A Minorization-Maximization Method for Optimizing Sum Rate in Non-Orthogonal Multiple Access Systems

A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of Non-Orthogonal Multiple Access Systems

Muhammad Fainan Hanif, Zhiguo Ding *Member, IEEE*, Tharmalingam Ratnarajah *Senior Member, IEEE*, and George K. Karagiannidis *Fellow, IEEE*

***Abstract*—Non-orthogonal multiple access (NOMA) systems have the potential to deliver higher system throughput, compared to contemporary orthogonal multiple access techniques. For a linearly precoded multiple-input single-output (MISO) system, we study the downlink sum rate maximization problem, when the NOMA principle is applied. Being a non-convex and intractable optimization problem, we resort to approximate it with a minorization-maximization algorithm (MMA), which is a widely used tool in statistics. In each step of the MMA, we solve a second-order cone program, such that the feasibility set in each step contains that of the previous one, and is always guaranteed to be a subset of the feasibility set of the original problem. It should be noted that the algorithm takes a few iterations to converge. Furthermore, we study the conditions under which the achievable rates maximization can be further simplified to a low complexity design problem, and we compute the probability of occurrence of this event. Numerical examples are conducted to show a comparison of the proposed approach against conventional multiple access systems.**

***Index Terms*—Non-orthogonal multiple access, orthogonal multiple access, convex optimization, zero forcing, spectral efficiency, connectivity, latency, complexity.**

I. Introduction

Efficient multiple access techniques in wireless systems has long been a sought after desirable feature. Several facets haven been considered, while dealing with the design of multiple access schemes. For example, spectral efficiency, reliability and quality of service, efficient utilization of radio resources, and recently, energy efficiency are some of the objectives, that form the basis of multiple access techniques in wireless communication systems. Non-orthogonal multiple access (NOMA) has been conceived as a breakthrough technology for fifth generation (5G) wireless networks [1]–[3]. The main themes of 5G networks, namely, reduced latency, high connectivity, and ultra-fast speeds are being attributed to devising systems working on the principles of NOMA [1]. NOMA uses power domain to multiplex additional users in the time/frequency/code, slot already occupied by a mobile device. The enabling techniques for NOMA are not new and find their roots in some old principles–*superposition coding* (SC) and *successive interference cancellation* (SIC). SC was first proposed by Cover in [4], as an achieveability scheme for a degraded broadcast channel. Likewise, various versions of SIC have been employed in the past in systems like Vertical-Bell Laboratories Layered Space-Time (V-BLAST) and Code Division Multiple Access (CDMA) [5], [6]. Therefore, in addition to being a candidate for the next generation of 5G wireless networks, it is very important that NOMA has also the potential to integrate well with existing multiple access paradigms.

In NOMA, the base station (BS) transmits a superposition coded signal, which is a sum of all messages of the users. The users are arranged with respect to their effective channel gains i.e., the one with the lowest gain is assumed to be at the bottom of the sequence, the one with the highest gain at the top, while the remaining are arranged in an increasing order between the two. NOMA ensures that the weaker users receive a higher fraction of the total power budget. When a stronger user is allowed to access the slot being occupied by a weaker one, its signal does not adversely impact the performance of the weaker user, as it is already experiencing a channel fade. At the same time, the stronger user can get rid of the interference due to the weaker one, by applying a SIC operation. In traditional orthogonal multiple access schemes, once the slot has been reserved for a user, other users are prohibited from accessing that. This, of course, has a negative impact on the aggregate system's throughput. The major outcome of sharing the same channel slot is that the sum rates are expected to improve, and with intelligent power allocation the weaker users can also be efficiently served.

A. Literature

To the best of our knowledge, as of today, NOMA has mostly been explored for single-input single-output (SISO) systems. For example, in [7] Ding *et al.* studied NOMA for the downlink of a cellular system, and by assuming fixed powers, they derived expressions for the aggregate ergodic sum rate and outage probability of a particular user. Interestingly, in that paper it was concluded that in the absence of a judiciously chosen target data rate, a user can always be in outage. For multiple-input multiple-output (MIMO) systems, Lan *et al.* [8], explored the impact of error propagation of SIC and user

The work of M. F. Hanif and Z. Ding was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/L025272/1. T. Ratnarajah was supported by the UK EPSRC under grant number EP/L025299/1.

M. F. Hanif and Z. Ding are with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, United Kingdom. Email: {m.f.hanif, z.ding}@lancaster.ac.uk and mfh21@uclive.ac.nz.

T. Ratnarajah is with the Institute for Digital Communications, School of Engineering, University of Edinburgh, Edinburgh EH9 3JL, United Kingdom. Email: T.Ratnarajah@ed.ac.uk.

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Khalifa University, PO Box 127788, Abu Dhabi, UAE, and also with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54 124, Thessaloniki, Greece. Email: geokarag@ieee.org.

velocity on the NOMA performance. Their results showed that even in the worst error propagation scenario, NOMA outperforms conventional orthogonal multiple access and can yield performance gains for different user mobility. Chen *et al.* [9], studied NOMA for the downlink of a wireless system, when BS and receivers are each equipped with two antennas. Traditional minimum-mean-squared-error (MMSE) precoding matrices have been used, which do not guarantee maximum throughput for a given user ordering. Similarly, Timotheou *et al.* [10], studied the power allocation for NOMA in a SISO system from a fairness point of view. Finally, Ding *et al.*, investigated MIMO-NOMA in [11], and derived outage probabilities for fixed and more sophisticated power allocation schemes.

B. Contributions

In this paper, we focus on the downlink of a multiple-input single-output (MISO) system, in which the transmit signals of each user are multiplied by a complex precoding vector. The goal is to design these vectors in order to maximize the total throughput of the system, while simultaneously satisfying the NOMA constraints. To solve this problem we rely on the approximation technique that has been commonly dubbed as *concave-convex procedure* (CCP)[1] or *minorization-maximization algorithm* (MMA)[2] [12]–[17]. Under the different name of sequential convex programming a parametric approach has been proposed in [18]. Recently, in the context of weighted sum rate maximization and physical layer multicasting, similar ideas were used by Hanif *et al.* and Tran *et al.* in [19], [20], respectively. In addition, Christopoulos *et al.* [21] have studied a successive convex approximation scheme, which is similar in spirit to the MMA approach, for multicast multigroup beamforming with applications to large scale antenna arrays. Due to the flexible nature of MMA approach, these ideas have also been used in image processing applications [22].

The main contributions of this paper can be summarized as follows:

- By incorporating decodability constraints to ensure that better users can perform SIC, we provide a novel mathematical programming based approach to solve the sum rate maximization problem in the downlink of a MISO system, relying on NOMA principles. Similarly, constraints are also included to guarantee that the desired signals of the weaker users are strong enough to render them non-zero data rates.
- Using the MMA concept, we develop an iterative algorithm that solves the NOMA sum rate maximization problem and obtains complex precoding vectors, which maximize the aggregate throughput. Unlike traditional approaches that rely on semidefinite programming (SDP), to deal with such optimization problems, the MMA based algorithm solves a second-order cone program (SOCP) in each step.
- We show that the proposed algorithm is provably convergent in few iterations. Moreover, a complexity analysis is also carried out to show that the worst case complexity of the SOCP, which we solve in each run, is just polynomial in design dimensions. Furthermore, under plausible assumptions, the algorithm converges to the Karush-Kuhn-Tucker (KKT) point of the original problem.
- We present an approximation to the original optimization program, with the main goal of complexity reduction. To provide more insight, we study conditions under which this approximation is tight. Moreover, for the special case of orthogonal precoding vectors, we provide a probabilistic insight regarding the tightness of the proposed approximation.
- Finally, numerical examples are presented to show the validity of the proposed algorithm. These results reveal that the NOMA transmission outperforms the conventional orthogonal multiple access schemes, particularly when the transmit signal-to-noise ratio (SNR) is low, and the number of users are greater than the number of BS antennas. We also investigate the scenario, where the proposed approximation exactly matches the original problem. In this case, it is shown that the distance between the users and the BS plays a crucial role and affects the system's throughput.

C. Structure

The rest of the paper is organized as follows. In Section II, we describe the system model and formulate the problem. In Section III, we present the preliminaries, needed to outline the algorithm in the next section. The algorithm is developed and analysed in Section IV, while a reduced complexity approximation is motivated and developed in Section V. Finally, numerical results and conclusions are presented in Sections VI and VII, respectively.

D. Notations

Bold uppercase and lowercase letters are used to denote matrices and vectors, respectively. The symbols $\mathbb{C}^n, \mathbb{R}^n$ and $\mathbb{R}^n_+$ are used for n-dimensional complex, real, and nonnegative real spaces, respectively. For a vector $\mathbf{e}$, its j^{th} coordinate is denoted by e_i. Furthermore, $\|\mathbf{e}\|_2$ is used to represent l_2 norm of a vector $\mathbf{e} \in \mathbb{C}^n$, which is defined as $\|\mathbf{e}\|_2 = \sum_{i=1}^{n} |e_i|^2$, where $|e_i|$ is the absolute value of e_i. $\mathcal{O}(.)$ is reserved for complexity estimates. Unless otherwise specified, calligraphic symbols are used to represent sets. $\lceil x \rceil$ is the ceiling function, which returns the smallest integer not less than x. $\nabla \mathbf{e}$ denotes gradient of a vector $\mathbf{e}$. $\min(.)$ gives the minimum of the quantities passed as its argument. $\Re(c)$ and $\Im(c)$ denote the real and imaginary parts of a complex number c, respectively. $\Pr(E)$ denotes the probability of event E. Any new or unconventional notation used in the paper is defined in the place where it occurs.

[1]If the original problem is a minimization instead of a maximization, the procedure has been referred to as convex-concave procedure (CCP).

[2]The MMA has also been called as majorization-minimization algorithm if the original problem is a minimization problem.

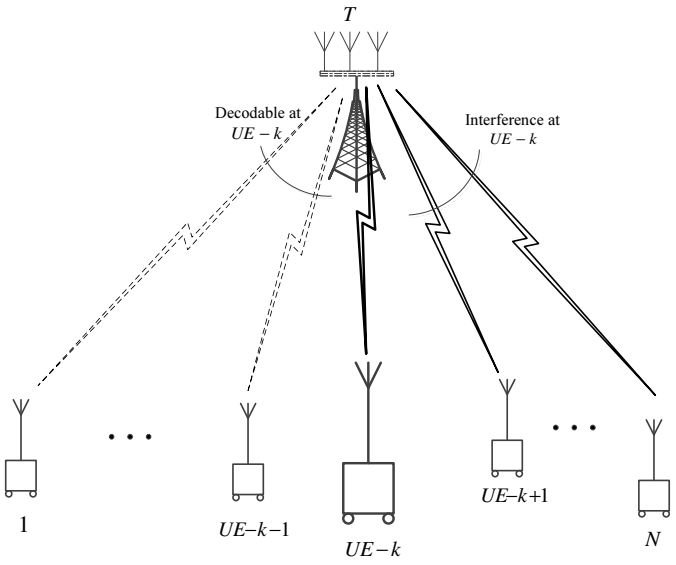


Fig. 1. The system setup. A BS with T antennas serves N users. The user UE-k receives interference from the users UE-$k+1$ to UE-N. The signals of remaining users from UE-1 to UE-$k-1$ are cancelled at UE-k.

