Reducing the Computational Complexity of Multicasting in Large-Scale Antenna Systems

Meysam Sadeghi, Luca Sanguinetti, Romain Couillet, Chau Yuen

Reducing the Computational Complexity of Multicasting in Large-Scale Antenna Systems

Meysam Sadeghi, *Student Member, IEEE*, Luca Sanguinetti, *Senior Member, IEEE*, Romain Couillet, *Senior Member, IEEE*, and Chau Yuen, *Senior Member, IEEE*

Abstract

In this paper, we study the physical layer multicasting to multiple co-channel groups in large-scale antenna systems. The users within each group are interested in a common message and different groups have distinct messages. In particular, we aim at designing the precoding vectors solving the so-called quality of service (QoS) and weighted max-min fairness (MMF) problems, assuming that the channel state information is available at the base station (BS). To solve both problems, the baseline approach exploits the semidefinite relaxation (SDR) technique. Considering a BS with N antennas, the SDR complexity is more than $\mathcal{O}(N^6)$, which prevents its application in large-scale antenna systems. To overcome this issue, we present two new classes of algorithms that, not only have significantly lower computational complexity than existing solutions, but also largely outperform the SDR based methods. Moreover, we present a novel duality between transformed versions of the QoS and the weighted MMF problems. The duality explicitly determines the solution to the weighted MMF problem given the solution to the QoS problem, and vice versa. Numerical results are used to validate the effectiveness of the proposed solutions and to make comparisons with existing alternatives under different operating conditions.

M. Sadeghi and C. Yuen are with Singapore University of Technology and Design (SUTD), Singapore. L. Sanguinetti is with the University of Pisa, Dipartimento di Ingegneria dell'Informazione, Italy (luca.sanguinetti@unipi.it) and also with the Large Systems and Networks Group (LANEAS), CentraleSupélec, Université Paris-Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France. R. Couillet is with the Telecommunication Department of CentraleSupélec, France. (romain.couillet@centralesupelec.fr).

L. Sanguinetti and R. Couillet have been supported by the ERC Starting Grant 305123 MORE. L. Sanguinetti is also funded by the 5GIOTTO project from University of Pisa.

I. Introduction

The advent of data-hungry services and applications has significantly increased the amount of data traffic of wireless networks [1]. A considerable amount of this traffic belongs to the services that are of interest to one or several groups of subscribers such as news headlines, financial data, regular system updates, and video broadcasting [1], [2]. The traditional unicast technology is highly inefficient for these services as it ignores the nature of such a traffic demand [2]–[4]. To address this issue, the multicasting technology has been included in different releases of the third generation partnership project (3GPP) [2].

Physical layer multicasting is an efficient multicasting technique designed for wireless networks [3], [4]. It has been widely studied in the literature either for single or multiple groups [3]–[14]. In single-group multicasting, a transmitter exploits its channel state information (CSI) to send out a common stream of data to one group of users, while in multi-group multicasting multiple independent streams of common data are sent to multiple distinct groups of users. In this context, two classes of problems have received particular attention, the so-called quality-of-service (QoS) problem and the weighted max-min-fairness (MMF)[1] problem. The former aims to minimize the total transmit power while satisfying target signal-to-interference-plus-noise ratios (SINRs) at the active user equipments (UEs). The latter seeks to maximize the minimum weighted SINR among all the UEs in the system, subject to a total transmit power constraint.

A seminal treatment of single-group multicasting for both QoS and MMF problems was first presented in [3]. Therein, it was proved that both QoS and MMF problems are NP-hard and then an approximate solution was presented employing semidefinite relaxation (SDR) technique [15]. This work is then extended to a multi-group single-cell scenario in [4]. It should be noted that, since in the multi-group case the SINR of every UE depends on the precoding vectors of all other groups, even finding a feasible solution for QoS and MMF problems might be a challenging task [4], [11]. Therefore, in [4] the SDR technique is followed by a randomization and a multi-group multicast power control phase. In [5], the MMF problem is studied under per-antenna power constraint for multi-group single-cell systems. The coordinated physical layer multicasting for single-group multi-cell scenario is investigated in [6]. Also, its application to coordinated physical layer multicasting for multi-group multi-cell scenario is studied in [14].

[1]For brevity, hereafter, we refer to the weighted MMF problem as MMF problem.

The aforementioned works (among many others) are based on the SDR technique, which is characterized by high computational complexity when the system dimensions grow large, especially for large antenna arrays. More precisely, consider a single-cell network wherein a BS with N antennas serves K UEs in G multicasting groups. Then, solving the QoS problem via SDR requires $\mathcal{O}(\sqrt{GN})$ iterations of an interior point method with each iteration requiring $\mathcal{O}(G^3N^6 + KGN^2)$ arithmetic operations [4]. The computational cost of finding an approximate solution for the MMF problem is even higher as its solution is achieved by iteratively solving different instances of the QoS problem. Therefore, the SDR-based solutions are not suitable for practical implementation when N, G, or K, grow large, as envisioned in large-scale antenna systems (or more commonly known as massive MIMO systems) wherein N can be of the order of hundreds [16]–[20].

In the context of massive MIMO multicasting, two possible approaches have been recently proposed, namely, the asymptotic approach and the successive convex approximation (SCA) approach [11], [21]–[26]. The former exploits the asymptotic orthogonality of the channel vectors, when N grows very large and K is kept fixed, to simplify the SINR expression of each UE and facilitate the design of asymptotically optimal beamforming schemes [21]–[23]. In particular, [22] investigates the MMF problem for the multi-group single-cell multicasting whereas the single-group multi-cell case is studied in [21]. The extension to a multi-group multi-cell network is considered in [23]. The main problem with the asymptotic approach is that an extremely large number of antennas is required to reach the asymptotic orthogonality condition. As a consequence, the performance of the asymptotically optimal precoders is poor when the system does not have an extremely large number of antennas, i.e., $N < 1000$ [23].

The SCA approach aims at iteratively solving the non-convex QoS and MMF problems by means of SCA of the original problems around a feasible point [27], [28]. More specifically, the algorithm starts from an initial feasible point, the non-convex constraints are approximated by convex functions around this point, and the resulting convex problem is solved to obtain the next iteration. This procedure is repeated until convergence to a stationary point. In [24], the SCA technique has been applied to reduce the computational complexity of beamforming design in single-group multicasting for large-scale antenna arrays. However, the SCA method is not suitable for multi-group multicasting communications as it requires an initial feasible point, which is hard to compute in these scenarios [11]. To handle this issue, a feasible point pursuit SCA (FPP-SCA) algorithm is proposed in [25] and applied to multi-group multicasting

in [11]. Therein, in order to guarantee the feasibility of the problem, slack variables are added to relax the constraints, and a penalty is used to ensure that slacks are sparingly used. The solution of the resulting optimization problem is then used for another round of approximation and the procedure is repeated until convergence. However, the method itself has two drawbacks. First, although the solution of the approximated problem is always feasible, it might not be a feasible solution of the original multicasting problem and it is sensitive to the initial point of the algorithm as it is detailed in [25]. Second, it is still computationally demanding when the number of antennas grows.

In this paper, we address all the aforementioned drawbacks for a multi-group single-cell large-scale antenna system by introducing a two-layer precoding scheme, which is tailored for large-scale antenna systems. Our main contributions are summarized as follows:

1) We present two algorithms for the QoS and MMF problems, that outperform most of the aforementioned solutions while guaranteeing a low computational complexity.
2) We reveal new duality results that allow to solve both QoS and MMF problems simultaneously. This is in sharp contrast with the existing algorithms for which the MMF problem is solved by iteratively solving different instances of the QoS problem.
3) We introduce a heuristic algorithm that significantly improves the computational complexity while only slightly reducing the performance of both QoS and MMF solutions.

The rest of the paper is organized as follows. Section II introduces the system model for a multi-group single-cell large-scale antenna array system and formulates the corresponding QoS and MMF problems. Section III introduces the proposed two-layer precoder, it provides a duality result between transformed versions of the QoS and MMF problems, and then it proposes two algorithms for solving both. Section IV introduces a heuristic solution to further reduce the computational complexity. Section V presents the numerical results whereas conclusions are drawn in Section VI.

Notations: The following notation is used throughout the paper. Scalars are denoted by lower case letters whereas boldface lower (upper) case letters are used for vectors (matrices). We denote by $\mathbf{0}$ a matrix of appropriate size where all its elements are zero. The transpose, conjugate transpose, real part, absolute value, and second norm operator are denoted by $(\cdot)^T$, $(\cdot)^H$, $\mathrm{Re}(\cdot)$, $|\cdot|$, and $\|\cdot\|$. The set of all positive real numbers is denoted by $\mathbb{R}^+$. A circular symmetric complex Gaussian random vector $\mathbf{x}$ is denoted by $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$, where $\mathbf{0}$ and $\mathbf{C}$ are its mean and covariance matrix, respectively. The inverse of an invertible function $f(.)$ is shown by $f^{-1}(.)$.

II. System Model and Problem Formulation

Consider a single-cell large-scale antenna array system in which a BS equipped with N antennas serves G multicasting groups. Denote by $\mathcal{G} = \{1, \ldots, G\}$ the set of indices of all groups and call $\mathcal{K}_j$ the set of UE indices associated with group j, with cardinality $K_j = |\mathcal{K}_j|$ and such that $\mathcal{K}_j \cap \mathcal{K}_i = \emptyset$, $j \neq i$, i.e., each UE is associated with a single group. Within this setting, we assume that $N > K - \min_{j\in\mathcal{G}} K_j$, where $K = \sum_{j=1}^{G} K_j$ is the total number of UEs in the network. Since large antenna systems are considered in this work, this technical assumption is naturally in place. A double index notation is used to refer to each UE as e.g., "user k in group j". Under this convention, let $\mathbf{g}_{jk} \in \mathbb{C}^N$ be the channel between UE k in group j and the BS and assume that $\mathbf{g}_{jk} = \sqrt{\beta_{jk}}\mathbf{h}_{jk}$, where $\mathbf{h}_{jk} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I}_N)$ is the small-scale fading channel and β_{jk} accounts for the large-scale channel attenuation (or path loss). We assume that the BS has perfect knowledge of the channel vectors $\{\mathbf{g}_{jk}\}$.

