is lower than that achieved by the current algorithm as the goals of the two algorithms are different. The comparison is still included here simply to demonstrate the different resource allocation modes possible in massive-MIMO systems. For instance, with $P_T = 4$ dB, the present algorithm schedules only about eight out of the ten users as compared to the six users scheduled by the all/none algorithm in [18]. Due to fewer users being scheduled on average, however, the average utility achieved is only about a third of that achieved by the proposed algorithm. As expected, however, the algorithm in [18] also utilizes significantly lower power, e.g., only about -4 dB power on average even when $P_T = 4$ dB. Next, we examine the long timescale problem in greater detail by studying the impact of parameters c and γ. Fig. 5 shows the utility against the number of transmit antennas (N) for different values of c associated with the RF-chains. It can be seen that the utility increases with N but the peak might decrease when c is large. Intuitively, when the per-antenna cost is high, it might be prudent to transmit using a few antennas only. Conversely, when c is small, it makes no sense to turn off any of the antennas and $N^\star = M$.

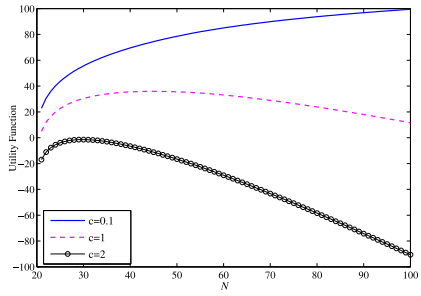

Fig. 5. Utility function $U_L(\mathbf{p}^\star(N), N)$ versus different BS antennas for various c values with $L = 20$ and $P_T = 13$ dB.

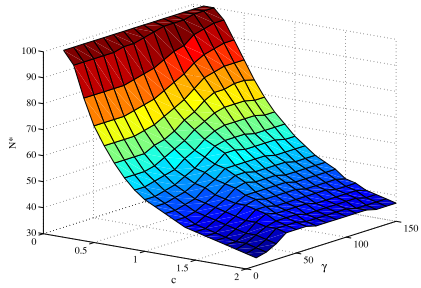

Fig. 6. Optimal number of antennas (N^*) for varying c values and SINR threshold (γ) at $L = 20$ and $P_T = 13$ dB. Total Monte Carlo run $= 1000$.

To obtain further intuition, Fig. 6 shows a plot of the optimal number of antennas $N^\star$ for different values of RF-chain cost c and SINR threshold γ. As expected, $N^\star$ decreases monotonically with c, irrespective of the value of γ. Interestingly, however, for low values of c, the optimal number of antennas $N^\star$ is close to M and increases monotonically for $1 < \gamma < 150$. In other words, as long as c is not too high, using more antennas may offset the loss in utility due to stricter QoS-constraints or higher

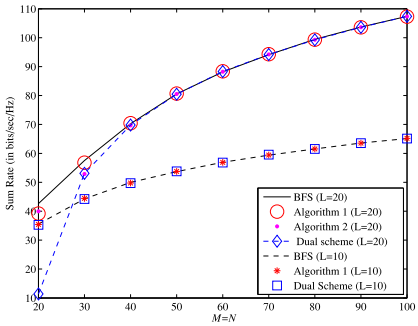

Fig. 7. Sum rate performance of different user scheduling schemes for varying number of antennas ($M = N$) for $L = 10$ and 20 at $P_{\max} = 10$ dB.

γ. However, the intuition does not hold when c is large, since the additional advantage obtained from using more antennas gets nullified culminating in lower utility.

B. Small Timescale: Instantaneous Resource Allocation

For the more difficult short timescale problem, we have assumed $\mathbf{D} \simeq \mathbf{I}_L$ (i.e., $\beta_k = 1\,\forall k$, as the related performance analysis would largely depend on the small-scale fading component per coherence interval basis rather than the large-scale fading component which remains constant over multiple coherence interval). We begin again by comparing the different user scheduling approaches proposed in Section III-A for varying number of transmit antennas with number of users $L = 10$ and 20 as shown in Fig. 7. The proposed scheduling schemes are shown to achieve the near-optimal performance as compared to the optimal BFS scheme. Observe that for the case when $L = 20$ and number of transmit antennas are less than 50, the performance of the proposed schemes are suboptimal as a reason of the poor effective channel gains occurring due to possible presence of users with bad channel condition. The relative dependency of $\tilde{\psi}_k^A$ on the channel of other users is shown to reduce substantially as the number of transmitting antennas increases, and hence, a near-optimal performance is achieved. The solution obtained directly applying (10) called the dual scheme, is seen to perform the worst for low number of transmit antennas for $L = 20$. Furthermore, the performance of the dual-scheme can be improved by applying the dual-relaxation scheme (see Algorithm 2). Thus, the motivation to use large number of transmit antennas is very well established, especially to obtain near-optimal performance in a ZF-precoding system with large number of users.