II. System Setup

We consider the downlink of a BS, equipped with T antennas and serving N single antenna users. NOMA principle is used for transmission purposes (please refer to Fig. 1). We further assume that the transmitted signal of each user equipment (UE) is linearly weighted with a complex vector. Specifically, to all N users, the BS transmits a superposition of the individual messages, $\mathbf{w}_i s_i$ for all i, where $\mathbf{w}_i \in \mathbb{C}^T$ and s_i are the complex weight vector and the transmitted symbol for UE-i, respectively. Therefore, under frequency flat channel conditions the received signal y_i at UE-i is

$$y_i = \mathbf{h}_i^H \left(\sum_{j=1}^{N} \mathbf{w}_j s_j \right) + n_i = \sum_{j=1}^{N} \mathbf{h}_i^H \mathbf{w}_j s_j + n_i, \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{h}_i = \sqrt{d_i^{-\gamma}} \mathbf{g}_i \in \mathbb{C}^T$ (column vector), with d_i being the distance between i^{th} UE and the BS, γ is the path loss exponent, $\mathbf{g}_i \sim \mathcal{CN}(0, \mathbf{I})$, and n_i represents circularly symmetric complex Gaussian noise with variance σ^2. Subsequently, NOMA proposes to employ SIC at individual UEs, based on the particular ordering. For instance, the works in [2], [7] use the fact that for a single-input single-output (SISO) system, once the channels are arranged in a particular order (increasing or decreasing), then a UE-k decodes all those UE-i signals, whose index $i < k$ (increasing order) and $i > k$ (decreasing order). An illustration of this process is also given in Fig. 1. However, simple SISO ordering cannot be transformed to the MISO setup. The present work does not focus on the optimal ordering problem, but in the design of the complex weighting vectors, $\mathbf{w}_i$, that maximize the aggregate throughput of the system, for a given UE ordering. Next, we assume that the channel state information (CSI) is perfectly known at all nodes.

A. Problem Formulation

We assume that the UE-1 is the weakest (and hence cannot decode any interfering signals), while UE-N is the strongest user, and is able to nullify all other UE interference by performing SIC. The other UEs are placed in an increasing order with respect to their index numbers. For instance, UE-m is placed before UE-n if index $m < n$. Increasing channel strengths can be used to order the users. It is pertinent to mention here that, unless otherwise indicated, by channel strengths we specifically mean to refer to the norm of channel vectors. But, as mentioned above, this ordering may not be optimal, and better rates may be achievable for different order of users. According to NOMA the achievable rate after SIC operation at the k^{th} user, with $k > i$ for all $i = 1, \ldots, k-1$, is [2], [7]

$$R_k^k = \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j=k+1}^{N} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2} \right), \; 1 \le k \le N-1. \tag{2}$$

An important observation should be noted here. For the above rate to be achievable at UE-k, it is necessary for all UE-j, with $j > k$, to satisfy

$$R_j^k = \log_2 \left(1 + \frac{|\mathbf{h}_j^H \mathbf{w}_k|^2}{\sum_{m=k+1}^{N} |\mathbf{h}_j^H \mathbf{w}_m|^2 + \sigma^2} \right) \ge R_{th} \quad j = k+1, \ldots, N \tag{3}$$

where R_j^k is the rate of UE-j to decode the message of k^{th} UE, and R_{th} is some target data rate for user R_k^k. In addition, to allocate non-trivial data rates to the weaker users, which present a lower decoding capability in a given order, the following condition must also be satisfied

$$|\mathbf{h}_k^H \mathbf{w}_1|^2 \ge \ldots |\mathbf{h}_k^H \mathbf{w}_{k-1}|^2 \ge |\mathbf{h}_k^H \mathbf{w}_k|^2 \ge |\mathbf{h}_k^H \mathbf{w}_{k+1}|^2 \ldots \ge |\mathbf{h}_k^H \mathbf{w}_N|^2. \tag{4}$$

As a further insight, (3) ensures that the signal-to-interference-plus-noise ratio (SINR) of UE-j to decode other user k's message (i.e. SINR corresponding to R_j^k) , where $j > k$, is higher compared to the SINR of UE-k needed to decode its own message (i.e., SINR corresponding to R_k^k). Once this condition is satisfied, all users, which are assumed to be at a 'higher' level in the given ordering, are able to perform SIC. Therefore, we propose to maximize the minimum of the SINRs of 'desired messages' and 'other users' messages'. To further exemplify, consider a three user system with UE-1 the lowest and the UE-3 the highest in the ordering. Now, assume that $\text{SINR}_1^1 \ge T_{th}$ and $\text{SINR}_w^1 < T_{th}$, $w = 2, 3$, where T_{th} is some threshold rate. In this scenario, both users 2 and 3 are unable to decode the message of UE-1 as the SINR_1^1 is at least as large as T_{th}, and therefore, SIC cannot be applied. Motivated by this, we aim at obtaining such precoders that ensure, we have $T_{th} \le \text{SINR}_w^1$, $w = 1, 2, 3$. Moreover, the sequence of inequalities in (4) helps to boost up the desired signal level of the 'lower' level users, and in the absence of this guarantee it is likely that most, if not all, radio resources are allocated to the users that receive very low or no interference. Elaborating a bit more, the gain inequalities in (4) are arranged in such a manner that the higher the ordered position of a user, the lower is its place in these inequalities. This helps

boost the SINRs needed to decode other users' messages. In order to understand it further, let us reconsider the three user example quoted above. We note that UE-3 decodes all three users, UE-1 is able to decode its own signal only, and the UE-2 can decode both UE-1's and its own messages. Particularly, the SINRs decoded at the three nodes read as follows

$$\text{SINR}_1^1 = \frac{|\mathbf{h}_1^H \mathbf{w}_1|^2}{|\mathbf{h}_1^H \mathbf{w}_2|^2 + |\mathbf{h}_1^H \mathbf{w}_3|^2 + \sigma^2} \tag{5}$$

$$\begin{cases} \text{SINR}_2^2 = \frac{|\mathbf{h}_2^H \mathbf{w}_2|^2}{|\mathbf{h}_2^H \mathbf{w}_3|^2 + \sigma^2}, \\ \text{SINR}_2^1 = \frac{|\mathbf{h}_2^H \mathbf{w}_1|^2}{|\mathbf{h}_2^H \mathbf{w}_2|^2 + |\mathbf{h}_2^H \mathbf{w}_3|^2 + \sigma^2} \end{cases} \tag{6}$$

$$\begin{cases} \text{SINR}_3^3 = \frac{|\mathbf{h}_3^H \mathbf{w}_3|^2}{\sigma^2}, \text{SINR}_3^2 = \frac{|\mathbf{h}_3^H \mathbf{w}_2|^2}{|\mathbf{h}_3^H \mathbf{w}_3|^2 + \sigma^2}, \\ \text{SINR}_3^1 = \frac{|\mathbf{h}_3^H \mathbf{w}_1|^2}{|\mathbf{h}_3^H \mathbf{w}_2|^2 + |\mathbf{h}_3^H \mathbf{w}_3|^2 + \sigma^2} \end{cases}. \tag{7}$$

Now, for the three users under consideration, the inequalities generated by the sequence in (4) are

$$\begin{cases} |\mathbf{h}_1^H \mathbf{w}_1|^2 \geq |\mathbf{h}_1^H \mathbf{w}_2|^2 \geq |\mathbf{h}_1^H \mathbf{w}_3|^2, \\ |\mathbf{h}_2^H \mathbf{w}_1|^2 \geq |\mathbf{h}_2^H \mathbf{w}_2|^2 \geq |\mathbf{h}_2^H \mathbf{w}_3|^2, \\ |\mathbf{h}_3^H \mathbf{w}_1|^2 \geq |\mathbf{h}_3^H \mathbf{w}_2|^2 \geq |\mathbf{h}_3^H \mathbf{w}_3|^2. \end{cases} \tag{8}$$

It can be seen that, due to the constraints in (8), the numerators (which define the powers of the signals to be decoded) are made greater than the individual interfering terms in the denominators of the SINRs defined in (5), (6) and (7). The interference from stronger users occupies a lower position in the decreasing sequence of powers, cf. (8). Based on the observations and examples presented above, the sum rate, R_{sum}, therefore, is given by

$$R_{sum} = \sum_{k=1}^{N-1} \log_2 \left(1 + \min\left(\text{SINR}_k^k, \ldots, \text{SINR}_N^k\right)\right) + \log_2 \left(1 + \frac{|\mathbf{h}_N^H \mathbf{w}_N|^2}{\sigma^2}\right), \tag{9}$$

where

$$\text{SINR}_i^k = \frac{|\mathbf{h}_i^H \mathbf{w}_k|^2}{\sum_{m=k+1}^{N} |\mathbf{h}_i^H \mathbf{w}_m|^2 + \sigma^2}, \qquad i = 1, \ldots, N. \tag{10}$$

The optimization problem can be formulated as

$$\underset{\mathbf{w}_i}{\text{maximize}} \quad R_{sum} \tag{11a}$$

$$\text{s. t.} \quad |\mathbf{h}_k^H \mathbf{w}_1|^2 \geq \ldots |\mathbf{h}_k^H \mathbf{w}_{k-1}|^2 \geq |\mathbf{h}_k^H \mathbf{w}_k|^2 \geq |\mathbf{h}_k^H \mathbf{w}_{k+1}|^2 \ldots \geq |\mathbf{h}_k^H \mathbf{w}_N|^2, \quad 1 \leq k \leq N \tag{11b}$$

$$\sum_{i=1}^{N} \|\mathbf{w}_i\|_2^2 \leq P_{th}, \tag{11c}$$

where $\mathbf{w}_i \in \mathbb{C}^T, \forall i$, and the constraint in (11c) represents that the total power, which is upper bounded to P_{th}. It is important to mention here that in the original NOMA [1], [2], its concept was applied only to two users. However, our work is much more general, and not only deals with multiple antenna systems, but also with a multiuser environment (cf. (11)).

III. Prerequisites

In order to solve the optimization problem in (11), we will eventually present an iterative algorithm. However, first it is necessary to transform the original problem and then to apply approximations that render it tractability. Before presenting a detailed treatment of our approach, we remark that minorization-maximization inspired approaches have been used in different contexts in earlier works, for instance, [19], [20].