Denoting by $\mathbf{w}_j \in \mathbb{C}^N$ the precoding vector associated with group j, the signal y_{jk} received at UE k can be written as:

$$y_{jk} = \mathbf{g}_{jk}^H \mathbf{w}_j s_j + \sum_{i=1, i\neq j}^{G} \mathbf{g}_{jk}^H \mathbf{w}_i s_i + n_{jk} \tag{1}$$

where $s_i \sim \mathcal{CN}(0,1)$ is the signal intended to group i, assumed independent across i, and $n_{jk} \sim \mathcal{CN}(0, \sigma_{jk}^2)$ accounts for the additive Gaussian noise. The SINR of UE k in group j, γ_{jk}, can be written as

$$\gamma_{jk} = \frac{\left|\mathbf{g}_{jk}^H \mathbf{w}_j\right|^2}{\sum\limits_{i=1, i\neq j}^{G} \left|\mathbf{g}_{jk}^H \mathbf{w}_i\right|^2 + \sigma_{jk}^2} \tag{2}$$

and the total average transmit power is $\sum_{j=1}^{G} \|\mathbf{w}_j\|^2$. Under the above assumptions, an instance of the QoS problem can be formulated as follows [4]:

$$\mathcal{Q}(\boldsymbol{\eta}): \quad \min_{\{\mathbf{w}_j\}} \quad \sum_{j=1}^{G} \|\mathbf{w}_j\|^2 \tag{3}$$

$$\text{s.t.} \quad \gamma_{jk} \geq \eta_{jk} \quad \forall j, k \tag{4}$$

where η_{jk} is the prescribed SINR of UE k in group j and $\boldsymbol{\eta} \in \mathbb{C}^K$ is the vector collecting all the $\{\eta_{jk}\}$. Accordingly, an instance of the MMF problem can be written as follows [4]:

$$\mathcal{F}(\boldsymbol{\eta}, P): \max_{\{\mathbf{w}_j\}} \min_j \min_k \frac{1}{\eta_{jk}} \gamma_{jk} \tag{5}$$

$$\text{s.t.} \quad \sum_{j=1}^{G} \|\mathbf{w}_j\|^2 \leq P \tag{6}$$

where P accounts for the power constraint at the BS, and $\frac{1}{\eta_{jk}}$ represents the weight of γ_{jk}. As mentioned before, $\mathcal{Q}(\boldsymbol{\eta})$ and $\mathcal{F}(\boldsymbol{\eta}, P)$ are NP-hard and the existing algorithms for computing their approximate solutions have either high computational complexity [3], [5], [11], or poor performance [21], [23], [26]. A two-layered architecture is proposed next to overcome these drawbacks.

III. The proposed Two-layer Precoding Scheme

In this section, we propose a simple and computationally efficient method to compute approximate solution to the QoS and MMF problems. The method is based on a two-layer precoding scheme: (i) the outer layer restricts the space of valid precoders to those cancelling the inter-group interference, thereby approximating the QoS and MMF problems by simpler (still non-convex) problems, denoted by QoS_{dec} and MMF_{dec}, for which trivial feasible points can be found; (ii) starting from these feasible points, the inner layer is designed to reach a suboptimal solution to the QoS_{dec} and MMF_{dec} problems, which are also feasible solutions of the original QoS and MMF problems. Section III-A presents the outer layer. Section III-B reveals an explicit duality between the QoS_{dec} and MMF_{dec} problems. Section III-C presents the inner layer and the algorithms developed. Section III-D evaluates the complexity of the proposed algorithms.

A. Outer Layer – Removing Multi-group Interference

Denote by $\mathbf{G}_i \in \mathbb{C}^{N \times K_i}$ the matrix collecting the channel vectors of all the K_i UEs in group i. The complete elimination of the multi-group interference $\sum_{i=1, i\neq j}^{G} \mathbf{g}_{jk}^H \mathbf{w}_i s_i$ in (1) is achieved by using the block-diagonalization zero-forcing (BDZF) technique [29], [30]. Consider a two-layer precoding vector for group j as follows

$$\mathbf{w}_j = \mathbf{F}_j \mathbf{c}_j \qquad \forall j \in \mathcal{G} \tag{7}$$

where $\mathbf{c}_j \in \mathbb{C}^{N-\tau_j}$ with $\tau_j = K - K_j$ is the inner layer, the design of which is discussed later, and $\mathbf{F}_j \in \mathbb{C}^{N \times (N-\tau_j)}$ is the outer layer. We design $\mathbf{F}_j$ as an isometric matrix whose

columns form a basis for the null space of $\mathbf{G}_{-j} = [\mathbf{G}_1, \ldots, \mathbf{G}_{j-1}, \mathbf{G}_{j+1}, \ldots, \mathbf{G}_G] \in \mathbb{C}^{N \times \tau_j}$, i.e., $\mathbf{G}_{-j}^H \mathbf{F}_j = \mathbf{0}_{\tau_j \times (N-\tau_j)}$. As proposed in [29], [30], $\mathbf{F}_j$ can be obtained through the singular value decomposition (SVD) of $\mathbf{G}_{-j}$. This requires $16(G-1)KN^2 + 24N\sum_{j=1}^{G}(K-K_j)^2$ floating point operations (flops)[2]. The same goal can be obtained with lower complexity (linear in the number of BS antennas N) using the QR-based decomposition approach as shown in [32]. This produces

$$\mathbf{G}_{-j} = \mathbf{Q}_j \mathbf{R}_j = \begin{bmatrix} \mathbf{Q}'_j & \mathbf{Q}''_j \end{bmatrix} \begin{bmatrix} \mathbf{R}'_j \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}'_j \mathbf{R}'_j \tag{8}$$

where $\mathbf{Q}''_j \in \mathbb{C}^{N \times (N-\tau_j)}$ gives the null space of $\mathbf{G}_{-j}$ such that $\mathbf{G}_{-j}^H \mathbf{Q}''_j = \mathbf{0}_{\tau_j \times (N-\tau_j)}$. Therefore we can use $\mathbf{Q}''_j$, as the outer layer of $\mathbf{w}_j$, i.e., $\mathbf{F}_j = \mathbf{Q}''_j$. Since the QR decomposition of an m by n matrix can be computed with $8mn^2 - \frac{8}{3}n^3$ flops [31], the total number of flops required to perform the BDZF technique reduces to $8N\sum_{j=1}^{G}(K-K_j)^2 - \frac{8}{3}\sum_{j=1}^{G}(K-K_j)^3$, which increases linearly with N. Plugging $\mathbf{F}_j = \mathbf{Q}''_j$ into (2) yields

$$\gamma_{jk} = |\overline{\mathbf{g}}_{jk}^H \mathbf{c}_j|^2 \tag{9}$$

where $\overline{\mathbf{g}}_{jk} = \frac{1}{\sigma_{jk}}(\mathbf{Q}''_j)^H \mathbf{g}_{jk} \in \mathbb{C}^{N-\tau_j}$ denotes the equivalent channel vector of UE k in group j. As $\forall j \ (\mathbf{F}_j)^H \mathbf{F}_j = \mathbf{I}_{N-\tau_j}$, the proposed outer layer does not change the transmit power, i.e., $\sum_{j=1}^{G} \|\mathbf{w}_j\|^2 = \sum_{j=1}^{G} \|\mathbf{c}_j\|^2$. Therefore, using the BDZF technique the QoS problem reduces to QoS_{dec}. Note that we define the QoS_{dec} as the transformed version of the QoS problem into G single-group multicasting QoS problems, $\{\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j)\}_{j=1}^{G}$, where the jth problem is given by

$$\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j): \quad \min_{\{\mathbf{c}_j\}} \quad \|\mathbf{c}_j\|^2 \tag{10}$$

$$\text{s.t.} \quad \left|\overline{\mathbf{g}}_{jk}^H \mathbf{c}_j\right|^2 \geq \eta_{jk} \quad \forall k \tag{11}$$

where $\boldsymbol{\eta}_j \in \mathbb{C}^{K_j}$ is the vector collecting all the quantities $\{\eta_{jk}\}$ in group j. To grasp the relation between the QoS_{dec} problem and the prescribed SINRs, i.e., $\{\boldsymbol{\eta}_j\}_{j=1}^{G}$, we denote an instance of the QoS_{dec} problem by $\overline{\mathcal{Q}}(\boldsymbol{\eta}) = \{\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j)\}_{j=1}^{G}$. Accordingly, using the BDZF technique, the

[2]The SVD calculation of an m by n matrix requires $16m^2n + 24mn^2$ flops [31].

MMF problem reduces to MMF_{dec}, where we show an instance of it by $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ and it is given as follows

$$\overline{\mathcal{F}}(\boldsymbol{\eta}, P): \max_{\{\mathbf{c}_j\}} \min_j \min_k \frac{1}{\eta_{jk}} \left|\overline{\mathbf{g}}_{jk}^H \mathbf{c}_j\right|^2 \tag{12}$$

$$\text{s.t.} \quad \sum_{j=1}^{G} \|\mathbf{c}_j\|^2 \leq P. \tag{13}$$

As mentioned in the introduction, finding a feasible point for $\mathcal{Q}(\boldsymbol{\eta})$ is hard [4]. The same holds for $\mathcal{F}(\boldsymbol{\eta}, P)$ since the common approach to solve $\mathcal{F}(\boldsymbol{\eta}, P)$ relies on iteratively solving $\mathcal{Q}(\boldsymbol{\eta})$. On the contrary, finding a feasible point for $\overline{\mathcal{Q}}(\boldsymbol{\eta}) = \{\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j)\}$ and, thus, for $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ is a straightforward task, thanks to the single-group nature of each subproblem $\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j)$. More specifically, for any given group j an initial feasible point can be computed by first choosing an arbitrary beamforming vector $\mathbf{c}_j$ and then rescaling it so as to meet the most violated SINR constraint with equality. Despite being simpler to solve than $\mathcal{Q}(\boldsymbol{\eta})$ and $\mathcal{F}(\boldsymbol{\eta}, P)$, both $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ are still NP-hard – as it easily follows observing that for $G = 1$ they reduce to the single-group problems studied in [3]. The main difficulty in solving $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ lies in the non-convexity of the SINR constraints. In Section III-C, the SCA technique is used to develop a possible solution capable of overcoming this issue.