Next, we compare the performance of the proposed MM and BCD algorithms with that of the BFS. The performance of the greedy algorithm from [28], referred to as JASUS, is also shown for comparison. The JASUS algorithm proposed in [28] considers a joint antenna selection and user scheduling for a fixed number of RF chains with the objective to maximize the sum-rate under the total power constraint. However, since JASUS was first proposed for a fixed value of N, we carry out a line search over all feasible values of N and pick the solution that yields the best utility value. Comparison is also included with the greedy iterative antenna selection (IAS) algorithm in [30]. It considers a joint antenna selection and user scheduling, where the objective is to maximize the sum-rate under a total power constraint w.r.t users power allocation and the power consumption of RF-circuits [digital-to-analog converter (DAC) mixers, filters] to guarantee a minimum SINR. Furthermore, since the user scheduling scheme in [30] requires equal received power allocation at all users, we instead use optimal user scheduling (via BFS) and power allocation within the IAS algorithm. Note that the values of M and L are kept small, since the BFS algorithm is no longer viable for larger values.

TABLE II
WORST-CASE COMPLEXITY FOR $L \leq N \leq M$

Algorithm 1	$\mathcal{O}(NL^2)$
Algorithm 2	$\mathcal{O}(NL)$
Algorithm 3	$\mathcal{O}(M^2L)$
MM Alg.	$\mathcal{O}(M) + \mathcal{O}(NL^2)$
BCD Alg.	$\mathcal{O}(M) + \mathcal{O}(NL^2)$
All/no user scheduling [18]	$\mathcal{O}(L^3)$
JASUS [28]	$\mathcal{O}(ML^3)$
IAS [30]	$\mathcal{O}(M^2L^2)$

It can be seen from Fig. 8(a) that our proposed algorithms are near-optimal for small γ. Such a behavior suggests that the loss of optimality due to the massive-MIMO approximation in Section IV-A, the possibility of getting stuck at a saddle point in Section IV-B, and due to the use of various heuristics employed in Section IV, is very small. However, as evident from Fig. 8(b) and as expected, the utility decreases with γ. However, the proposed algorithm decreases slowly with γ, in contrast to the existing algorithms that show larger variations. Nevertheless, since all the existing algorithms rely on approximations, they do eventually become suboptimal for large values of γ. Indeed, JASUS and modified IAS algorithms are far from optimal for most values of γ and M. It can be seen from Fig. 8(a) and (b), that MM algorithm is slightly better than the BCD algorithm at low total power budget (P_T) and also at high SINR threshold (γ). In conclusion, the loss in the utility from ignoring a few users is not significant. Consequently, the system considered in Fig. 8(a) and (b) is not a massive MIMO system. Nevertheless, it helps us understand how suboptimal the proposed algorithm is, as compared to the optimal BFS algorithm.

Next, Fig. 8(c) compares the performance of the proposed MM and the BCD methods for a large system with $M = 100$ and $L = 50$. It can be observed that the utility increases almost linearly with increase in the total power budget P_T. It is remarked, however, that the BCD-based algorithm was observed to be more sensitive to initialization. That is, while the MM algorithm yielded almost the same utility regardless of the initialization, the BCD-based algorithm required a good initialization point. Henceforth, the antenna set obtained from [28] was used as an initialization point to obtain $\mathbf{\Delta}^0$ for the MM and the BCD schemes. Finally, Fig. 9 compares the optimal number of antennas required and the respective utility value achieved at different cost values $c = [0.1\ 0.05\ 0.01\ 0.005\ 0.001\ 0.0005\ 0.0001]$.

It can be observed that the MM-based algorithm requires fewer antennas to achieve the same utility as compared to the BCD and the N-Search JASUS methods. Furthermore, the number of optimal antennas for the last 2 or 3 values of "c" are the same for the different schemes. Finally, the complexity of the proposed algorithms are summarized in Table II.

Note that the big-$\mathcal{O}$ notation includes the constant $\log(\frac{1}{\epsilon})$, where ϵ is the desired precision and $\log(\frac{1}{\epsilon})$ is the number of steps required by the bisection algorithm used to solve for λ in Algorithms 1 and 3. Moreover, Algorithm 1 has also been utilized in the MM, BCD, JASUS, and IAS algorithms.