A. Equivalent Transformations

The problem in (11) is non-convex, and it seems that it is not possible to directly approximate it, since the only convex constraint is the power constraint. Therefore, several steps need to be invoked before we can present an algorithm, which solves this problem approximately. To this end, we first introduce the vector $\mathbf{r} \in \mathbb{R}_+^N$ and observe that (11) can be equivalently written as

$$\underset{\mathbf{w}_i, \mathbf{r}}{\text{maximize}} \quad \left(\prod_{k=1}^{N} r_k\right)^{\frac{1}{N}} \tag{12a}$$

$$\text{s. t.} \quad r_k - 1 \leq \min\left(\text{SINR}_k^k, \ldots, \text{SINR}_N^k\right), \quad k = 1, \ldots, N-1 \tag{12b}$$

$$r_N - 1 \leq \frac{|\mathbf{h}_N^H \mathbf{w}_N|^2}{\sigma^2} \tag{12c}$$

$$\text{(11b) \& (11c)}, \tag{12d}$$

where $\mathbf{w}_i \in \mathbb{C}^T, \forall i, \mathbf{r} \in \mathbb{R}_+^N, r_i, i = 1, \ldots, N$, the components of $\mathbf{r}$, represent the optimization variables, and the objective has been obtained by considering that $\log(\cdot)$ is a non-decreasing function, and the geometric mean of the vector $\mathbf{r}$, i.e., $\left(\prod_{k=1}^{N} r_k\right)^{1/N}$, is concave and increasing[3]. It is well known that the geometric mean is readily expressible as a system of second-order cone (SOC) constraints [23]. So this step has no negative impact on the tractability of the objective function. However, the overall problem still remains intractable. In the proposition to follow, we will present an equivalent formulation which is later used to solve this problem.

Proposition 1: The optimization problem in (12) can be equivalently expressed as

$$\underset{\mathbf{w}_i, \mathbf{r}, \bar{\mathbf{w}}, \mathbf{v}}{\text{maximize}} \quad \left(\prod_{k=1}^{N} r_k\right)^{\frac{1}{N}} \tag{13a}$$

$$\text{s. t.} \quad \begin{cases} \bar{w}_k r_k - \bar{w}_k \leq |\mathbf{h}_k^H \mathbf{w}_k|^2 \\ \sum_{j=k+1}^{N} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \leq \bar{w}_k \end{cases} \quad k = 1, \ldots, N-1 \tag{13b}$$

$$r_k v_j - v_j \leq |\mathbf{h}_j^H \mathbf{w}_k|^2, \quad \sum_{m=k+1}^{N} |\mathbf{h}_j^H \mathbf{w}_m|^2 + \sigma^2 \leq v_j, \quad j = k+1, \ldots, N \tag{13c}$$

$$r_N - 1 \leq \frac{|\mathbf{h}_N^H \mathbf{w}_N|^2}{\sigma^2} \tag{13d}$$

$$\mathcal{T}(1, N), \ldots, \mathcal{T}(k, N), \ldots, \mathcal{T}(N, N) \text{ \& (11c)}. \tag{13e}$$

[3]It is not necessary to explicitly constrain the vector $\mathbf{r}$ to be positive, since for non-zero data rates this condition holds.

where $\mathbf{w}_i \in \mathbb{C}^T, \forall i, \mathbf{r} \in \mathbb{R}_+^N, \bar{\mathbf{w}} \in \mathbb{R}_+^{N-1}$, and $\mathbf{v} \in \mathbb{R}_+^{0.5(N^2-N)}$.

Proof: Please, refer to Appendix A. ∎

B. Approximation of the non-convex constraints

Next, we approximate the equivalent formulation in (13). To this end, note that, excluding the power constraint, the first set of constraints in (13b), (13c), and the constraints in (13d) and (13e) are all non-convex. The rest of the constraints are convex, and in fact admit SOC representation. Consider the second set of constraints in (13b) i.e.,

$$\sum_{j=k+1}^{N} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \leq \bar{w}_k \Leftrightarrow \left\| \begin{pmatrix} \mathbf{h}_k^H \mathbf{w}_{k+1} \\ \vdots \\ \mathbf{h}_k^H \mathbf{w}_N \\ \sigma \\ \frac{\bar{w}_k - 1}{2} \end{pmatrix} \right\|_2 \leq \frac{\bar{w}_k + 1}{2}, \quad k = 1, \ldots, N-1. \quad (14)$$

Similarly,

$$\sum_{m=k+1}^{N} |\mathbf{h}_j^H \mathbf{w}_m|^2 + \sigma^2 \leq v_j \Leftrightarrow \left\| \begin{pmatrix} \mathbf{h}_j^H \mathbf{w}_{k+1} \\ \vdots \\ \mathbf{h}_j^H \mathbf{w}_N \\ \sigma \\ \frac{v_j - 1}{2} \end{pmatrix} \right\|_2 \leq \frac{v_j + 1}{2}, \quad j = k+1, \ldots, N. \quad (15)$$

Now, in order to tackle the non-convex constraints, the so called CCP is used. The CCP has been widely used in neural computing [15], and has recently found applications in wireless signal processing [19], [20]. The CCP has also been referred to as minorization-maximization algorithm (MMA) [12], [14].

First, the procedure of handling the first set of non-convex constraints in (13b) is considered. The approximation of the other non-convex constraints closely follows the same technique. Consider the k^{th} constraint

$$\bar{w}_k r_k - \bar{w}_k \leq |\mathbf{h}_k^H \mathbf{w}_k|^2. \quad (16)$$

This is non-convex because of the bilinear term on the left side and the quadratic term on the right side of the inequality. An equivalent transformation of the above inequality is

$$\bar{w}_k r_k - \bar{w}_k \leq (\theta_{k,k}^i)^2 + (\theta_{k,k}^r)^2 = \|\boldsymbol{\Theta}_{k,k}\|_2^2, \quad \theta_{k,k}^r = \Re\left(\mathbf{h}_k^H \mathbf{w}_k\right), \theta_{k,k}^i = \Im\left(\mathbf{h}_k^H \mathbf{w}_k\right) \quad (17)$$

where $\boldsymbol{\Theta}_{k,k} = [\theta_{k,k}^r, \theta_{k,k}^i]^{\mathrm{T}}$ and $f(\boldsymbol{\Theta}_{k,k}) \triangleq |\mathbf{h}_k^H \mathbf{w}_k|^2$. Since the function in the right side of (17) is a convex one, it follows that [24]

$$f(\boldsymbol{\Theta}_{k,k}) = \|\boldsymbol{\Theta}_{k,k}\|_2^2 \geq \|\boldsymbol{\Theta}_{k,k}^t\|_2^2 + 2(\boldsymbol{\Theta}_{k,k}^t)^{\mathrm{T}}(\boldsymbol{\Theta}_{k,k} - \boldsymbol{\Theta}_{k,k}^t) \triangleq g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t), \quad (18)$$

where the right side of the inequality in (18) is the first order Taylor approximation of the function $\|\boldsymbol{\Theta}_{k,k}\|_2^2$ around $\boldsymbol{\Theta}_{k,k}^t$. Clearly, this formulation is linear in the variable $\boldsymbol{\Theta}_{k,k}$, and will be used instead of the original norm-squared function. Three important properties follow here

$$f(\boldsymbol{\Theta}_{k,k}) \geq g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t), \quad \text{for all } \boldsymbol{\Theta}_k \quad (19a)$$

$$f(\boldsymbol{\Theta}_{k,k}^t) = g(\boldsymbol{\Theta}_{k,k}^t, \boldsymbol{\Theta}_{k,k}^t), \quad (19b)$$

$$\nabla f(\boldsymbol{\Theta}_{k,k})|_{\boldsymbol{\Theta}_{k,k}^t} = \nabla g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t)|_{\boldsymbol{\Theta}_{k,k}^t} \quad (19c)$$

where the notation $(\cdot)|_{\boldsymbol{\Theta}_{k,k}^t}$ is used to represent the value of the function at $\boldsymbol{\Theta}_{k,k}^t$. The basic idea of the approximation algorithm presented below is to maximize the minorant $g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t)$ over the variable $\boldsymbol{\Theta}_{k,k}$, in order to obtain the next iterate term, $\boldsymbol{\Theta}_{k,k}^{t+1}$, i.e.,

$$\boldsymbol{\Theta}_{k,k}^{t+1} = \max_{\boldsymbol{\Theta}_{k,k}} g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t). \quad (20)$$

Using these considerations, it can be easily concluded that

$$f(\boldsymbol{\Theta}_{k,k}^{t+1}) = f(\boldsymbol{\Theta}_{k,k}^{t+1}) - g(\boldsymbol{\Theta}_{k,k}^t, \boldsymbol{\Theta}_{k,k}^{t+1}) + g(\boldsymbol{\Theta}_{k,k}^t, \boldsymbol{\Theta}_{k,k}^{t+1}) \quad (21)$$

$$\overset{(a)}{\geq} g(\boldsymbol{\Theta}_{k,k}^t, \boldsymbol{\Theta}_{k,k}^{t+1}) \overset{(b)}{\geq} g(\boldsymbol{\Theta}_{k,k}^t, \boldsymbol{\Theta}_{k,k}^t) \overset{(c)}{=} f(\boldsymbol{\Theta}_{k,k}^t), \quad (22)$$

where (a) follows from $f(\boldsymbol{\Theta}_{k,k}) \geq g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t)$, (b) is due to (20), and the final equality (c) is due to (19b).

Now, to deal with the bilinear product on the left side of (16), first we observe that for nonnegative $\bar{w}_k, r_k$ it holds that

$$\bar{w}_k r_k = \frac{1}{4}\left[(\bar{w}_k + r_k)^2 - (\bar{w}_k - r_k)^2\right]. \quad (23)$$

The quadratic term being subtracted in the above inequality can be well approximated by a first order Taylor series around $\bar{w}_k^t, r_k^t$. Thus, by combining the additional constraints in (17), the overall approximation of the constraint in (16) reads as

$$0.25(\bar{w}_k + r_k)^2 - \bar{w}_k - 0.25[(\bar{w}_k^t - r_k^t)^2 + 2(\bar{w}_k^t - r_k^t)\{\bar{w}_k - \bar{w}_k^t - r_k + r_k^t\}] \leq \|\boldsymbol{\Theta}_{k,k}^t\|_2^2 + 2(\boldsymbol{\Theta}_{k,k}^t)^{\mathrm{T}}(\boldsymbol{\Theta}_{k,k} - \boldsymbol{\Theta}_{k,k}^t) \triangleq g(\boldsymbol{\Theta}_{k,k}, \boldsymbol{\Theta}_{k,k}^t), \boldsymbol{\Theta}_{k,k} = [\theta_{k,k}^r, \theta_{k,k}^i]^{\mathrm{T}}, \theta_{k,k}^r = \Re\left(\mathbf{h}_k^H \mathbf{w}_k\right), \theta_{k,k}^i = \Im\left(\mathbf{h}_k^H \mathbf{w}_k\right) \quad (24)$$

which is convex in the variables of interest.