Before delving into this, next we further detail the characteristics of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ and establish a duality and direct relations between the two problems: the solution of $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ can be obtained from that of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ (and vice versa). These results will be used in Section III-C and Section IV to compute an approximate solution to $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ by means of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ without the need of iteratively solving instances of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ as for existing alternatives [4]–[6].

B. On the duality between $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$

Let $\{\mathbf{c}_j^\star(\boldsymbol{\eta})\}$ and $P^\star(\boldsymbol{\eta})$ denote the set of optimal precoding vectors and the optimal objective value of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, respectively. Similarly, let $\{\mathbf{c}_j^\circ(\boldsymbol{\eta}, P)\}$ and $t^\circ(\boldsymbol{\eta}, P)$ denote the set of optimal precoding vectors and the optimal objective value of $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$. We then start providing the following result:

Lemma 1. *For $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ we have:*

$$\mathbf{c}_j^\star(\alpha\boldsymbol{\eta}) = \mathbf{c}_j^\circ\left(\boldsymbol{\eta}, P^\star(\alpha\boldsymbol{\eta})\right) \quad \forall j \in \mathcal{G} \tag{14}$$

with $\alpha = t^{\circ}\left(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta})\right)$. *Also, we have that:*

$$\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P) = \mathbf{c}_j^{\star}\left(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta}\right) \quad \forall j \in \mathcal{G} \tag{15}$$

with $P = P^{\star}\left(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta}\right)$.

Proof. The proof proceeds by contradiction. First, notice that, by definition, $\{\mathbf{c}_j^{\star}(\alpha\boldsymbol{\eta})\}$ is a feasible solution of $\overline{\mathcal{F}}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta}))$ with an objective value equal to α. Now, let us assume there exists a set of precoding vectors $\{\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta}))\}$ for which $t^{\circ}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta})) > \alpha$. Clearly, $\{\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta}))\}$ is also a feasible solution of $\overline{\mathcal{Q}}(\alpha\boldsymbol{\eta})$ for which all the SINR constraints are over satisfied. Hence, there exists a constant $\nu < 1$ such that $\{\nu\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta}))\}$ meets all the SINR constraints of $\overline{\mathcal{Q}}(\alpha\boldsymbol{\eta})$ with equality while providing a smaller objective value than $P^{\star}(\alpha\boldsymbol{\eta})$. This, however, contradicts our assumption and proves that (14) is valid with $\alpha = t^{\circ}(\boldsymbol{\eta}, P^{\star}(\alpha\boldsymbol{\eta}))$. A similar line of reasoning can be used to prove (15). By definition the set of precoding vectors $\{\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P)\}$ is a feasible solution of $\overline{\mathcal{Q}}(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta})$ with an objective value equal to P. Let us assume there exists $\{\mathbf{c}_j^{\star}(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta})\}$ with $P^{\star}(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta}) < P$. Then, one can use the remaining power, $P - P^{\star}(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta})$ to rescale $\{\mathbf{c}_j^{\star}(t^{\circ}(\boldsymbol{\eta}, P)\boldsymbol{\eta})\}$ and improve $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$. This is in contradiction with our assumption and completes the proof. □

Also, the following lemma can be simply proved from the definition of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$:

Lemma 2. *For* $\overline{\mathcal{Q}}(\alpha\boldsymbol{\eta})$ *and* $\forall\alpha \in \mathbb{R}^{+}$ *we have*

$$P^{\star}(\alpha\boldsymbol{\eta}) = \alpha P^{\star}(\boldsymbol{\eta}) \tag{16}$$

and $\forall j \in \mathcal{G}$

$$\mathbf{c}_j^{\star}(\alpha\boldsymbol{\eta}) = \sqrt{\alpha}\mathbf{c}_j^{\star}(\boldsymbol{\eta}). \tag{17}$$

We are now ready to state the following explicit duality between $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$:

Theorem 1. *Given the set of optimal precoding vectors and the optimal objective value of* $\overline{\mathcal{Q}}(\boldsymbol{\eta})$*, i.e.,* $\{\mathbf{c}_j^{\star}(\boldsymbol{\eta})\}$ *and* $P^{\star}(\boldsymbol{\eta})$*, the set of optimal precoding vectors and the optimal objective value of* $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$*, i.e.* $\{\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P)\}$ *and* $t^{\circ}(\boldsymbol{\eta}, P)$*, are determined as*

$$\mathbf{c}_j^{\circ}(\boldsymbol{\eta}, P) = \sqrt{\frac{P}{P^{\star}(\boldsymbol{\eta})}}\mathbf{c}_j^{\star}(\boldsymbol{\eta}) \;\; \forall j \in \mathcal{G} \tag{18}$$

$$t^{\circ}(\boldsymbol{\eta}, P) = \frac{P}{P^{\star}(\boldsymbol{\eta})} \tag{19}$$

and vice versa as

$$\mathbf{c}_j^\star(\boldsymbol{\eta}) = \frac{1}{\sqrt{t^\circ(\boldsymbol{\eta}, P)}} \mathbf{c}_j^\circ(\boldsymbol{\eta}, P) \;\; \forall j \in \mathcal{G} \tag{20}$$

$$P^\star(\boldsymbol{\eta}) = \frac{P}{t^\circ(\boldsymbol{\eta}, P)}. \tag{21}$$

Proof. Starting with (18) we have that

$$\sqrt{\frac{P}{P^\star(\boldsymbol{\eta})}} \mathbf{c}_j^\star(\boldsymbol{\eta}) \overset{(a)}{=} \mathbf{c}_j^\star \left(\frac{P}{P^\star(\boldsymbol{\eta})} \boldsymbol{\eta} \right) \overset{(b)}{=} \mathbf{c}_j^\circ \left(\boldsymbol{\eta}, P^\star \left(\frac{P}{P^\star(\boldsymbol{\eta})} \boldsymbol{\eta} \right) \right) \tag{22}$$

$$\overset{(c)}{=} \mathbf{c}_j^\circ \left(\boldsymbol{\eta}, \frac{P}{P^\star(\boldsymbol{\eta})} P^\star(\boldsymbol{\eta}) \right) = \mathbf{c}_j^\circ(\boldsymbol{\eta}, P) \tag{23}$$

where (a) follows from (17), (b) holds because of (14) and (c) is obtained using (16). The equality (19) follows from

$$P \overset{(a)}{=} P^\star \left(t^\circ(\boldsymbol{\eta}, P) \boldsymbol{\eta} \right) \overset{(b)}{=} t^\circ(\boldsymbol{\eta}, P) P^\star(\boldsymbol{\eta}) \tag{24}$$

where (a) exploits $P = P^\star \left(t^\circ(\boldsymbol{\eta}, P) \boldsymbol{\eta} \right)$ (see Lemma 1), and (b) is due to (16). The equality in (20) follows from replacing (19) in (18). □

Theorem 1 reveals the relation between the optimal precoding vectors and the optimal objective values of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$. However, as they are NP-hard, any arbitrary algorithm with polynomial complexity can provide an approximate set of precoding vectors, rather than the optimal one. Therefore, it is interesting to establish a relation between the precoding vectors and the objective values of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ while they are achieved from any arbitrary sub-optimal algorithm. This relation is given in Propositions 1 and 2.

Proposition 1. *Assume* $\{\mathbf{c}_{j,\mathrm{app}}^\star(\boldsymbol{\eta})\}$ *is a set of precoding vectors of* $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ *and* $P_{\mathrm{app}}^\star(\boldsymbol{\eta})$ *is its associated objective value achieved by any arbitrary algorithm. Then, the set of precoding vectors* $\{\sqrt{\frac{P}{P_{\mathrm{app}}^\star(\boldsymbol{\eta})}} \mathbf{c}_{j,\mathrm{app}}^\star(\boldsymbol{\eta})\}$ *(or* $\{\sqrt{\frac{P}{P_{\mathrm{app}}^\star(\boldsymbol{\eta})}} \mathbf{F}_j \mathbf{c}_{j,\mathrm{app}}^\star(\boldsymbol{\eta})\}$*) is a feasible answer for* $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ *(or* $\mathcal{F}(\boldsymbol{\eta}, P)$*), and provides an objective value* $t_{app}^\circ(\boldsymbol{\eta}, P)$ *such that* $t_{app}^\circ(\boldsymbol{\eta}, P) \in [\frac{P}{P_{\mathrm{app}}^\star(\boldsymbol{\eta})}, \frac{P}{P^\star(\boldsymbol{\eta})}]$.

Proof. Please refer to the Appendix A. □

Proposition 2. *Assume* $\{\mathbf{c}_{j,\mathrm{app}}^\circ(\boldsymbol{\eta}, P)\}$ *is a set of precoding vectors of* $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ *and* $t_{app}^\circ(\boldsymbol{\eta}, P)$ *is its associated objective value achieved by any arbitrary algorithm. Then, the set of precoding vectors* $\{\frac{1}{\sqrt{t_{app}^\circ(\boldsymbol{\eta}, P)}} \mathbf{c}_{j,\mathrm{app}}^\circ(\boldsymbol{\eta}, P)\}$ *(or* $\{\frac{1}{\sqrt{t_{app}^\circ(\boldsymbol{\eta}, P)}} \mathbf{F}_j \mathbf{c}_{j,\mathrm{app}}^\circ(\boldsymbol{\eta}, P)\}$*), is a feasible answer for* $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ *(or* $\mathcal{Q}(\boldsymbol{\eta})$*), and provides an objective value* $P_{\mathrm{app}}^\star(\boldsymbol{\eta})$ *such that* $P_{\mathrm{app}}^\star(\boldsymbol{\eta}) \in [\frac{P}{t^\circ(\boldsymbol{\eta}, P)}, \frac{P}{t_{app}^\circ(\boldsymbol{\eta}, P)}]$.