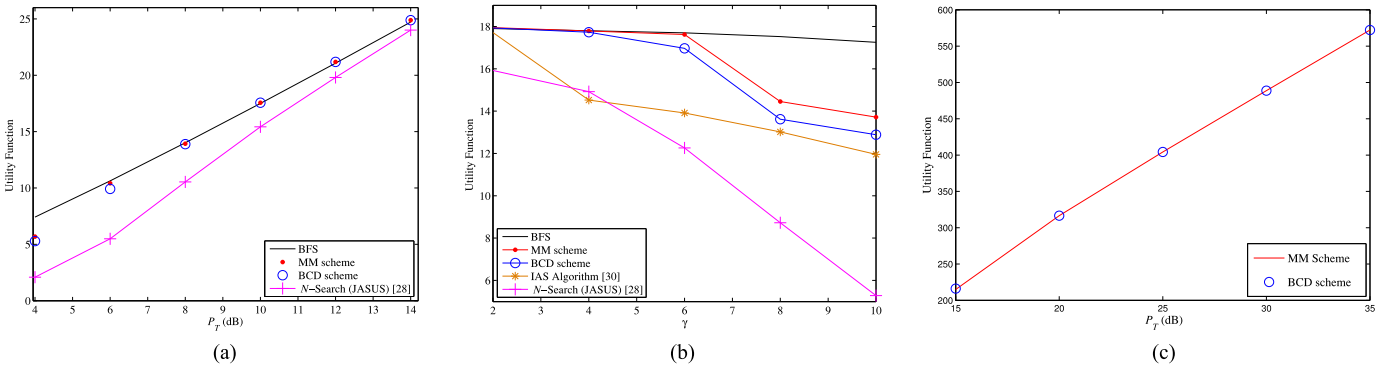

Fig. 8. (a) Utility function of all the schemes for varying power (P_T) at $M = 10$, $L = 6$, and $c = 0.1$. (b) Utility function w.r.t. varying SINR threshold (γ) for $M = 10,\ L = 6,\ P_T = 10$ dB, and $c = 0.1$. (c) Utility function for varying power at $M = 100, L = 50$ and $c = 0.1$.

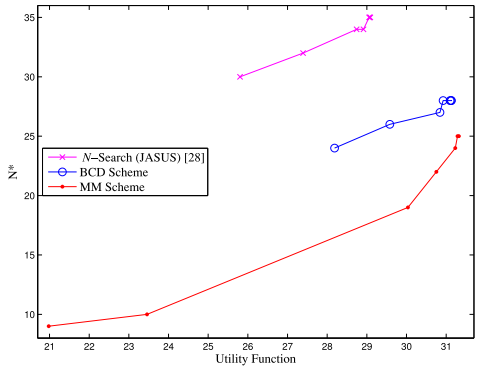

Fig. 9. Optimal antenna selected w.r.t. the optimal utility achieved for $M = 100$, $L = 10$, $P_T = 10$ dB, and $\gamma = 2$ at $c = [0.1\ 0.05\ 0.01\ 0.005\ 0.001\ 0.0005\ 0.0001]$.

VI. Conclusion

This article considered the problem of joint antenna selection and user scheduling in a cellular system enabled with massive-MIMO. Different from the existing works, a general utility maximization framework is considered, where the goal is to optimally select part of transmit antennas limited by the RF-chains installed at the BS and simultaneously allocate power, i.e., schedule certain number of mobile users that satisfies the minimum QoS. The problem is considered under both, long and short time-scales. While, the user selection requirement renders the long time-scale problem combinatorial, a polynomial time algorithm is proposed that still yields the exact solution. For solving the more challenging short time-scale problem, convex relaxation-based algorithms are developed that were shown to significantly outperform the standard greedy approaches while incurring low complexity. Furthermore, the utility increases with the number of users to a certain point and starts decreasing within the set of feasible users for a given total power. Detailed simulations are provided in order to demonstrate the efficacy of the proposed algorithms.

Appendix A
Proof of Lemma 1

Let $\mathcal{U}^\star := \{1, \ldots, K\}$ be the optimal set of scheduled users and $\mathbf{p}^\star \in \mathbb{R}^K_{++}$ be the optimal power allocations. By way of contradiction, let us assume that $\mathbf{1}^T\mathbf{p}^\star =: P^\star < P_{\max}$. Consider the power allocation $\mathbf{p}' \in \mathbb{R}^K_{++}$ such that $p'_k = p^\star_k + \frac{p^\star_k}{P^\star}(P_{\max} - P^\star)$, so that $\mathbf{1}^T\mathbf{p}' = P_{\max}$ and $p'_k P^\star = p^\star_k P_{\max}$. Let us consider the SINR difference for the allocations $\mathbf{p}'$ and $\mathbf{p}^\star$ for the kth user, given by