Following similar procedure the remaining non-convex constraints in (13c), (13d) and (13e) can be approximated as follows. The j-th constraint in (13c) and that in (13d) can be written as

$$0.25(r_k + v_j)^2 - v_j - 0.25[(r_k^t - v_j^t)^2 + 2(r_k^t - v_j^t)\{r_k - r_k^t - v_j + v_j^t\}] \leq \bar{g}(\boldsymbol{\Theta}_{j,k}, \boldsymbol{\Theta}_{j,k}^t) \quad (25)$$

$$\sigma^2(r_N - 1) \leq \bar{g}(\boldsymbol{\Theta}_{N,N}, \boldsymbol{\Theta}_{N,N}^t), \quad (26)$$

where

$$\boldsymbol{\Theta}_{j,k} = [\theta_{j,k}^r, \theta_{j,k}^i]^{\mathrm{T}}, \boldsymbol{\Theta}_{N,N} = [\theta_{N,N}^r, \theta_{N,N}^i]^{\mathrm{T}},$$
$$\bar{g}(\boldsymbol{\Theta}_{j,k}, \boldsymbol{\Theta}_{j,k}^t) = \|\boldsymbol{\Theta}_{j,k}^t\|_2^2 + 2(\boldsymbol{\Theta}_{j,k}^t)^{\mathrm{T}}(\boldsymbol{\Theta}_{j,k} - \boldsymbol{\Theta}_{j,k}^t),$$
$$\bar{g}(\boldsymbol{\Theta}_{N,N}, \boldsymbol{\Theta}_{N,N}^t) = \|\boldsymbol{\Theta}_{N,N}^t\|_2^2 + 2(\boldsymbol{\Theta}_{N,N}^t)^{\mathrm{T}}(\boldsymbol{\Theta}_{N,N} - \boldsymbol{\Theta}_{N,N}^t)$$

and $r_k^t, v_j^t, \boldsymbol{\Theta}_{j,k}^t, \boldsymbol{\Theta}_{N,N}^t$ represent the points around which the quadratic terms have been linearized.

Finally, the last set of non-convex constraints in (13e) can be tackled similarly. To demonstrate it, we linearize the first set of constraints in (44), i.e.,

$$|\mathbf{h}_k^H \mathbf{w}_N|^2 \leq \min_{m \in [1, N-1]} \tilde{g}(\boldsymbol{\Phi}_{k,m}, \boldsymbol{\Phi}_{k,m}^t), \quad (27)$$

TABLE I
NOMA/MISO SUM RATE MAXIMIZATION

given randomly generated $\mathbf{\Lambda}^0$ feasible to (11).
$t := 0$.
repeat
1- *Solve* (28) labelled as (Pb_t).
2- *Set* $\mathbf{\Lambda}^{t+1} = \mathbf{\Lambda}^t$.
3- *Update* $t := t + 1$.
until convergence or required number of iterations.

where
$\mathbf{\Phi}_{k,m} = [\phi^r_{k,m}, \phi^i_{k,m}]^{\mathrm{T}}, \phi^r_{k,m} = \Re(\mathbf{h}_k^H \mathbf{w}_m), \phi^i_{k,m} = \Im(\mathbf{h}_k^H \mathbf{w}_m)$
$\tilde{g}(\mathbf{\Phi}_{k,m}, \mathbf{\Phi}^t_{k,m}) = \|\mathbf{\Phi}^t_{k,m}\|_2^2 + 2(\mathbf{\Phi}^t_{k,m})^{\mathrm{T}}(\mathbf{\Phi}_{k,m} - \mathbf{\Phi}^t_{k,m})$
and $\mathbf{\Phi}^t_{k,m}$ is the linearization point. Based on this procedure, the notation $\bar{\mathcal{T}}(k^t, N^t)$ is used to represent the approximation of the form given in (27) for the remaining inequalities.

IV. The Proposed Solution

Having set up the stage as above, in this section the procedure that provides a tractable approximation to the sum rate maximization problem is outlined.

A. The Procedure

Using the above equivalent transformations and approximations, in the t^{th} iteration of the algorithm outlined in Table I, the following optimization problem is solved

$$\underset{\substack{\mathbf{w}_i, \mathbf{r}, \bar{\mathbf{w}}, \\ \mathbf{v}, \mathcal{A}}}{\text{maximize}} \quad \left(\prod_{k=1}^{N} r_k \right)^{\frac{1}{N}} \tag{28a}$$

(Pb_t) s. t. (14) & (24), $k = 1, \ldots, N-1$ (28b)

(15) & (25) $j = k+1, \ldots, N$, (26) (28c)

$\bar{\mathcal{T}}(1^t, N^t), \ldots, \bar{\mathcal{T}}(k^t, N^t), \ldots, \bar{\mathcal{T}}(N^t, N^t)$ & (11c), (28d)

where $\mathbf{w}_i \in \mathbb{C}^T, \forall i, \mathbf{r} \in \mathbb{R}_+^N, \bar{\mathbf{w}} \in \mathbb{R}_+^{N-1}, \mathbf{v} \in \mathbb{R}_+^{0.5(N^2-N)}$, and for all $j, k, m, \mathcal{A} \triangleq \{\mathbf{\theta}_{k,k} \in \mathbb{R}^{2N-2}, \mathbf{\theta}_{j,k} \in \mathbb{R}^{N^2-N}, \mathbf{\theta}_{N,N} \in \mathbb{R}^2, \mathbf{\Phi}_{k,m} \in \mathbb{R}^{2N^2-2N}\}$ represents the collection of all auxiliary variables. For the sake of notational convenience, all parameters about which the quadratic terms are linearized in iterate t are defined as

$$\mathbf{\Lambda}^t \triangleq [\bar{w}_k^t, r_k^t, v_j^t, \mathbf{\theta}^t_{k,k}, \mathbf{\theta}^t_{j,k}, \mathbf{\theta}^t_{N,N}, \mathbf{\Phi}^t_{k,m}]. \tag{29}$$

The MMA (CCP) algorithm used to solve (28) has been summarized in Table I. Note, that the convergence criteria can vary. For NOMA sum rate maximization, this algorithm terminates when the difference between two successive values of sum rate is less than a threshold. This aspect is discussed in more detail in section VI. Before concluding this section, we stress that (28) is an approximation, and need not to coincide with the global solution of (13).

B. Properties of the Proposed Algorithm

Before describing various characteristics of the algorithm presented above, let us define the feasible set, the objective and the set of optimization variables in the t^{th} iteration, respectively, as

$$\mathcal{F}_t = [\mathbf{w}_i \text{ for all } i, \mathbf{r}, \bar{\mathbf{w}}, \mathbf{v}, \mathcal{A} | \text{constraints in } (Pb_t) \text{ are satsified}] \tag{30}$$

$$\mathcal{O}_t = \max\left[(28a) | \{\mathbf{w}_i \text{ for all } i, \mathbf{r}, \bar{\mathbf{w}}, \mathbf{v}, \mathcal{A}\} \in \mathcal{F}_t\right] \tag{31}$$

$$\mathcal{V}_t = [\mathbf{w}_i \text{ for all } i, \mathbf{r}, \bar{\mathbf{w}}, \mathbf{v}, \mathcal{A}]. \tag{32}$$

1) Convergence:

Proposition 2: The sequence of variables $\{\mathcal{V}_t\}_{t \geq 0}$ is feasible i.e., it belongs to $\mathcal{F}_0$, where $\mathcal{F}_0$ is the feasibility set of the original problem (13).

Proof: Please, refer to Appendix B. ■

Proposition 3: The algorithm in Table I returns a non-decreasing sequence of objective values i.e., $\mathcal{O}_{t+1} \geq \mathcal{O}_t$, and hence it converges.

Proof: In order to prove this proposition, we note that $\mathcal{F}_{t+1} \supseteq \mathcal{F}_t$. From (22) it is clear that the surrogate functions used in place of non-convex terms are non-decreasing with iteration number i.e., $SF^{t+1} \geq SF^t$, where SF is a generic representation of these functions used in the paper and is valid for all of them. Therefore, $\mathcal{F}_{t+1} \supseteq \mathcal{F}_t$, is an immediate consequence, and the statement in Proportion 3 follows. Hence, $\{\mathcal{O}_t\}_{t \geq 0}$ is non-decreasing, and possibly converges to positive infinity. ■

In Proposition 3 the property (20) is used. It is important to note that the outcome of this proposition remains valid as long as the surrogate is increasing and does not rely on explicit maximization. The increasing behaviour of all SFs can be shown by following arguments similar to those outlined in [18], [19]. As a remark, we point out that when the feasibility set is convex and compact, the algorithm converges to a finite value.

2) KKT Conditions: Under a couple of technical assumptions the accumulation point of the algorithm satisfies the KKT conditions, as summarized in the proposition given below.

Proposition 4: As the iteration number t tends to infinity, the algorithm in Table I converges to the KKT point of (13).

Proof: Please, refer to Appendix C. ■

V. A Reduced Complexity Approximation

In the original sum rate function given in (9), it has been ensured that users with high SNRs are able to decode the messages of the weaker ones in the superposition coded signal, and hence apply SIC to remove interference from them. Optimal ordering of users depends upon physical parameters like, transmit antennas, precoding vectors etc. However, in certain situations the channel ordering alone may be sufficient to support the stronger users to decode the weaker ones. In scenarios, this is true, only the first term in the $\min(\cdot)$ function

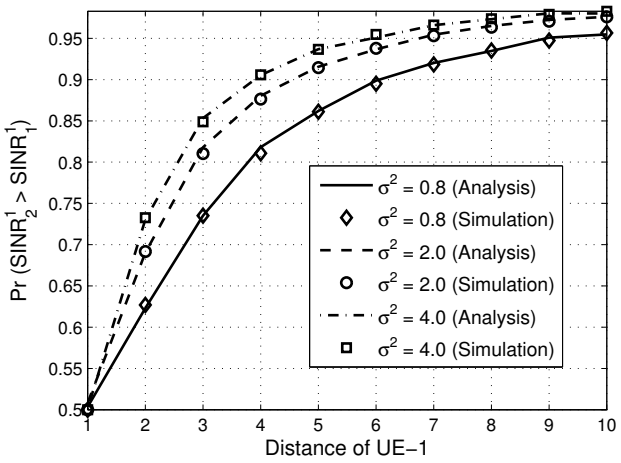


Fig. 2. Variation of $\Pr(\text{SINR}_2^1 > \text{SINR}_1^1)$ with the distance of UE-1. $T = 6, N = 4, \gamma = 2.0$ and distance of UE-2 $= 1$ is fixed.