Proof. Please refer to the Appendix B. □

Note that the relation between the QoS and MMF problems was first discovered in [4], but it was not given in an explicit form. Therefore, the existing works in the literature, as [4]–[6], [11], solve the MMF problem by iteratively solving specific instances of the QoS problem. By virtue of the large number of antennas available in large-scale antenna systems and the BDZF technique, Theorem 1, Proposition 1, and Proposition 2, state that $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ and $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ (also $\mathcal{F}(\boldsymbol{\eta}, P)$ and $\mathcal{Q}(\boldsymbol{\eta})$) can be solved simultaneously. It is also interesting to observe that the upper bound of the objective value of $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ achieved via Proposition 1 is equal to (19). Also, the lower bound of the objective value of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ achieved via Proposition 2 is equal to (21).

C. Inner Layer – Successive Convex Approximation

In the sequel, the SCA technique is applied to solve $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$. We begin with $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, and rewrite $|\bar{\mathbf{g}}_{jk}^H \mathbf{c}_j|^2$ as

$$|\bar{\mathbf{g}}_{jk}^H \mathbf{c}_j|^2 = \mathbf{c}_j^H \mathbf{X}_{jk} \mathbf{c}_j \tag{25}$$

where $\mathbf{X}_{jk} = \bar{\mathbf{g}}_{jk} \bar{\mathbf{g}}_{jk}^H$ is a rank-one positive semi-definite matrix. Thus, for any arbitrary vector $\mathbf{z}_j \in \mathbb{C}^{N-\tau_j}$ we have that $(\mathbf{c}_j - \mathbf{z}_j)^H \mathbf{X}_{jk} (\mathbf{c}_j - \mathbf{z}_j) \geq 0$ from which it follows

$$\mathbf{c}_j^H \mathbf{X}_{jk} \mathbf{c}_j \geq 2\mathfrak{Re}(\mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{c}_j) - \mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{z}_j. \tag{26}$$

Now, for any $\mathbf{z}_j$ the non-convex SINR constraint $\mathbf{c}_j^H \mathbf{X}_{jk} \mathbf{c}_j \geq \eta_{jk}$ can be replaced with a tighter convex constraint given by

$$2\mathfrak{Re}(\mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{c}_j) - \mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{z}_j \geq \eta_{jk}. \tag{27}$$

By replacing (11) with (27), we obtain

$$\widetilde{\mathcal{Q}}_j(\boldsymbol{\eta}_j, \mathbf{z}_j): \min_{\{\mathbf{c}_j\}} \quad \|\mathbf{c}_j\|^2 \tag{28}$$

$$\text{s.t.} \quad 2\mathfrak{Re}(\mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{c}_j) - \mathbf{z}_j^H \mathbf{X}_{jk} \mathbf{z}_j \geq \eta_{jk} \ \ \forall k \tag{29}$$

which represents a convex approximation of $\overline{\mathcal{Q}}_j(\boldsymbol{\eta}_j)$ for a specific instance of $\mathbf{z}_j$. Now, we can introduce Algorithm 1 and its following proposition for the QoS problem.

Algorithm 1 The QoS BDZF-SCA Algorithm

1: Compute $\mathbf{F}_j \ \forall j \in \mathcal{G}$.
2: **for** $j = 1$ to G **do**
3: Select an arbitrary $\mathbf{z}_j^{(1)}$ and rescale it such that $\forall k \ \mathbf{z}_j^{(1)H}\mathbf{X}_{jk}\mathbf{z}_j^{(1)} \geq \eta_{jk}$.
4: **repeat**
5: Solve:

$$\widetilde{\mathcal{Q}}_j(\boldsymbol{\eta}_j, \mathbf{z}_j^{(i)}) : \min_{\mathbf{c}_j^{(i)}} \quad \|\mathbf{c}_j^{(i)}\|^2 \tag{30}$$

$$\text{s.t.} \ \ 2\mathfrak{Re}(\mathbf{z}_j^{(i)H}\mathbf{X}_{jk}\mathbf{c}_j^{(i)}) - \mathbf{z}_j^{(i)H}\mathbf{X}_{jk}\mathbf{z}_j^{(i)} \geq \eta_{jk} \quad \forall k. \tag{31}$$

6: Let $\mathbf{c}_j^{(i)}$ denote the optimal value obtained from $\widetilde{\mathcal{Q}}_j(\boldsymbol{\eta}_j, \mathbf{z}_j^{(i)})$, then set $\mathbf{z}_j^{(i+1)} \leftarrow \mathbf{c}_j^{(i)}$.
7: **until** Convergence
8: **end for**
9: Compute the precoding vectors $\mathbf{w}_j = \mathbf{F}_j\mathbf{c}_j \ \forall j \in \mathcal{G}$.

Proposition 3. *Algorithm 1 converges to a point satisfying the KKT conditions of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, while providing a feasible solution for $\mathcal{Q}(\boldsymbol{\eta})$.*

Proof. Please refer to the Appendix C. □

Now let us consider $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ and $\mathcal{F}(\boldsymbol{\eta}, P)$. A solution to these two problems can be achieved by first applying Algorithm 1 and then using Proposition 1. Besides, we can directly apply the SCA technique to $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ and find a solution to these two problems, similar to Algorithm 1. The latter approach is presented in Algorithm 2 and we have the following proposition for Algorithm 2.

Proposition 4. *Algorithm 2 converges to a KKT point of $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$, while providing a feasible solution to $\mathcal{F}(\boldsymbol{\eta}, P)$.*

Proof. The proof follows the same lines as the proof of Proposition 3. □

Algorithm 2 The MMF BDZF-SCA Algorithm

1: Compute $\mathbf{F}_j \ \forall j \in \mathcal{G}$.

2: Select an arbitrary set $\mathbf{z}^{(1)} := \{\mathbf{z}_j^{(1)}\}_{j=1}^G$ such that $\sum_{j=1}^G \|\mathbf{z}_j^{(1)}\|^2 \leq P$.

3: **repeat**

4: Solve:

$$\widetilde{\mathcal{F}}(\boldsymbol{\eta}, P, \mathbf{z}^{(i)}): \max_{\{\mathbf{c}_j^{(i)}\}} \quad \min_j \min_k \frac{1}{\eta_{jk}} \left[2\mathfrak{Re}(\mathbf{z}_j^{(i)H}\mathbf{X}_{jk}\mathbf{c}_j^{(i)}) - \mathbf{z}_j^{(i)H}\mathbf{X}_{jk}\mathbf{z}_j^{(i)}\right] \tag{32}$$

$$\text{s.t.} \quad \sum_{j=1}^G \|\mathbf{c}_j^{(i)}\|^2 \leq P. \tag{33}$$

5: Let $\{\mathbf{c}_j^{(i)}\}_{j=1}^G$ denote the optimal value obtained from $\widetilde{\mathcal{F}}(\boldsymbol{\eta}, P, \mathbf{z}^{(i)})$, then $\forall j \in \mathcal{G}$ set $\mathbf{z}_j^{(i+1)} \leftarrow \mathbf{c}_j^{(i)}$.

6: **until** Convergence

7: Generate the precoding vectors of $\mathcal{F}(\boldsymbol{\eta}, P)$, $\forall j \in \mathcal{G}$ $\mathbf{w}_j = \mathbf{F}_j\mathbf{c}_j$.

D. Computational Complexity

The computational load of Algorithms 1 and 2 is now assessed in terms of the number of required flops as follows. Note that both algorithms consist of three steps. The first step computes $\{\mathbf{F}_j; \forall j \in \mathcal{G}\}$ and requires $8N\sum_{j=1}^G (K-K_j)^2 - \frac{8}{3}\sum_{j=1}^G (K-K_j)^3$ flops using the QR-decomposition [31], [32]. The second step aims at designing the inner layer precoding vectors $\{\mathbf{c}_j; \forall j \in \mathcal{G}\}$ – as detailed in lines 2 to 8 (2 to 6) of Algorithm 1 (Algorithm 2). Since $\widetilde{\mathcal{Q}}(\boldsymbol{\eta}, \mathbf{z})$ and $\widetilde{\mathcal{F}}(\boldsymbol{\eta}, P, \mathbf{z})$ are both convex, they can be solved at each iteration using standard techniques with a worst case complexity of $\mathcal{O}(N^3)$ [25]. Therefore, the number of flops required by the second step is $\mathcal{O}(IN^3)$ with I being the number of iterations required to converge. As it will be observed in Section V, only a few iterations are needed to reach a satisfying solution even for large N. The third step calculates the composite precoding vectors $\mathbf{w}_j = \mathbf{F}_j\mathbf{c}_j$ and requires $8GN^2 - 8(G-1)KN$ flops. In large-scale antenna array systems, i.e., where $N \gg K$, the overall complexity of the proposed algorithm is dominated by the second step and it is of $\mathcal{O}(N^3)$. Taking into account that the complexity of SDR based techniques is greater than $\mathcal{O}(N^6)$ [4], a reduction by a factor of N^3 is achieved through Algorithms 1 and 2.