$$\begin{aligned}
&\frac{(N-K)\alpha_k p'_k}{1+(\beta_k-\alpha_k)P_{\max}} - \frac{(N-K)\alpha_k p^\star_k}{1+(\beta_k-\alpha_k)P^\star} \\
&\quad= \frac{\alpha_k(N-K)}{(1+(\beta_k-\alpha_k)P_{\max})(1+(\beta_k-\alpha_k)P^\star)} \\
&\qquad\times [p'_k(1+(\beta_k-\alpha_k)P^\star) - p^\star_k(1+(\beta_k-\alpha_k)P_{\max})] \\
&\quad= \frac{\alpha_k(N-K)(p'_k-p^\star_k)}{(1+(\beta_k-\alpha_k)P_{\max})(1+(\beta_k-\alpha_k)P^\star)} > 0. \quad (25)
\end{aligned}$$

In other words, the SINR of the kth user is always better when using the power allocation $\mathbf{p}'$ as opposed to when using $\mathbf{p}^\star$. Therefore, the allocation $\mathbf{p}'$ is not only feasible with respect to the QoS-constraints, but also yields a higher sum-rate. However, such a result is absurd since $\mathbf{p}^\star$ was assumed to the optimal solution to (6). Therefore, our original hypothesis must be false, and it must hold that $\mathbf{1}^T\mathbf{p}^\star = P_{\max}$.

Appendix B
Proof of Lemma 2

For every value of $K \leq K_{\max}$, there exists a feasible solution to (7). Indeed, in the worst case when $K = K_{\max}$, allocating the minimum power $p_k = \frac{\gamma}{(N-K)\psi_k}$ to the kth user for $1 \leq k \leq K$ still yields a feasible power allocation. Having ensured that a feasible power allocation exists, the required result can be established by contradiction. Given K, let the set of users that yield the maximum sum-rate in (7) be denoted by $\mathcal{U}'$. By way of contradiction, let there be some $t \in \{K+1, \ldots, L\}$ such that $t \in \mathcal{U}'$. Consequently, there must exist a user $u \in \{1, \ldots, K\}$ such that $u \notin \mathcal{U}'$. Since ψ_ks are sorted in decreasing order, it holds that $\psi_u > \psi_t$. As a result, the optimal power p_t allocated to user t is also feasible for the user u since $p_t \geq \frac{\gamma}{(N-K)\psi_t} > \frac{\gamma}{(N-K)\psi_u}$. It is, therefore, possible to replace the user t in $\mathcal{U}'$ with the user u, while allocating it the same power p_t. Such a replacement results in the objective function of (7) increasing by a positive quantity $\log[1+(N-K)\psi_u p_t] - \log[1+(N-K)\psi_t p_t]$, while the constraints are still satisfied. This result is contradictory, since the set of K users that maximize the sum-rate is $\mathcal{U}'$. Therefore, the original hypothesis is false, which implies that

there cannot be any $t \in \{K+1, \ldots, L\}$ that belongs to $\mathcal{U}'$. In other words, given K, the set of K users that maximize the sum-rate is given by $\mathcal{U} = \{1, 2, \ldots, K\}$.

REFERENCES

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

[2] 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.

[3] E. G. 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.

[4] C. X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[5] D. Wang, C. Ji, X. Gao, S. Sun, and X. You, "Uplink sum-rate analysis of multi-cell multi-user massive MIMO system," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2013, pp. 5404–5408.

[6] 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.

[7] E. G. Larsson and L. Van der Perre, "Massive MIMO for 5G," *IEEE 5G Tech Focus*, vol. 1, no. 1, pp. 1–4, Mar. 2017.

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

[9] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012.

[10] 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.

[11] M. Benmimoune, E. Driouch, W. Ajib, and D. Massicotte, "Novel transmit antenna selection strategy for massive MIMO downlink channel," *Wireless Netw.*, vol. 23, no. 8, pp. 2473–2484, 2017.

[12] 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.

[13] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[14] D. A. Gore, R. U. Nabar, and A. Paulraj, "Selecting an optimal set of transmit antennas for a low rank matrix channel," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 5, pp. 2785–2788.

[15] R. Rajkumar, C. Lee, J. Lehoczky, and D. Siewiorek, "A resource allocation model for QoS management," in *Proc. IEEE Real-Time Syst. Symp.*, Dec. 1997, pp. 298–307.

[16] R. Rajkumar, C. Lee, J. P. Lehoczky, and D. P. Siewiorek, "Practical solutions for QoS-based resource allocation problems," in *Proc. IEEE Real-Time Syst. Symp.*, Dec. 1998, pp. 296–306.