for a user k needs to be retained and the objective becomes

$$R'_{sum} = \sum_{k=1}^{N-1} \log_2(1 + \text{SINR}_k^k) + \log_2\left(1 + \frac{|\mathbf{h}_N^H \mathbf{w}_N|^2}{\sigma^2}\right). \quad (33)$$

From (33) it can be seen that $N(N-1)/2$ SINR terms do not appear in the simplified sum rates. In turn, this means that in the formulation of (13), there are not $N^2 - N$ inequality constraints, and clearly, a complexity improvement is expected. Before moving on to the complexity analysis section, for completeness, the updated optimization problem solved in the t^{th} run of the algorithm in Table I, can be written as

$$\underset{\substack{\mathbf{w}_i, \mathbf{r}, \bar{\mathbf{w}}, \\ \mathcal{A}_p}}{\text{maximize}} \quad \left(\prod_{k=1}^{N} r_k\right)^{\frac{1}{N}} \quad (34a)$$

$$(Pb'_t) \qquad \text{s.t. (14) \& (24), } k = 1, \ldots, N-1, \text{ (26)} \quad (34b)$$

$$\bar{\mathcal{T}}(1^t, N^t), \ldots, \bar{\mathcal{T}}(k^t, N^t), \ldots, \bar{\mathcal{T}}(N^t, N^t) \text{ \& (11c)}, \quad (34c)$$

where $\mathbf{w}_i \in \mathbb{C}^T, \forall i, \mathbf{r} \in \mathbb{R}_+^N, \bar{\mathbf{w}} \in \mathbb{R}_+^{N-1}$, and now $\mathcal{A}_p \triangleq \{\boldsymbol{\theta}_{k,k}, \boldsymbol{\theta}_{N,N}, \boldsymbol{\phi}_{k,m}\}$ has a reduced cardinality compared to the original set of the variable set $\mathcal{A}$.

In order to provide more insight into the approximation used above let us consider the following lemma.

Lemma 1: Suppose that $\mathbf{h}_{k+1} = c_{k+1}\mathbf{h}_k, k = 1, \ldots, N-1$, so that $\mathbf{h}_n = c_n c_{n-1} \ldots c_{k+1}\mathbf{h}_k$, where $k+1 \leq n \leq N$ and the magnitudes of the complex constants $c_n c_{n-1} \ldots c_{k+1} \triangleq c_n^{k+1}$ is greater than one. Under this assumption, when (9) reduces to (33), then

$$\|\mathbf{h}_1\|_2 < \|\mathbf{h}_2\|_2 \ldots < \|\mathbf{h}_N\|_2. \quad (35)$$

Proof: Please, refer to Appendix D. ∎

From Lemma 1 it can be expected that, at least approximately, when that channels are clearly ordered, i.e., the magnitudes of successive channel vectors differ significantly and the channel ratio inequalities as given above are satisfied, the problems in (34) and (28) are equivalent. To further highlight, we evaluate below the probability of an event of interest.

Lemma 2: Consider a random unitary precoding matrix, i.e., $\mathbf{W}^H\mathbf{W} = \mathbf{I}$,[4] where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$ and $\mathbf{W}$ is independent of the channel matrices. For $i \geq j$

$$\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right) \quad (36)$$

is given by

$$\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right) = 1 - e^{(\lambda_i+\lambda_j)\sigma^2}\lambda_i\sigma^2 \psi\left((\lambda_i+\lambda_j)\sigma^2, 2(N-k)\right) - e^{(\lambda_i+\lambda_j)\sigma^2}(N-k) \psi\left((\lambda_i+\lambda_j)\sigma^2, 2(N-k)+1\right), \quad (37)$$

where

$$\psi(\lambda, m) = (-1)^m \frac{\lambda^{m-1}\text{Ei}(-\lambda)}{(m-1)!} + e^{-\lambda}\sum_{l=0}^{m-2}\frac{(-1)^l\lambda^l}{(m-1)\cdots(m-1-l)} \quad (38)$$

and $\text{Ei}(x)$ is the exponential integral [25].

Proof: Please, refer to Appendix E. ∎

When $\lambda_j >> \lambda_i$, (37) can be approximated as

$$\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right) \approx 1 - e^{\lambda_j\sigma^2}\lambda_i\sigma^2\psi\left(\lambda_j\sigma^2, 2(N-k)\right) - e^{\lambda_j\sigma^2}(N-k)\psi\left(\lambda_j\sigma^2, 2(N-k)+1\right). \quad (39)$$

It can be seen from (39) that when d_i decreases (and thus λ_i decreases), $\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right)$ increases as well. Hence, the stronger the channel of UE-i compared to UE-j, the higher the probability $\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right)$. In order to further investigate the probability under consideration, in Fig. 2, the variation of $\Pr(\text{SINR}_2^1 > \text{SINR}_1^1)$ is depicted in terms of the distance of UE-1. With a decrease in the channel strength of UE-1, $\Pr(\text{SINR}_2^1 > \text{SINR}_1^1)$ increases, thereby justifying the use of SINR_1^1 instead of SINR_2^1. The probability $\Pr(\text{SINR}_2^1 > \text{SINR}_1^1)$ varies inversely with the noise variance for a given UE-1 distance. It can also be seen that for higher values of noise variance, the interference term dominates and the probability that SINR_1^1 remains below SINR_2^1 is increased. In addition, this figure also validates the analytical results derived above.

A. Complexity

In each iteration of the procedure presented in Table I, we solve an SOCP. The total number of iterations are fixed and only variables are updated in each run of the algorithm. Hence, the worst case regarding the complexity is determined by the SOCP in each step. Therefore, to provide a complexity estimate, the worst case complexity of the SOCP given by (28) or (34) is estimated. It is well known that for general interior-point methods the complexity of the SOCP depends upon the number of constraints, variables and the dimension of each SOC constraint [23]. The total number of constraints in the formulations of (28) and (34) are $0.5N^3 + 0.5N^2 + 2N + c$ and $0.5N^3 - 0.5N^2 + 3N + c$, respectively, where the non-negative integer constant, c, refers to the SOC constraints with different

[4]Extending this lemma without the orthogonality constraint remains an open problem.

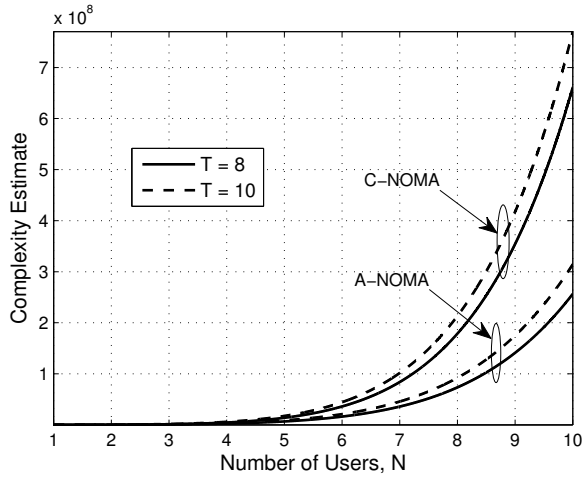


Fig. 3. Variation of the per iteration complexity of the exact and approximate NOMA formulations with the number of users N for given values of T and $c = 10$.

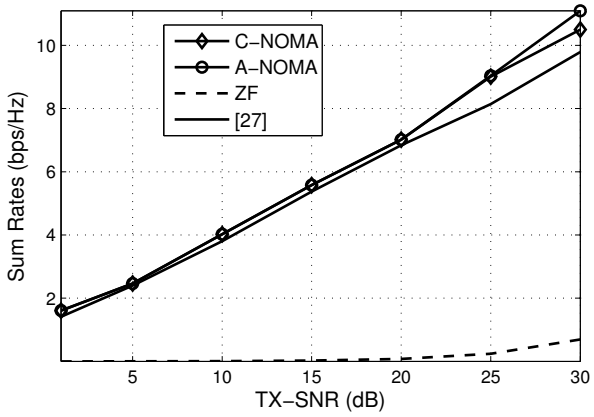


Fig. 4. Achievable sum rates vs. normalized transmit power called TX-SNR. We take $T = N = 3, D_0 = 50, \gamma = 2$ and $\sigma = 2$.

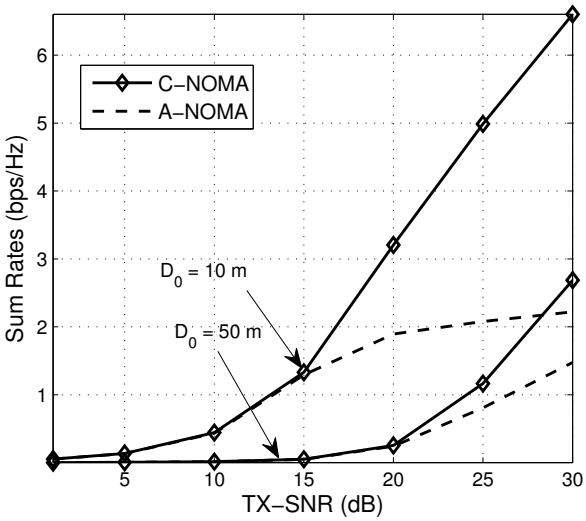


Fig. 5. Variation of average sum rates with TX-SNR for different D_0. $N = T = 4, \sigma = 1$ and $\gamma = 2.0$.

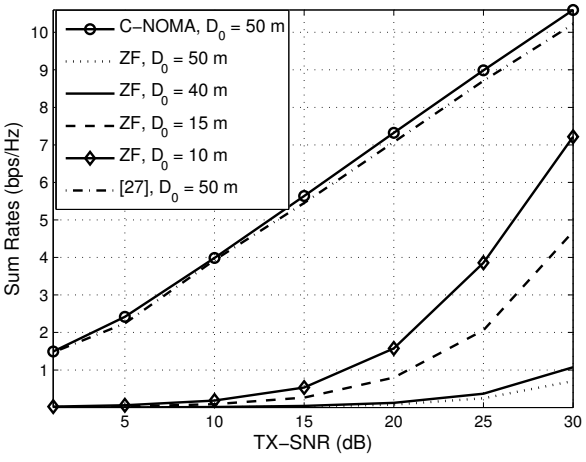


Fig. 6. The achievable sum rates as a fucntion of TX-SNR for different values of D_0. The parameters taken are $N = T = 4, \sigma = 1$ and $\gamma = 2.0$.