IV. A HEURISTIC INNER LAYER OF ORDER $\mathcal{O}(N)$

In the previous section, it was shown that the complexity of the proposed algorithms is of $\mathcal{O}(N^3)$, which is due to the application of SCA technique to find the inner layer precoding vectors, i.e., $\{\mathbf{c}_j\}_{j=1}^{G}$,. Therefore, the inner layer retrieval may still be computationally expensive when N is relatively large. Moreover, it requires optimization packages for solving the convex problems $\widetilde{\mathcal{Q}}(\boldsymbol{\eta}, \mathbf{z})$ and $\widetilde{\mathcal{F}}(\boldsymbol{\eta}, P, \mathbf{z})$, which may not be available on every hardware platform. Therefore, in what follows, we present a simple, yet effective, heuristic algorithm for computing the inner layer precoding vectors of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ with a complexity of $\mathcal{O}(N)$. Then, by employing Proposition 1 and the solution obtained for $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, we compute an approximate solution for $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$. Therefore, the complexity of simultaneously finding an inner layer precoder for both problems becomes $\mathcal{O}(N)$.

The proposed heuristic algorithm aims at computing the precoding vector $\mathbf{c}_j$ $\forall j \in \mathcal{G}$, in K_j sequential steps. The algorithm has two main parts, the ordering part and the successive precoder design part. Assuming that the K_j UEs in group j are labeled from 1 to K_j, the ordering part will re-label them by a bijective function $f_j : \{1, \ldots, K_j\} \to \{\mu_{j1}, \ldots, \mu_{jK_j}\}$, where $\mu_{jk} = f_j(i)$ means that the UE who was labeled as i is now re-labeled as μ_{jk} and will be served in kth step of the algorithm, $k \in \{1, \ldots, K_j\}$. Therefore, the new labels, $\{\mu_{jk}\}_{k=1}^{K_j}$, will determine the order by which the UEs in group j are served in each step. The successive precoder design part, designs the precoding vector of group j in K_j steps such that in kth step the requested SNR of UE μ_{jk} is met with minimum power consumption while the SNR of the previous $k-1$ ordered UEs, i.e., $\{\mu_{jt}\}_{t=1}^{k-1}$, is not violated. We will detail the successive precoder design and the user ordering in the following two subsections.

A. The Successive Precoder

Assume that $\forall j \in \mathcal{G}$ the UE ordering is given, i.e., $\{\mu_{jk}\}_{k=1}^{K_j}$ is known. Denote by $\mathbf{c}_j^{(k)}$ the precoding vector $\mathbf{c}_j$ at kth step, then it is computed as follows:

$$\mathbf{c}_j^{(k)} = \mathbf{c}_j^{(k-1)} + \alpha_j^{(k)} \mathbf{d}_j^{(k)} \quad k \in \{1, \ldots, K_j\} \tag{34}$$

where $\mathbf{d}_j^{(k)} \in \mathbb{C}^{N-\tau_j}$ is a unit norm vector and $\alpha_j^{(k)} \in \mathbb{C}$. In what follows, we explain how $\mathbf{d}_j^{(k)}$ and $\alpha_j^{(k)}$ should be designed such that the SNR constraint of μ_{jk} is met with minimum power consumption while the SNR of $\{\mu_{jt}\}_{t=1}^{k-1}$ is not violated.

We start by initializing the precoding vector $\mathbf{c}_j^{(1)}$ of UE μ_{j1} such that its own SNR constraint, i.e., $\left|\overline{\mathbf{g}}_{j\mu_{j1}}^H \mathbf{c}_j^{(1)}\right|^2 \geq \eta_{j\mu_{j1}}$, is met with equality. This yields $\mathbf{c}_j^{(1)} = \frac{\sqrt{\eta_{j\mu_{j1}}}}{\|\overline{\mathbf{g}}_{j\mu_{j1}}\|^2}\overline{\mathbf{g}}_{j\mu_{j1}}$. For $k \in \{2, \ldots, K_j\}$, the vectors $\mathbf{d}_j^{(k)}$ must be chosen such that the previously satisfied $k-1$ SINR constraints are not violated. This is achieved by selecting $\mathbf{d}_j^{(k)}$ orthogonal to $\{\overline{\mathbf{g}}_{j\mu_{ji}}\}_{i=1}^{k-1}$, i.e., $\overline{\mathbf{g}}_{j\mu_{ji}}^H \mathbf{d}_j^{(k)} = 0$ for $i = 1, \ldots, k-1$. To this end, $\{\mathbf{d}_j^{(k)}\}_{k=2}^{K_j}$ are computed using the Gram–Schmidt procedure, which produces $\mathbf{d}_j^{(k)} = \frac{\mathbf{u}_j^{(k)}}{\|\mathbf{u}_j^{(k)}\|}$ with

$$\mathbf{u}_j^{(k)} = \overline{\mathbf{g}}_{j\mu_{jk}} - \sum_{i=1}^{k-1} \frac{\mathbf{u}_j^{(i)H} \overline{\mathbf{g}}_{j\mu_{jk}}}{\|\mathbf{u}_j^{(i)}\|^2} \mathbf{u}_j^{(i)} \tag{35}$$

being the component of $\overline{\mathbf{g}}_{j\mu_{jk}}$ orthogonal to the subspace spanned by $\{\mathbf{u}_j^{(i)}\}_{i=1}^{k-1}$. Once the unit norm vectors $\mathbf{d}_j^{(k)}$ are computed, we proceed with the design of coefficients $\{\alpha_j^{(k)}\}_{k=2}^{K_j}$. In particular, each $\alpha_j^{(k)}$ is chosen such that the power consumption in step k, given by $||\mathbf{c}_j^{(k)}||^2 = ||\mathbf{c}_j^{(k-1)}||^2 + |\alpha_j^{(k)}|^2$, is minimized while satisfying the kth SINR constraint. More precisely, $\alpha_j^{(k)}$ must be computed as the solution of the following problem:

$$\min_{\alpha_j^{(k)}} \quad |\alpha_j^{(k)}|^2 \quad \text{s.t.} \quad \left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k)}\right|^2 \geq \eta_{j\mu_{jk}}. \tag{36}$$

As shown in the Appendix D (see also [12]), the optimal $\alpha_j^{(k)} = |\alpha_j^{(k)}| \exp(\mathrm{i}\, \angle\alpha_j^{(k)})$ is computed as:

$$\angle\alpha_j^{(k)} = -\angle\rho_j^{(k)} \tag{37}$$

$$|\alpha_j^{(k)}| = \frac{-\left|\rho_j^{(k)}\right| + \sqrt{\left|\rho_j^{(k)}\right|^2 - \left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)}\right|^2 \left(\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k-1)}\right|^2 - \eta_{j\mu_{jk}}\right)}}{\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)}\right|^2} \tag{38}$$

with $\rho_j^{(k)} = \overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)} \mathbf{c}_j^{(k-1)^H} \overline{\mathbf{g}}_{j\mu_{jk}}$. In the sequel, the above results are used to sort the UEs according to a *worst-first* policy, which is observed to achieve close-to-optimal performance by means of numerical results in Section V.

B. User Ordering

At this stage, we are only left with the computation of the UE ordering indices $\{\mu_{jk}\}$. A possible solution is illustrated in [12], for the QoS problem in single-group multicasting systems. More specifically, denote by $\mathcal{S}_j^{(k-1)} = \{\mu_{j1}, \ldots, \mu_{j(k-1)}\}$ the set of indices of the ordered UEs at

step $k-1$ and call $\mathcal{R}_j^{(k-1)}$ the set of indices of the remaining ones, i.e., $\mathcal{R}_j^{(k-1)} = \{1\ldots,K_j\} \setminus \{f_j^{-1}(\mu_{jt})\}_{t=1}^{k-1}$. Then, in [12] the set $\mathcal{S}_j^{(k)}$ is computed as $\mathcal{S}_j^{(k)} = \mathcal{S}_j^{(k-1)} \cup \{\mu_{jk}\}$ with

$$\mu_{jk} = \underset{i\in\mathcal{R}_j^{(k-1)}}{\arg\min} \quad \frac{|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}|^2}{\eta_{ji}} \tag{39}$$

corresponding to the UE index in $\mathcal{R}_j^{(k-1)}$ that has the weakest ratio (or also the most violated constraint) between the provided SNR up to step k, given by $|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}|^2$, and the requested one given by η_{ji}. The above procedure has the following two drawbacks. Firstly, it needs to calculate at each stage k the quantities $|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}|^2\ \forall i \in \mathcal{R}_j^{(k-1)}$, which requires $\mathcal{O}(K_j^2 N)$ flops for group j. This is costly if N and K_j are large. Secondly, it does not take into account the extra amount of power $|\alpha_j^{(k)}|^2$ required at stage k to meet the SNR constraint of the selected UE. To see how this comes about, consider a generic UE $i \in \mathcal{R}_j^{(k-1)}$ such that at stage k the ratio $|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}|^2/\eta_{ji}$ takes a very high value. This might happen, for example, because its own channel vector $\overline{\mathbf{g}}_{ji}$ is almost collinear to the channel vectors of the UEs selected in the previous $k-1$ stages. According to (39), such a UE will be selected at the very end of the procedure. This, however, would result in a huge power consumption because the Gram-Schmidt procedure will only have a restricted number of degrees of freedom to make $\mathbf{c}_j^{(K_j)}$ orthogonal to $\overline{\mathbf{g}}_{ji}$ for $i = 1,\ldots,K_j-1$ and at the same time to meet the requested SINR. In other words, the procedure in (39) sorts the UEs according to a *best-first* criterion such that higher priority is given to the UEs requiring low power to meet their SNR constraints.