[17] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[18] K. Senel, E. Björnson, and E. G. Larsson, "Joint transmit and circuit power minimization in massive MIMO with downlink SINR constraints: When to turn on massive MIMO?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1834–1846, Mar. 2019.

[19] A. F. Molisch, M. Z. Win, Y.-S. Choi, and J. H. Winters, "Capacity of mimo systems with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1759–1772, Jul. 2005.

[20] I. Bahceci, T. M. Duman, and Y. Altunbasak, "Antenna selection for multiple-antenna transmission systems: Performance analysis and code construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2669–2681, Oct. 2003.

[21] A. Gorokhov, "Antenna selection algorithms for MEA transmission systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 3, pp. III-2857–III-2860.

[22] M. Gharavi-Alkhansari and A. B. Gershman, "Fast antenna subset selection in MIMO systems," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 339–347, Feb. 2004.

[23] A. Gorokhov, D. A. Gore, and A. J. Paulraj, "Receive antenna selection for MIMO spatial multiplexing: Theory and algorithms," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2796–2807, Nov. 2003.

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

[25] X. Gao, O. Edfors, J. Liu, and F. Tufvesson, "Antenna selection in measured massive MIMO channels using convex optimization," in *Proc. IEEE Globecom Workshops*, Dec. 2013, pp. 129–134.

[26] S. Mahboob, R. Ruby, and V. C. Leung, "Transmit antenna selection for downlink transmission in a massively distributed antenna system using convex optimization," in *Proc. IEEE 7th Int. Conf. Broadband, Wireless Computing, Commun. Appl.*, 2012, pp. 228–233.

[27] 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.

[28] M. Benmimoune, E. Driouch, W. Ajib, and D. Massicotte, "Joint transmit antenna selection and user scheduling for massive MIMO systems," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2015, pp. 381–386.

[29] Y. Dong, Y. Tang, and K. Z. Shenzhen, "Improved joint antenna selection and user scheduling for massive MIMO systems," in *Proc. IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, 2017, pp. 69–74.

[30] R. Hamdi, E. Driouch, and W. Ajib, "Resource allocation in downlink large-scale MIMO systems," *IEEE Access*, vol. 4, pp. 8303–8316, 2016.

[31] A. Liu and V. K. Lau, "Joint power and antenna selection optimization for energy-efficient large distributed MIMO networks," in *Proc. IEEE Int. Conf. Commun. Syst.*, 2012, pp. 230–234.

[32] X. Guozhen, L. An, J. Wei, X. Haige, and L. Wu, "Joint user scheduling and antenna selection in distributed massive MIMO systems with limited backhaul capacity," *China Commun.*, vol. 11, no. 5, pp. 17–30, 2014.

[33] J. Akhtar and K. Rajawat, "Quality-of-service constrained user and antenna selection in downlink massive-MIMO systems," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshop*, 2019, pp. 1–6.

[34] J. Akhtar and K. Rajawat, "QOS-based antenna and user selection in large-scale fading for massive-MIMO systems," in *Proc. IEEE 19th Int. Workshop Signal Process. Advances Wireless Commun.*, Jun. 2018, pp. 1–5.

[35] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[36] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[37] K. Zheng, S. Ou, and X. Yin, "Massive MIMO channel models: A survey," *Int. J. Antennas Propag.*, vol. 2014, 2014, Art. no. 848071.

[38] Y. Pei, T.-H. Pham, and Y.-C. Liang, "How many RF chains are optimal for large-scale MIMO systems when circuit power is considered?" in *Proc. IEEE Global Commun. Conf.*, 2012, pp. 3868–3873.

[39] D. W. K. Ng and R. Schober, "Spectral efficiency in large-scale MIMO-OFDM systems with per-antenna power cost," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2012, pp. 289–294.

[40] D. Ha, K. Lee, and J. Kang, "Energy efficiency analysis with circuit power consumption in massive MIMO systems," in *Proc. IEEE Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2013, pp. 938–942.

[41] R. Guruprasad, K. Son, and S. Dey, "Power-efficient base station operation through user QoS-aware adaptive RF chain switching technique," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 244–250.

[42] R. Zi, X. Ge, J. Thompson, C. Wang, H. Wang, and T. Han, "Energy efficiency optimization of 5G radio frequency chain systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 758–771, Apr. 2016.

[43] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[44] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.

[45] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[46] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, Jun. 2001.

[47] P. Raghavan and C. D. Tompson, "Randomized rounding: A technique for provably good algorithms and algorithmic proofs," *Combinatorica*, vol. 7, pp. 365–374, Dec. 1987.

[48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.