N. This happens because of the equivalent SOC representation of the geometric mean, given in the objective function, also see [23]. Therefore, for both problems the number of iterations needed to reduce the duality gap to a small constant is upper bounded by $\mathcal{O}(\sqrt{0.5N^3 + 0.5N^2 + 2N + c})$ and $\mathcal{O}(\sqrt{0.5N^3 - 0.5N^2 + 3N + c})$, respectively [23]. In order to calculate the dimension of all SOCs in (28) we provide an upper bound because the sums of the dimensions for some constraints have been bounded from above by definite integrals of increasing functions. This estimate is found to be $\lceil 1.833N^3 + 3N^2 + 8N + NT + 3c - 5.83333 \rceil$ for (28). The interior-point method's per iteration worst case complexity estimate of (28) is $\mathcal{O}((3.5N^2 + 1.5N + 2NT + c - 1)^2(\lceil 1.833N^3 + 3N^2 + 8N + NT + 3c - 5.83333 \rceil))$, where $3.5N^2 + 1.5N + 2NT + c - 1$ is the number of optimization variables in (28). Likewise, the interior-point method's per iteration complexity to solve the SOCP in (34) is given by $\mathcal{O}((2N^2 + 3N + 2NT + c - 1)^2(1.5N^3 - N^2 + 10.5N + NT + 3c - 4))$, where $2N^2 + 3N + 2NT + c - 1$ and $1.5N^3 - N^2 + 10.5N + NT + 3c - 4$ are the optimization variables and the total dimension of the SOC constraints in (34).

To provide further insight, we plot the per iteration complexity estimates of the SOCPs in Fig. 3. The SOCP in (28) is called complete NOMA (C-NOMA), while the one in (34) is dubbed as approximate NOMA (A-NOMA). The figure quantifies the increase in the complexity as a function of both N and T.

VI. Numerical Results

In this section we investigate the performance of the proposed solution to the NOMA sum rate maximization problem. For a given set of antennas T and users N, the channels as $\mathbf{h}_i = \sqrt{d_i^{-\gamma}}\mathbf{g}_i$ are generated, where $\mathbf{g}_i \sim \mathcal{CN}(0, \mathbf{I})$, and the distances of all users are fixed, such that they are equally spaced between distances of 1 and D_0 from the BS. It should be noted here that in simulations the user distances are fixed and the average is taken over the fast fading component of

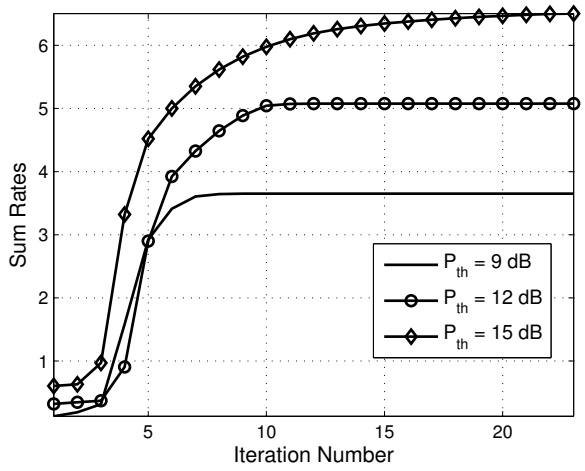


Fig. 7. Iterations required for convergence in C-NOMA approach. We take $T = N = 5, \sigma = 1, D_0 = 10$ and $\gamma = 2$.

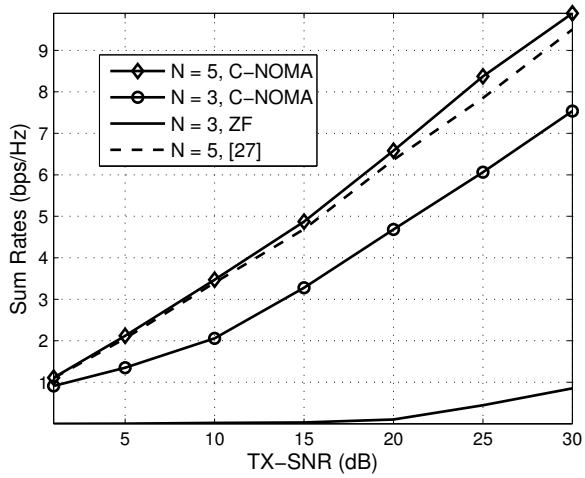


Fig. 8. The effect of TX-SNR on achievable average sum rates when $N > T$. The parameters used are $T = 3, \sigma = 1, \gamma = 2.0$ and $D_0 = 50$.

the channel vectors. For each set of results the values of γ and D_0 are mentioned, while it is assumed that $\sigma = 1$ for all users. Similarly, the transmit power is normalized with respect to noise, whose variance is taken to be unity. For the simulations the CVX package [26] is used. For sum rates comparison, in addition to traditional techniques, we will also use the approach developed in Christensen *et al.* [27].

In Fig. 4, we plot the average sum rates versus the transmit power for a three user system and a BS equipped with three antennas. We take $\gamma = 2, D_0 = 50$, and therefore, the three users are placed at $1, 25.5$ and 50 meters from the BS, respectively. Unless specifically pointed out, γ and D_0 retain the same value. It is noted that for transmit power up to 25 dB, the sum rates of the complete NOMA (C-NOMA) formulation and the its approximation (A-NOMA) are equal. This observation is because of the distance effect, the ordering of the channels $\|\mathbf{h}_1\|_2 \leq \|\mathbf{h}_2\|_2 \leq \|\mathbf{h}_3\|_2$ is valid for all realizations of $\mathbf{g}_i$. As a consequence, $\text{SINR}_1^1 < \min\left(\text{SINR}_2^1, \text{SINR}_3^1\right)$, and $\text{SINR}_2^2 < \text{SINR}_3^2$. Therefore, the objective function in (9) matches with that in (33). Because of the wide range of multiplicative distance factor, this observation can be attributed as a result of the Lemma 1. Several other factors like Gaussian noise, transmit SNR, etc. can also influence the results. For example, once the transmit power crosses a certain value (25 dB in our case), the ordering of users need not to be the optimal one and hence the two curves deviate from each other. The A-NOMA approach produces better rates because the interference free rates of the last user are boosted more compared to the C-NOMA. This comes with a degradation in the sum rates of the $N-1$ users (excluding UE-N) as we will see in the next experiment. Interestingly, the competing zero-forcing (ZF) solution performs very poorly for lower SNRs, and only produces significant sum rates, when the transmit power is sufficiently high. This poor performance of the ZF scheme can be attributed to the distance effect, which makes the channel matrix poorly-conditioned [28]. At higher transmit SNRs this poor condition of the channel matrix is partially circumvented and hence a notable increase in ZF rates is observed. In Fig. 4, we have also plotted the sum rates obtained using the technique proposed in [27]. The curve lies below the sum rates obtained by the proposed algorithm. Although the approach in [27] is numerical in nature, the higher data rates of the proposed algorithm in paper may be attributed to the nature of the cost function in which the strongest user sees no interference, while this does not hold true in the case of the cost function in [27].

The next set of results presented in Fig. 5 depict the average sum rates of all users excluding UE-N as a function of transmit power, with $N = T = 4$. Basically, Fig. 5 can be seen as complementing the observations made in Fig. 4, where also at high transmit SNR A-NOMA has better total sum rates, compared to C-NOMA. It is seen that for low SNRs the curves for A-NOMA and C-NOMA overlap. As the the transmit power is further boosted, C-NOMA outperforms A-NOMA. The reason for the equality of the rates in both techniques is the same as mentioned above. However, at higher transmit SNR the C-NOMA provides better data rates, because of lack of optimality in the users' ordering, the beamformers of A-NOMA will not necessarily produce optimum $\min\limits_{k \leq j \leq N} (\text{SINR}_k^j)$ for all k. In addition, we have also included curves, when D_0 is decreased from 50 to 10 meters. It is evident that because of the shorter distance the net effect of distance attenuation, which orders the channels, is diminished. Hence, the gap between the graphs of C-NOMA and A-NOMA is enlarged. Nonetheless, overall higher data rates are reported in this case because of better channel conditions for the users.

It can be concluded from the previous discussions, that distance plays an important role in determining the aggregate data rates of the NOMA system. Therefore, to further explore its impact we set $N = T = 4$ and plot the curves for the sum rates of C-NOMA and ZF, with $\gamma = 2$. The sum rates of the ZF scheme are shown in Fig. 6 as the distance D_0 is decreased from 50 m to 10 m. As the distance is decreased, the effect of path loss is minimized and we have better conditioned channel matrices. Therefore, the sum rates of ZF are considerably

enhanced at $D_0 = 10$ m. We have also compared our rates with those obtained using [27] in Fig. 6.

In order to investigate the convergence of the proposed algorithm, we consider a downlink system with $T = 5$ antennas, serving $N = 5$ users. As a stopping criteria, we use successive values of the sum rate returned by the algorithm. The algorithm exits from the main sequential iteration loop, when the difference between two consecutive values of the sum rate is less than or equal to 10^{-2}. With this criterion, as shown in Fig. 7, the algorithm converges within 25 iterations for the three values of transmit SNR shown in the figure. Moreover, as expected, with higher transmit power, we obtain better sum rate.

As a multiuser system is considered, the proposed approach is expected to deliver acceptable spectral efficiency when $N > T$. The results reported in Fig. 8 show the performance of C-NOMA, when the number of users N is greater than the number of transmit antennas $T = 3$. For comparison we have also included the sum rates achieved by the C-NOMA and ZF solutions with $N = 3$ users only. To obtain these two curves, we randomly pick three users to be served with C-NOMA and ZF precoders. It is evident that with fewer users C-NOMA underperforms. Since, in this case, users are randomly chosen, it is likely that the effective multiuser diversity [29] is lost and we see a downward trend in achievable data rates. The lower data rates of [27] can be attributed to the reason mentioned above in the description of Fig. 4.

In order to investigate the decoding capability of users placed higher in the order, consider a three user system such that the transmitter has three antennas. The quantity $\kappa \triangleq \log\left(\frac{\text{SINR}_3^1}{\text{SINR}_1^1}\right)$, indicates the capability of the third user (strongest user in the given ordering) to decode the first user (weakest user in the given ordering). We calculate the average value of κ, first when the third user is at a distance of 1 m from the base station, and later, when it is moved away to a distance of 5 m from the base station. The variation of average value of κ with transmit SNR is shown in Table II. It is seen that with an increase in the transmit SNR and the distance of user three from the transmitter, κ decreases. This observation coupled with the SINR expressions in (5), (6) and (7), shows that distance, transmit power and other factors that appear in these expressions control whether $\min(\text{SINR}_1^1, \text{SINR}_2^1, \text{SINR}_3^1) = \text{SINR}_1^1$, holds or not. For instance, when the third user is sufficiently far from the base station, or the value of the noise variance is such that we may have $\text{SINR}_3^1 < \text{SINR}_1^1$, the C-NOMA and the A-NOMA will not be equal anymore. This fact was observed earlier in Fig. 4, where at a transmit SNR of 25 dB, the curves for the C-NOMA and A-NOMA deviated from each other. Then, clearly due to a bigger feasible set, the A-NOMA produced better rates, albeit with no guarantee on performing SIC. In summary, the results reported in Fig. 4 indicate that the SIC capability of a higher order user depends on several factors. Therefore, for a given scenario, knowing an optimal user ordering is indeed an important problem to explore.