Unlike [12], we make use of the power increase (38) at each stage k to order the UEs within each group according to a *worst-first* criterion. As mentioned before, this choice is motivated by the fact that the Gram-Schmidt procedure in (35) progressively reduces the available degrees of freedom as k tends to K_j. Therefore, since power consumption is dominated by UEs with the worst conditions (according to some criterion), higher priority should be given to these UEs. Mathematically, we propose to compute the index μ_{jk} at step k as follows:

$$\mu_{jk} = \underset{i\in\mathcal{R}_j^{(k-1)}}{\arg\max}\, |\alpha_{ji}^{(k)}|^2 \tag{40}$$

with

$$|\alpha_{ji}^{(k)}| = \frac{-\left|\rho_{ji}^{(k)}\right| + \sqrt{\left|\rho_{ji}^{(k)}\right|^2 - \left|\overline{\mathbf{g}}_{ji}^H \mathbf{d}_j^{(k)}\right|^2 \left(\left|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}\right|^2 - \eta_{ji}\right)}}{\left|\overline{\mathbf{g}}_{ji}^H \mathbf{d}_j^{(k)}\right|^2} \tag{41}$$

and $\rho_{ji}^{(k)} = \overline{\mathbf{g}}_{ji}^H \mathbf{d}_j^{(k)} \mathbf{c}_j^{(k-1)^H} \overline{\mathbf{g}}_{ji}$. As seen, μ_{jk} corresponds to the UE index in $\mathcal{R}_j^{(k-1)}$ for which the incremental power $|\alpha_{ji}^{(k)}|^2$ at stage k takes the maximum value. Note that the computational cost of this operation is still $O(K_j^2 N)$ flops as for [12]. To further reduce the computational burden, we propose an alternative approach that exploits the inherent characteristic of large-scale antenna systems. As N is large, each user $i \in \mathcal{R}_j^{(k-1)}$ can use the excess degree of freedom, provided by the large number of antennas, to chose $\mathbf{d}_j^{(k)}$ as collinear as possible to $\overline{\mathbf{g}}_{ji}$ while almost nulling the interference generated to the other UEs, i.e., $|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}| \approx 0$. Therefore, by replacing $|\overline{\mathbf{g}}_{ji}^H \mathbf{d}_j^{(k)}|^2$ with $||\overline{\mathbf{g}}_{ji}||^2$ and neglecting the term $|\overline{\mathbf{g}}_{ji}^H \mathbf{c}_j^{(k-1)}|$, the right-hand-side of (41) reduces to $\frac{\eta_{ji}}{\|\overline{\mathbf{g}}_{ji}\|^2}$. This means that UEs in group j can be ordered by simply sorting the following ratios in a descending order:

$$\left\{ \frac{\eta_{j1}}{\|\overline{\mathbf{g}}_{j1}\|^2}, \ldots, \frac{\eta_{jK_j}}{\|\overline{\mathbf{g}}_{jK_j}\|^2} \right\} \quad \forall j \in \mathcal{G}. \tag{42}$$

In other words, higher priority should be given to those UEs that have bad channel conditions compared to the target SNRs. In doing so, no greedy strategy is required for UE ordering, thereby reducing the total number of flops to $\mathcal{O}(K_j N)$. Based on the above discussion, a heuristic solution is proposed in Algorithm 3 for the inner layer. Numerical results are used in Section V to make comparisons among the above ordering policies in different settings. As it will be seen, the ordering policy of (42) largely outperforms the strategy proposed in [12].

C. The Proposed Heuristic Inner Layer Precoder

Collecting the results achieved in Sections IV-A and IV-B, we present the following heuristic algorithm to design the inner layer precoder of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$. To emphasis on the simplicity of Algorithm 3 and to enable the reproducibility of our results, its MATLAB code is provided in [33].

Algorithm 3 A heuristic algorithm of the inner layer for solving $\overline{\mathcal{Q}}(\boldsymbol{\eta})$

1: **for** $j = 1$ to G **do**
2: Sort the UEs in group j in descending order based on $\{\frac{\eta_{ji}}{\|\overline{\mathbf{g}}_{ji}\|^2}\}$ and label the list as $\{\mu_{j1}, \ldots, \mu_{jK_j}\}$, respectively.
3: Compute $\{\mathbf{d}_j^{(k)}\}_{k=1}^{K_j}$ using the Gram–Schmidt procedure in (35).
4: Set $\mathbf{c}_j^{(1)} = \frac{\sqrt{\eta_{j1}}}{\|\mathbf{g}_{j\mu_{j1}}\|^2}\mathbf{g}_{j\mu_{j1}}$.
5: **for** $k = 2$ to K_j **do**
6: **if** $|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k-1)}|^2 < \eta_{j\mu_{jk}}$ **then**
7: Compute $\alpha_j^{(k)}$ through (37) and (38).
8: Update $\mathbf{c}_j^{(k)} = \mathbf{c}_j^{(k-1)} + \alpha_j^{(k)}\mathbf{d}_j^{(k)}$.
9: **end if**
10: **end for**
11: **end for**

The complexity of Algorithm 3 can be evaluated as follows. Observe that evaluating the terms $\{\eta_{ji}/\|\overline{\mathbf{g}}_{ji}\|^2\}$ for group j requires $4(N - \tau_j)$ flops whereas sorting a list of size K_j needs $\mathcal{O}(K_j \log(K_j))$ flops. Therefore, the flop counts for UE ordering in line 2 is $4K_j(N - \tau_j) + \mathcal{O}(K_j \log(K_j))$. The Gram–Schmidt procedure of line 4 can be performed through the QR decomposition, which requires $8(N - \tau_j)K_j^2 - \frac{8}{3}K_j^3$ flops [31]. The computation of $\mathbf{c}_j^{(1)}$ needs $2(N - \tau_j + 1)$ flops. The condition $|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k-1)}|^2 < \eta_{jk}$ in line 7 avoids to waste power for those UEs whose requested SNR constraints are already met (more details on this are given in the Appendix). Lines 7 to 10 require $\mathcal{O}(N - \tau_j)$ flops, and as the condition of line 7 is true at most $K_j - 1$ times, the flops required by lines 6 to 11 is $\mathcal{O}(K_j(N - \tau_j))$. Summing all the above terms together, the complexity of Algorithm 3 is found to be $\mathcal{O}(N)$, thereby reducing the complexity of the inner layer precoder by a factor of N^2.

Note that by jointly employing Proposition 1 and Algorithm 3, the approximated precoding vectors for $\overline{\mathcal{F}}(\boldsymbol{\eta}, P)$ can be computed as

$$\mathbf{c}_{j,\mathrm{BDZF-HEU}}^{\circ}(\boldsymbol{\eta}, P) = \sqrt{\frac{P}{P_{\mathrm{BDZF-HEU}}^{\star}(\boldsymbol{\eta})}}\mathbf{c}_{j,\mathrm{BDZF-HEU}}^{\star}(\boldsymbol{\eta}) \qquad j \in \mathcal{G} \tag{43}$$

where $\{\mathbf{c}_{j,\mathrm{BDZF-HEU}}^{\star}(\boldsymbol{\eta})\}$ and $P_{\mathrm{BDZF-HEU}}^{\star}(\boldsymbol{\eta})$ denotes the precoding vectors and the total power consumption as obtained with Algorithm 3. Therefore the precoding vectors for $\mathcal{F}(\boldsymbol{\eta}, P)$ are given by $\{\mathbf{F}_j \mathbf{c}_{j,\mathrm{BDZF-HEU}}^{\circ}(\boldsymbol{\eta}, P)\}$.

V. Numerical results

Monte Carlo simulations are used to assess the performance of the proposed algorithms and to make comparisons with existing alternatives. In particular, we consider the algorithm presented in [4], which employs the SDR technique followed by a randomization and multicast multigroup power control (MMPC) policy.[3] Comparisons are also made with the asymptotic results of [23], the FPP based algorithm presented in [11], and the heuristic algorithms developed in [12]. A single cell system with radius of 900 meters is considered with UEs being distributed uniformly and randomly in the cell excluding an inner circular area of radius 100 meters. For each value of N, the average values of power consumption or minimum SINR of the system are obtained from 100 different channel realizations and UE distributions. We assume (if not otherwise specified) that there are $G = 3$ multicasting groups, each counting $K_j = 10$ UEs (such that $K = 30$). The channel vector $\mathbf{g}_{jk}$ between UE k in group j and the BS is modeled as $\mathbf{g}_{jk} = \sqrt{\beta_{jk}}\mathbf{h}_{jk}$ where $\mathbf{h}_{jk} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ represents the small scale fading and β_{jk} accounts for the large scale attenuation given by $\beta_{jk} = -128.1 - 37.6 \log_{10} d_{jk}$ dB with d_{jk} being the distance between the UE and the BS expressed in kilometers [34]. The noise power spectral density is assumed to be -174 dBm/Hz, and the channel bandwidth is 20 MHz [26]. All the simulations are performed on a 64-bit Linux operating system with Intel Xeon processor E5-1680 v3.

Fig. 1 compares the average power consumption of the ordering policies proposed in [12] with those given by (40) and (42), for $G = 1$, $K = 20$ and $\eta = 63$, 127, and 255 (which correspond to a target rate for each UE of $6, 7$ and 8 bit/s/Hz, respectively). The proposed ordering policies are seen to outperform the ordering of [12]. Moreover, the simple ordering policy of (42) has even a slightly better performance than (40). Note that, as the ordering belongs to the heuristic inner layer of the proposed precoder and as the outer layer removes the effect of inter-group interference, the same conclusion holds for $G > 1$. Based on the above results, the simpler ordering policy presented in (42) will be used in the remainder of this section.

Fig. 2 depicts the average power consumption of the QoS problem versus the number of antennas N at the BS. We assume that $\eta_{jk} = 255$ for all UEs (corresponding to 8 bit/s/Hz/UE), and it is chosen in agreement with 5G rate requirements [35], but the conclusions generically hold for all other values of $\boldsymbol{\eta}$. The performance of Algorithm 1, and the combination of BDZF and Algorithm 3 is compared to other existing algorithms. As the QoS problem is NP-hard, a

[3]For the randomization phase, 100 samples are generated using the Gaussian randomization method [4].

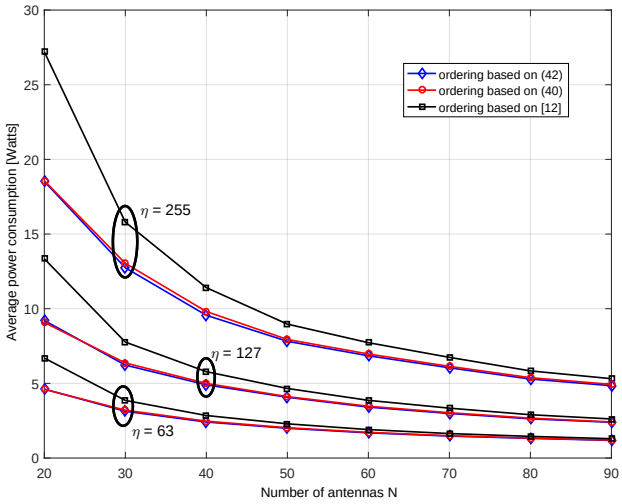


Fig. 1: Average power consumption of the QoS problem, comparing different ordering policies for $G = 1$ and $K = 20$.

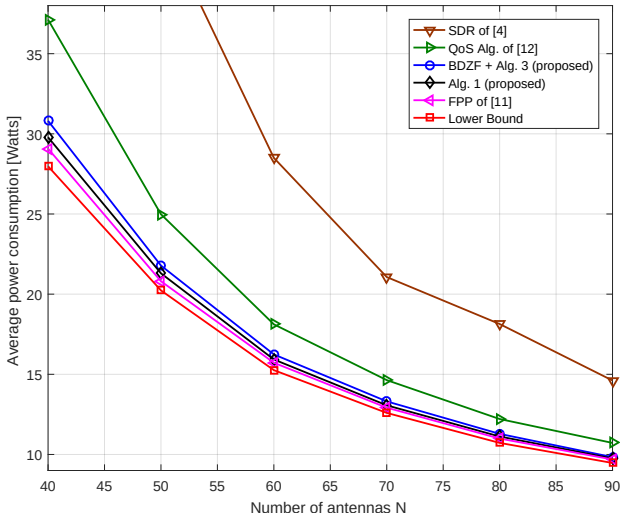


Fig. 2: Average power consumption of the QoS problem with $\eta_{jk} = 255 \ \ \forall j, k$.

lower bound of the problem is also presented as a benchmark [4]. Observe that, the proposed algorithms outperform the SDR-based solution in [4] and the heuristic one in [12], while they have nearly the same performance as [11]. However, this is achieved at a much lower complexity and computational cost as detailed next. Note that for $N \geq 60$ both algorithms are at most 6% away from the lower bound and this gap reduces as N grows large, while for SDR technique

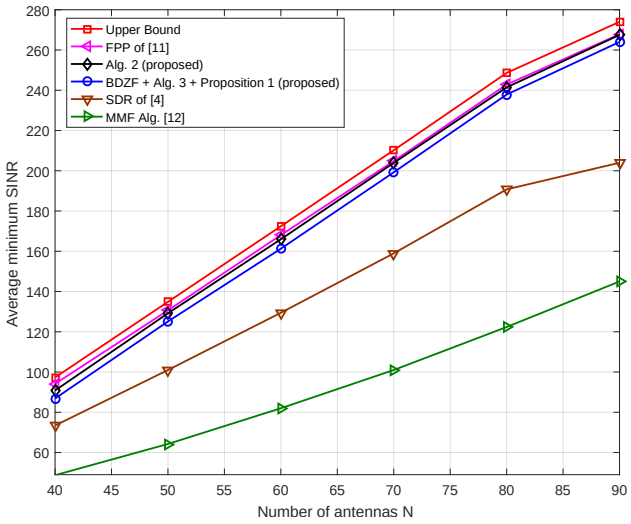


Fig. 3: Average minimum SINR of the MMF problem for $P = 10$ Watt.

this gap is 87% and reduces slowly by adding more antennas.

Fig. 3 illustrates the average minimum SINR of the MMF problem versus N. The available power at the BS is considered to be 10 Watt. In this figure, the performance achieved by Algorithm 2, and the combination of BDZF, Algorithm 3, and Proposition 1 is compared to other existing algorithms. Also, the upper bound of the problem is depicted as a benchmark. Similar to the results of Fig. 2, the proposed algorithms largely outperform [4] and [12], while nearly having the same performance as [11]. However, this is achieved for a computational cost that is significantly smaller than other algorithms as detailed next. Observe that Algorithm 2 is within 5% of the upper bound with just $N = 50$ antennas. Also, Algorithm 3 jointly with BDZF and Proposition 1 achieve the same target with $N = 70$ antennas.

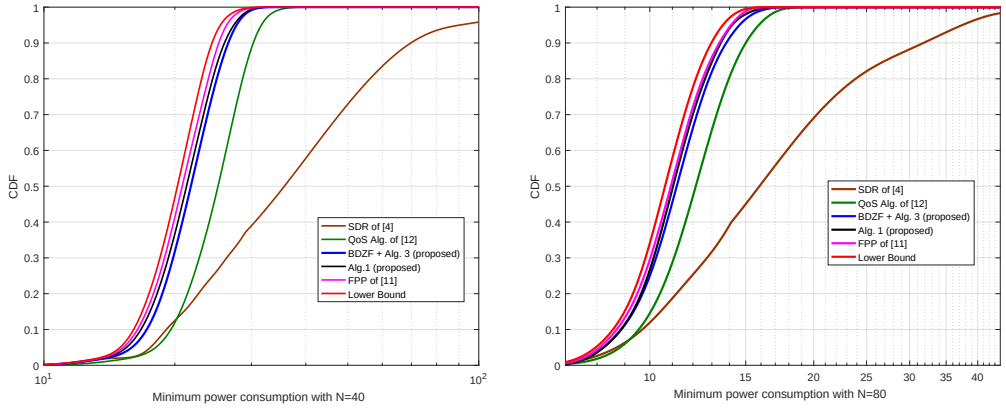
To assess the computational complexity of the investigated algorithms more intuitively, beside the complexity analysis of Section III-D and Section IV-C, we also present the computation time required to approximately solve $\mathcal{Q}(\boldsymbol{\eta})$ and $\mathcal{F}(\boldsymbol{\eta}, P)$ versus N, in Table I. The table presents the average time (in seconds) required to solve the QoS and MMF problems. The second and third columns report the average time required by the SDR and the FPP algorithms to solve an instance of the QoS problem. The fourth and fifth columns present the average required time by the same algorithms to solve an instance of the MMF problem. Note the increase in time from the QoS problem to the MMF problem, as in the SDR and FPP algorithms the MMF is

TABLE I: The average time (in seconds) required to solve QoS, MMF, or both of them.

	QoS Problem		MMF Problem		QoS and MMF (both problems simultaneously)	
	SDR - [4]	FPP - [11]	SDR - [4]	FPP - [11]	Alg. 1 + Proposition 1	BDZF + Alg.3 + Proposition 1
$N = 40$	55	41	419	356	11.3	2.5×10^{-3}
$N = 50$	67	51	579	450	11.6	2.8×10^{-3}
$N = 60$	84	61	798	507	11.7	3.1×10^{-3}
$N = 70$	110	75	1151	617	11.9	3.5×10^{-3}
$N = 80$	146	87	1549	727	12.2	4.0×10^{-3}
$N = 90$	182	107	2050	865	12.5	4.5×10^{-3}

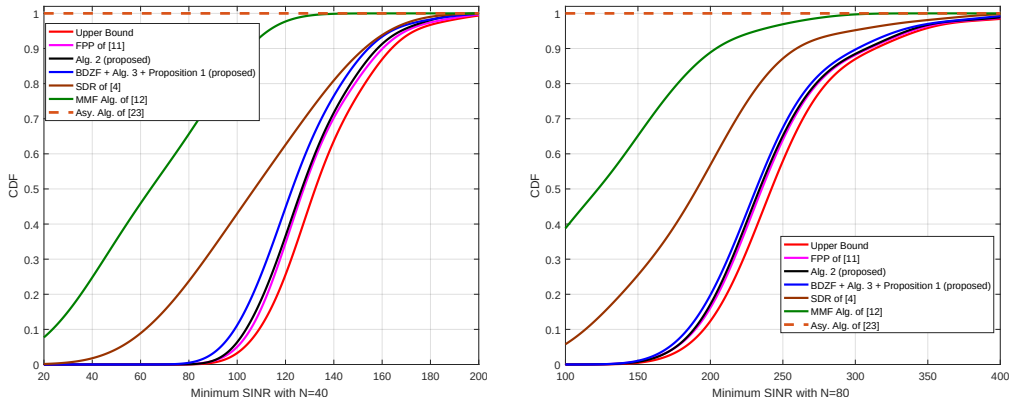
solved by iteratively applying the QoS algorithm. The sixth and seventh columns of the table present the average time required to solve both QoS and MMF problems simultaneously using the proposed algorithms. Note that not only we solve both problems at the same time with good performance, but also the required time has reduced significantly. As an example, for $N = 90$, the combination of BDZF, Algorithm 3, and Proposition 1, solves both problems in less than 5 milliseconds, while the SDR and the FPP algorithms require 2050 or 865 seconds, respectively, just to solve the MMF problem. At the same time, as shown in Figs. 2 and 3, the solution provided by joint application of the BDZF, Algorithm 3, and Proposition 1 is nearly as good as the solution achieved from the FPP Algorithm, and significantly outperforms the SDR Algorithm.

Figs. 4 and 5 present the cumulative distribution function (CDF) of the precoder power consumption and the minimum SINR of the system for the QoS and MMF problems, respectively. For the QoS problem, the requested SINR by each user is assumed to be 255, and for the MMF problem the available power at the BS is considered to be 10 Watt. Unlike Figs. 2 and 3 that provide the average of minimum power consumption or the average of minimum SINR of the system, these figures provide a clear vision on the distribution of these quantities, for the existing and proposed algorithms. It is seen that, as we increase N from 40 to 80, the CDF curves of our proposed algorithms become closer to the optimal bound, and also improve significantly in terms of performance, thanks to the large number of antennas. As an example, for the QoS problem with $N = 40$, the power consumption is greater than 20 Watt 60% of the times, while with $N = 80$ our proposed algorithms always meet the requested SINRs with less than 17 Watt. Also, for the MMF problem none of the algorithms can provide a minimum SINR bigger than 200 with $N = 40$, while for $N = 80$ our proposed algorithms can provide a minimum SINR



(a) $N = 40$. (b) $N = 80$.