TABLE II
AVERAGE κ VERSUS TX-SNR

TX-SNR dB	κ UE-3 at 1 m	κ UE-3 at 5 m
1.0	6.6516	4.2409
5.0	6.0522	4.0985
10.0	4.5938	3.5091
15.0	3.0706	2.4917
20.0	1.6947	1.4773
25.0	0.6991	0.566
30.0	0.0085	0.0018

VII. CONCLUSION

In this paper, we have studied the sum rate maximization problem of a MISO downlink system based on NOMA. Specifically, we approximate the originally non-convex optimization problem with a MM method. For the proposed algorithm, we have solved an SOCP with polynomial computational complexity in each step. For the scenarios considered, the algorithm is numerically shown to converge within a few iterations. Furthermore, we developed a reduced complexity approximation and explore the conditions under which it is tight. Finally, we provide an insight into the tightness of the proposed approximation. Our experimental results reveal that the NOMA has a superior performance compared to conventional orthogonal multiple access schemes. High data rates are obtained with small transmit power. The distance attenuation has a very low impact on NOMA performance. NOMA particularly outperforms ZF when the number of users is higher than the transmit antennas, thus making it an ideal candidate for enabling multiple access in the next generation 5G networks.

APPENDIX A
PROOF OF PROPOSITION 1

In order to show the equivalence of (12) with (13), the original formulation in (12) is *factored* into several different constraints, and so, these factors can be processed individually. We first focus on the constraints in (12b), and then move to the remaining intractable constraints. Without loss of generality, it holds that

$$r_k - 1 \leq \min\left(\text{SINR}_k^k, \ldots, \text{SINR}_N^k\right) \Leftrightarrow$$
$$r_k - 1 \leq \begin{cases} \text{SINR}_k^k \\ \min\left(\text{SINR}_{k+1}^k, \ldots, \text{SINR}_N^k\right), \end{cases} \tag{40}$$

for $k = 1, \ldots, N-1$. The constraint in (12b) has been purposely written as that in (40), since the first term SINR_k^k is different from the remaining ones. Hence, it is necessary to deal with the first term and the remaining $N - k$ terms passed as argument of the $\min(\cdot)$ function. By introducing,

$\bar{\mathbf{w}} \in \mathbb{R}_+^{N-1}$, it holds that

$$r_k - 1 \leq \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j=k+1}^N |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2} \Leftrightarrow \begin{cases} \bar{w}_k r_k - \bar{w}_k \leq |\mathbf{h}_k^H \mathbf{w}_k|^2 \\ \sum_{j=k+1}^N |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \leq \bar{w}_k, \end{cases} \quad (41)$$

where $\bar{w}_k$ is the k^{th} component of the vector $\bar{\mathbf{w}}$, and the expression of SINR_k^k is used. Likewise, for an arbitrary SINR_j^k, $k+1 \leq j \leq N$, belonging to the remaining terms in the $\min(\cdot)$ function, we introduce the new variable, $\mathbf{v} \in \mathbb{R}_+^{0.5(N^2-N)}$, and write the corresponding constraint as the following system of inequalities

$$r_k v_j - v_j \leq |\mathbf{h}_j^H \mathbf{w}_k|^2, \quad \sum_{m=k+1}^N |\mathbf{h}_j^H \mathbf{w}_m|^2 + \sigma^2 \leq v_j, \quad (42)$$

where v_j is the j^{th} element of $\mathbf{v}$. Note, that even if the constraints in (12b) have been transformed, the problem remains intractable. From the inequalities in (11b), it holds that

$$|\mathbf{h}_k^H \mathbf{w}_1|^2 \geq \ldots |\mathbf{h}_k^H \mathbf{w}_{k-1}|^2 \geq |\mathbf{h}_k^H \mathbf{w}_k|^2 \geq |\mathbf{h}_k^H \mathbf{w}_{k+1}|^2 \ldots \geq |\mathbf{h}_k^H \mathbf{w}_N|^2 \quad (43)$$

$$\Leftrightarrow \begin{cases} |\mathbf{h}_k^H \mathbf{w}_N|^2 \leq \min_{m\in[1,N-1]} |\mathbf{h}_k^H \mathbf{w}_m|^2 \\ \ldots \\ |\mathbf{h}_k^H \mathbf{w}_{k+1}|^2 \leq \min_{m\in[1,k]} |\mathbf{h}_k^H \mathbf{w}_m|^2 \triangleq \mathcal{T}(k,N) \\ \ldots \\ |\mathbf{h}_k^H \mathbf{w}_2|^2 \leq |\mathbf{h}_k^H \mathbf{w}_1|^2. \end{cases} \quad (44)$$

Similarly, equivalent transformations, $\mathcal{T}(1,N)$ and $\mathcal{T}(N,N)$ for $k = 1, N$ can be obtained. Now putting together all transformations developed above, we arrive at the equivalent transformation given in (13).

B Proof of Proposition 2

Without loss of generality, we focus on the function $f(\Theta_{k,k})$, its approximation $g(\Theta_{k,k}, \Theta_{k,k}^t)$ and the constraint in which it appears. The same arguments will be applicable to all non-convex functions, their convex minorants and the respective constraints. Therefore, it holds that

$$0.25(\bar{w}_k + r_k)^2 - \bar{w}_k - g(\bar{w}_k, r_k, \bar{w}_k^t, r_k^t) \leq g(\Theta_{k,k}, \Theta_{k,k}^t), \quad (45)$$

where $g(\bar{w}_k, r_k, \bar{w}_k^t, r_k^t) \triangleq 0.25\left[(\bar{w}_k^t - r_k^t)^2 + 2(\bar{w}_k^t - r_k^t)\{\bar{w}_k - \bar{w}_k^t - r_k + r_k^t\}\right]$ is the approximation of the original function $(\bar{w}_k - r_k)^2$. Note, that this constraint is a convex approximation of that in (16). Now, let us assume that the tuple $(\bar{w}_k^t, r_k^t, \Theta_{k,k}^t)$ is feasible to (16). Clearly, the same point also satisfies (45) as a consequence of (19b). Since $g(\bar{w}_k, r_k, \bar{w}_k^t, r_k^t) \leq (\bar{w}_k - r_k)^2$ and $f(\Theta_{k,k}) \geq g(\Theta_{k,k}, \Theta_{k,k}^t)$, it follows that

$$0.25(\bar{w}_k + r_k)^2 - \bar{w}_k - 0.25(\bar{w}_k - r_k)^2 - f(\Theta_{k,k}) \quad (46)$$

$$\leq 0.25(\bar{w}_k + r_k)^2 - \bar{w}_k - g(\bar{w}_k, r_k, \bar{w}_k^t, r_k^t) - g(\Theta_{k,k}, \Theta_{k,k}^t). \quad (47)$$

Hence, $(\bar{w}_k^{t+1}, r_k^{t+1}, \Theta_{k,k}^{t+1})$ should satisfy (16) because

$$0.25(\bar{w}_k^{t+1} + r_k^{t+1})^2 - \bar{w}_k^{t+1} - 0.25(\bar{w}_k^{t+1} - r_k^{t+1})^2 - f(\Theta_{k,k}^{t+1}) \quad (48)$$

$$\leq 0.25(\bar{w}_k^{t+1} + r_k^{t+1})^2 - \bar{w}_k^{t+1} - g(\bar{w}_k^{t+1}, r_k^{t+1}, \bar{w}_k^t, r_k^t) - g(\Theta_{k,k}^{t+1}, \Theta_{k,k}^t) \leq 0. \quad (49)$$

The above conclusion holds for all k and $\{\mathcal{V}_t\}_{t\geq 0}$, as the algorithm was initialized with $\boldsymbol{\Lambda}^{(0)} \in \mathcal{F}_0$.

C Proof of Proposition 4

The following assumptions are made before outlining the arguments.

Assumption 1: We assume that as $t \to \infty$, the sequence of variables $\{\mathcal{V}_t\}_{t\geq 0}$ generated by the algorithm in Table I converges to a value $\mathcal{V}^*$.

Assumption 2: The constraints in the approximate problem (28) or (34) are qualified at the accumulation point.

Without explicitly mentioning the constraints, we use abstract notation to prove the claim made in Proposition 4. First let us give a generic representation to all convex constraints in (28) as $\mathcal{C}_a(\mathcal{V})_s \leq 0, a = 1, \ldots, L_1$, where $(\mathcal{V})_s$ denotes the subset of $\mathcal{V}_t$ containing the corresponding variables that appear in these constraints. Similarly, let us define as $\mathcal{C}_b^t(\mathcal{V})_p \leq 0, b = L_1 + 1, \ldots, L_2$ the constraints obtained by approximating the non-convex functions with convex minorants in (28), and $(\mathcal{V})_p \subseteq \mathcal{V}_t$. Let $\eta_a^*, \bar{\eta}_b^* \in \mathbb{R}_+$ for all a, b, denote the dual variables at convergence. The KKT conditions of the problem in (28) at $(\mathcal{V}^*)_s, (\mathcal{V}^*)_p$ then read as

$$\nabla \mathbf{r}^* + \sum_{a=1}^{L_1} \eta_a^* \nabla \mathcal{C}_a(\mathcal{V}^*)_s + \sum_{b=L_1+1}^{L_2} \bar{\eta}_b^* \nabla \mathcal{C}_b^t(\mathcal{V}^*)_p = 0 \quad (50)$$

$$\eta_a^* \mathcal{C}_a(\mathcal{V}^*)_s = 0, \ a = 1, \ldots, L_1, \quad \bar{\eta}_b^* \mathcal{C}_b^t(\mathcal{V}^*)_p = 0, \ b = L_1 + 1, \ldots, L_2. \quad (51)$$

Since all convex minorants satisfy the properties in (19), it is easy to conclude that the KKT conditions given above will reduce to those of the problem in (28). Similar conclusion also holds for the simplified problem in (34).

D Proof of Lemma 1

For (33) to be valid for all $1 \leq k \leq N-1$, it holds that

$$\text{SINR}_k^k < \min_{i\in[k+1,N]} \text{SINR}_i^k. \quad (52)$$

For an arbitrary $k \in [1, N-1]$ and $i = n$, let us consider the following inequality,

$$\frac{|\mathbf{h}_n^H \mathbf{w}_k|^2}{\sum_{m=k+1}^N |\mathbf{h}_n^H \mathbf{w}_m|^2 + \sigma_n^2} > \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{m=k+1}^N |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_k^2}, \quad (53)$$

where we have assumed that the noise variances at the n^{th} and the k^{th} nodes are σ_n^2 and σ_k^2, respectively. By substituting the assumptions made in the lemma, $\mathbf{h}_n = c_n c_{n-1} \ldots c_{k+1} \mathbf{h}_k \triangleq$

$c_n^{k+1}\mathbf{h}_k$, where $k+1 \leq n \leq N$. After some simple manipulations

$$\sum_{m=k+1}^{N} |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_k^2 > \sum_{m=k+1}^{N} |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_n^2/|c_n^{k+1}|^2 \Leftrightarrow |c_n^{k+1}| > \frac{\sigma_n}{\sigma_k}. \quad (54)$$

Now, if $\sigma_n = \sigma_k$, using the condition on $|c_n^{k+1}|$, we obtain $\|\mathbf{h}_n\|_2 > \|\mathbf{h}_k\|_2$ for all n. Repeating the same argument for all k, the required proof follows.

E Proof of Lemma 2

The SINR_i^k, $i \geq k$, can be written as

$$\text{SINR}_i^k = \frac{|\mathbf{h}_i^H \mathbf{w}_k|^2}{\sum_{m=k+1}^{N} |\mathbf{h}_i^H \mathbf{w}_m|^2 + \sigma^2}. \quad (55)$$

If a random unitary matrix is used for precoding, $|\mathbf{h}_i^H \mathbf{w}_k|^2$ is still complex Gaussian distributed, since a unitary transformation of Gaussian vectors is still complex Gaussian distributed. In addition, $|\mathbf{h}_i^H \mathbf{w}_k|^2$ and $|\mathbf{h}_i^H \mathbf{w}_l|^2$, $k \neq l$ are independent. Define $x_{ik} \triangleq |\mathbf{h}_i^H \mathbf{w}_k|^2$ and $y_{ik} \triangleq \sum_{m=k+1}^{N} |\mathbf{h}_i^H \mathbf{w}_m|^2$. Therefore x_{ik} is an exponentially distributed random variable, with $\lambda_i \triangleq d_i^{\gamma}$, i.e., $f_{x_{ik}}(x) = \lambda_i e^{-\lambda_i x}$. Similarly, y_{ik} follows the Chi-square distribution, i.e.,

$$f_{y_{ik}}(y) = \frac{\lambda_i^{N-k} y^{(N-k-1)}}{(N-k-1)!} e^{-\lambda_i y}. \quad (56)$$

Consequently the cumulative distribution function of SINR_i^k can be calculated from the following

$$\Pr\left(\text{SINR}_i^k \leq \theta\right) = \Pr\left(\frac{x_{ik}}{y_{ik}+\sigma^2} \leq \theta\right) \quad (57)$$

$$= \int_0^{\infty} \left(1 - e^{-\lambda_i \theta\left(y+\sigma^2\right)}\right) f_{y_{ik}}(y) dy \quad (58)$$

$$= 1 - \frac{e^{-\lambda_i \theta \sigma^2}}{(N-k-1)!} \times \int_0^{\infty} e^{-(1+\theta)\lambda_i y} (\lambda_i y)^{(N-k-1)} d\lambda_i y. \quad (59)$$

Applying [25, Eq. (3.351.3)], the pdf of SINR_i^k can be obtained as follows:

$$F_{\text{SINR}_i^k}(z) = 1 - \frac{e^{-\lambda_i \sigma^2 z}}{(1+z)^{N-k}}. \quad (60)$$

Again, following the unitary transformation of Gaussian variables, the desired probability can be evaluated as

$$\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right) = \int_0^{\infty} \left(1 - \frac{e^{-\lambda_j \sigma^2 z}}{(1+z)^{N-k}}\right) f_{\text{SINR}_i^k}(z) dz \quad (61)$$

$$= 1 - \int_0^{\infty} \left(\frac{\lambda_i \sigma^2 e^{-(\lambda_i+\lambda_j)\sigma^2 z}}{(1+z)^{2(N-k)}} + \frac{(N-k) e^{-(\lambda_i+\lambda_j)\sigma^2 z}}{(1+z)^{2(N-k)+1}}\right) dz. \quad (62)$$

Applying [25, Eq. (3.351.4)], the above probability can be expressed as

$$\Pr\left(\text{SINR}_i^k > \text{SINR}_j^k\right) = 1 - e^{(\lambda_i+\lambda_j)\sigma^2} \lambda_i \sigma^2 \psi\left((\lambda_i+\lambda_j)\sigma^2, 2(N-k)\right) - e^{(\lambda_i+\lambda_j)\sigma^2} (N-k) \psi\left((\lambda_i+\lambda_j)\sigma^2, 2(N-k)+1\right), \quad (63)$$

and the proof is completed.

References

[1] "DOCOMO 5G WHITE PAPER, 5G radio access: requirements, concept and technologies," NTT DOCOMO, INC., Tech. Rep., July 2014.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[3] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 66–70.

[4] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[5] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. URSI International Symposium on Signals, Systems, and Electronics*, Sep. 1998, pp. 295–300.

[6] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a DS/CDMA system," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 5, pp. 796–807, June 1994.

[7] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[8] Y. Lan, A. Benjebboiu, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop SU-MIMO," in *Proc. IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2014, pp. 1–5.

[9] X. Chen, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Impact of rank optimization on downlink non-orthogonal multiple access (NOMA) with SU-MIMO," in *Proc. IEEE International Conference on Communication Systems (ICCS)*, Nov. 2014, pp. 233–237.

[10] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[11] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, (submitted). [Online]. Available: http://arxiv.org/abs/1503.05367

[12] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[13] D. R. Hunter and R. Li, "Variable selection using MM algorithms," *Annals of Statistics*, vol. 33, no. 4, pp. 1617–1642, 2005.

[14] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[15] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[16] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables," in *Proc. Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[17] P. Stoica and Y. Selen, "Cyclic minimizers, majorization techniques, and the expectation-maximization algorithm: a refresher," *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 112–114, Jan. 2004.

[18] A. Beck, A. B. Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Opt.*, vol. 47, no. 1, pp. 29–51, 2010.

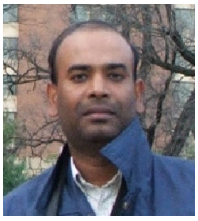
[19] M. F. Hanif, L.-N. Tran, A. Tolli, M. Juntti, and S. Glisic, "Efficient solutions for weighted sum rate maximization in multicellular networks with channel uncertainties," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5659–5674, Nov. 2013.

[20] L.-N. Tran, M. F. Hanif, and M. Juntti, "A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 114–117, Jan. 2014.

[21] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Multicast multigroup beamforming for per-antenna power constrained large-scale arrays," in *Proc. IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2015, pp. 271–275.
[22] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.
[23] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Applications, Special Issue on Linear Algebra in Control, Signals and Image Processing*, pp. 193–228, Nov. 1998.
[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U. K.: Cambridge University Press, 2004.
[25] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. U. S.: Academic Press, 2007.
[26] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx/, Feb. 2015.
[27] S. S. Christensen, R. Agarwal, E. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
[28] A. F. Molisch, *Wireless Communications*, 2nd ed. New York, NY, USA: Wiley, 2010.
[29] 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.

Muhammad Fainan Hanif received the Ph.D. degree in electrical and electronic engineering from the University of Canterbury, Christchurch, New Zealand, in 2011. Later, he was as a postdoctoral researcher at the Centre for Wireless Communications, Department of Communications Engineering, University of Oulu, Finland. Since Feb. 2015, he has been working as a Research Associate within the Computing and Communications Department of the Lancaster University, United Kingdom. His research interests include convex optimization applications in signal processing and modern communication systems.

Zhiguo Ding (S'03-M'05) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005. From Jul. 2005 to Aug. 2014, he was working in Queen's University Belfast, Imperial College and Newcastle University. Since Sept. 2014, he has been with Lancaster University as a Chair Professor. Dr Ding's research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Networks*, *IEEE Wireless Communication Letters*, *IEEE Communication Letters*, and *Journal of Wireless Communications and Mobile Computing*. He received the best paper award in IET Comm. Conf. on Wireless, Mobile and Computing, 2009, IEEE Communication Letter Exemplary Reviewer 2012, and the EU Marie Curie Fellowship 2012-2014.



Tharmalingam Ratnarajah (A'96-M'05-SM'05) is currently with the Institute for Digital Communications, University of Edinburgh, Edinburgh, UK, as a Professor in Digital Communications and Signal Processing. His research interests include signal processing and information theoretic aspects of 5G wireless networks, full-duplex radio, mmWave communications, random matrices theory, interference alignment, statistical and array signal processing and quantum information theory. He has published over 260 publications in these areas and holds four U.S. patents. He is currently the coordinator of the FP7 projects HARP (3.2M€) in the area of highly distributed MIMO and ADEL (3.7M€) in the area of licensed shared access. Previously, he was the coordinator of FP7 Future and Emerging Technologies project CROWN (2.3M€) in the area of cognitive radio networks and HIATUS (2.7M€) in the area of interference alignment. Dr Ratnarajah is a Fellow of Higher Education Academy (FHEA), U.K., and an associate editor of the IEEE Transactions on Signal Processing.

George K. Karagiannidis (M'96-SM'03-F'14) was born in Pithagorion, Samos Island, Greece. He received the University Diploma (5 years) and PhD degree, both in electrical and computer engineering from the University of Patras, in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher at the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004, he joined the faculty of Aristotle University of Thessaloniki, Greece where he is currently Professor in the Electrical & Computer Engineering Dept. and Director of Digital Telecommunications Systems and Networks Laboratory. In January 2014, he joined Khalifa University, UAE, where he is currently Professor in the Electrical & Computer Engineering Dept. and Coordinator of the ICT Cluster.

His research interests are in the broad area of digital communications systems with emphasis on communications theory, energy efficient MIMO and cooperative communications, satellite communications, cognitive radio, localization, smart grid and optical wireless communications.

He is the author or co-author of more than 250 technical papers published in scientific journals and presented at international conferences. He is also author of the Greek edition of a book on "Telecommunications Systems" and co-author of the book "Advanced Optical Wireless Communications Systems", Cambridge Publications, 2012. He is co-recipient of the Best Paper Award of the Wireless Communications Symposium (WCS) in the IEEE International Conference on Communications (ICC'07), Glasgow, U.K., June 2007.

Dr. Karagiannidis has been a member of Technical Program Committees for several IEEE conferences such as ICC, GLOBECOM, VTC, etc. In the past he was Editor for Fading Channels and Diversity of the IEEE Transactions on Communications, Senior Editor of IEEE Communications Letters and Editor of the EURASIP Journal of Wireless Communications & Networks. He was Lead Guest Editor of the special issue on "Optical Wireless Communications" of the IEEE Journal on Selected Areas in Communications and Guest Editor of the special issue on "Large-scale multiple antenna wireless systems".

Dr. Karagiannidis has been selected as a Thomson Reuters Highly Cited Researcher. Since January 2012 he is the Editor-in Chief of IEEE Communications Letters.