Fig. 4: CDF of minimum power consumption (QoS problem) of the system.

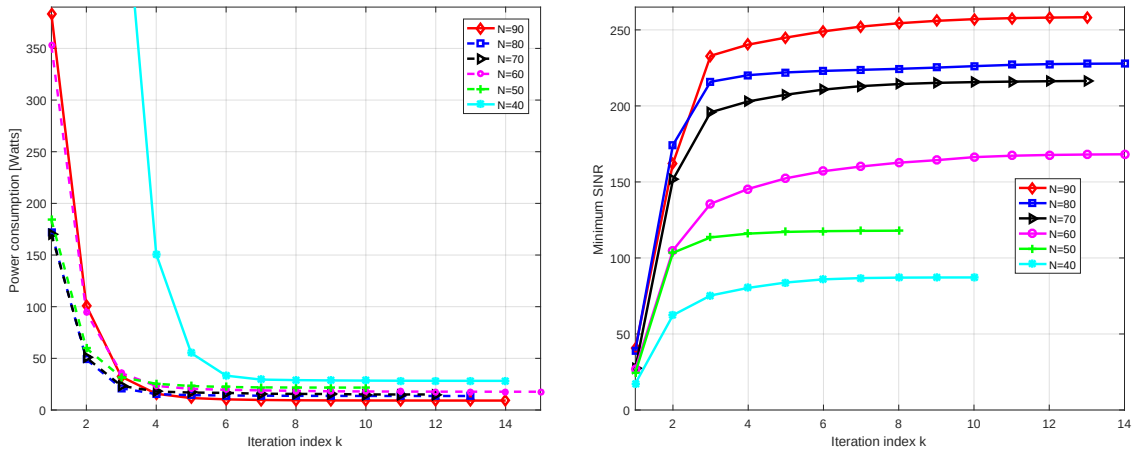


(a) $N = 40$. (b) $N = 80$.

Fig. 5: CDF of minimum SINR (MMF problem) of the system with $P = 10$.

bigger than 200 in 80% of the times. Fig. 5 also contains the CDF of the asymptotic approach of [23]. Notice that the asymptotic approach can never provide an SINR which is bigger than 20 (or 100) with $N = 40$ (or $N = 80$) antennas. The insufficiency of the asymptotic approach is detailed in [23].

In Section III, we have elaborated the BDZF-SCA approach and proved the convergence of Algorithms 1 and 2, but we have not specified the number of iterations required by each algorithm to converge. Table II presents the average number of iterations required by Algorithms 1 and 2 to achieve convergence for different values of G, K and N. Denoting the objective value achieved



(a) Algorithm 1. (b) Algorithm 2.

Fig. 6: The convergence behavior of Algorithms 1 and 2 for different number of antennas N.

at the kth iteration of either Algorithm 1 or 2 as $\varepsilon(k)$, the convergence condition of Table II is $\frac{|\varepsilon(k+1)-\varepsilon(k)|}{\varepsilon(k)} < 10^{-3}$. Also, Fig. 6 illustrates $\varepsilon(k)$ for both algorithms at each iteration index k for different number of antennas N. As it is seen, both algorithms converge in a few iterations for any value of N.

TABLE II: Average number of iterations required by Algorithms 1 and 2.

	$(G=2, K=10)$		$(G=3, K=10)$		$(G=3, K=15)$	
	Algorithm 2	Algorithm 1	Algorithm 2	Algorithm 1	Algorithm 2	Algorithm 1
$N=40$	15.12	16.56	15.38	16.90	15.95	17.00
$N=50$	15.14	16.61	15.50	16.94	16.12	17.11
$N=60$	15.21	16.74	15.54	16.97	16.23	17.14
$N=70$	15.40	16.88	15.60	17.04	16.46	17.40
$N=80$	15.44	16.94	15.60	17.08	16.52	17.51
$N=90$	15.53	17.10	15.62	17.12	16.68	17.56

So far, we have assumed that the BS has perfect knowledge of the channel vectors $\{\mathbf{g}_{jk}\}$. Although some of the analysis can in principle be extended (to some extent) to the scenario with imperfect CSI, we believe that this is out of the scope of this work and thus it is left for the future. To partially fulfill this lack, we now investigate the impact of imperfect CSI on the performance of the proposed algorithms. A time-division-protocol (TDD) is employed such that

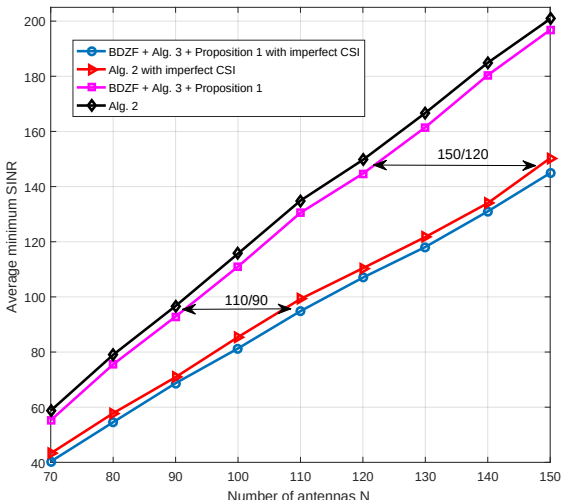


Fig. 7: Evaluating the impact of imperfect CSI at the BS for the MMF Problem with $G = 4$, $K_j = 15\ \forall j$ and $P = 10$ Watt. Channel estimation is performed using an MMSE estimator that exploit UL pilot signals of length $\tau_p = K$ and power equal to 1 Watt.

channel estimation can be performed in the uplink on the basis of UE pilot signals and then used in the downlink. We assume that pilots of length $\tau_p = K$ are used, with power equal to 1 Watt. The estimates of channel vectors $\{\mathbf{g}_{jk}\}$ are computed at the BS using an MMSE estimator. This yields [36]

$$\hat{\mathbf{g}}_{jk} = \frac{\sqrt{\tau_p}\beta_{jk}}{\sigma^2 + \tau_p \beta_{jk}} \left(\sqrt{\tau_p}\mathbf{g}_{jk} + \mathbf{n}\right) \quad \forall j, k \tag{44}$$

where $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the additive Gaussian noise. Fig. 7 reports the performance of the proposed algorithms for MMF when perfect and imperfect CSI (as given in (44)) is available at the BS. We assume that $G = 4$ and $K_j = 15$ for $j = 1, \ldots, 4$. As expected, imperfect CSI degrades the performance of the proposed algorithms. However, such a performance loss can be compensated by using more antennas at the BS. Quantitively speaking, with imperfect CSI N must be roughly increased by a factor of $15\% - 30\,\%$ compared to the perfect CSI case. The higher N, the larger the factor. For example, to achieve the same performance of the perfect CSI case with $N = 90$ and 120, then 110 and 150 antennas are respectively needed with imperfect CSI, corresponding to a 22% and 25% increase.

VI. Conclusions

Multicasting is an efficient technology to transmit distinct common data streams to multiple groups of users. The existing multicasting algorithms are computationally expensive, especially for the large-scale antenna systems. There has been a struggle to present algorithms that can reduce this complexity. However, these algorithms are still too complex to fit in the next generation of wireless systems that enjoy hundreds of antennas, or they have a very poor performance. In this paper, we designed new algorithms, which are tailored for physical layer multicasting in large-scale antenna systems. These algorithms not only provide good performance, but also this good performance is achieved at a very low computational cost and complexity. Particularly, we used the large number of antennas to cancel the inter-group interference and simplify the QoS and MMF problems. This enabled us to present efficient algorithms for these two problems. Moreover, we showed that the two problems can be solved simultaneously and at no extra cost, while the baseline method is to solve the MMF problem by iteratively solving the QoS problem.

Appendix A
Proof of Proposition 1

Starting with the power constraint of (13) and replacing $\mathbf{c}_j$ with $\sqrt{\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})$ we have

$$\sum_{j=1}^{G}\|\sqrt{\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\|^2=\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}\sum_{j=1}^{G}\|\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\|^2=P \tag{45}$$

which proves the feasibility of the proposed solution. For the achievable objective value of $\overline{\mathcal{F}}(\boldsymbol{\eta},P)$ (or $\mathcal{F}(\boldsymbol{\eta},P)$) using $\{\sqrt{\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\}$, we have

$$t^\circ_{app}(\boldsymbol{\eta},P)=\min_j\min_k\frac{1}{\eta_{jk}}\left|\overline{\mathbf{g}}^H_{jk}\sqrt{\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\right|^2=\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}\min_j\min_k\frac{1}{\eta_{jk}}\left|\overline{\mathbf{g}}^H_{jk}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\right|^2. \tag{46}$$

Denote $\lambda:=\min_j\min_k\frac{1}{\eta_{jk}}\left|\overline{\mathbf{g}}^H_{jk}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\right|^2$. As $\{\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\}$ is a set of the precoding vectors of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, $\lambda\geq 1$. Therefore

$$t^\circ_{app}(\boldsymbol{\eta},P)=\frac{P}{\frac{1}{\lambda}P^\star_{\rm app}(\boldsymbol{\eta})}\geq\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}. \tag{47}$$

As $\frac{1}{\lambda}P^\star_{\rm app}(\boldsymbol{\eta})$ is an objective value of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ that can be achieved by $\{\frac{1}{\sqrt{\lambda}}\mathbf{c}^\star_{j,{\rm app}}(\boldsymbol{\eta})\}$, it is bigger than or equal to the optimal objective value of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$, i.e., $\frac{1}{\lambda}P^\star_{\rm app}(\boldsymbol{\eta})\geq P^\star(\boldsymbol{\eta})$, and we have

$$\frac{P}{P^\star_{\rm app}(\boldsymbol{\eta})}\leq t^\circ_{app}(\boldsymbol{\eta},P)\leq\frac{P}{P^\star(\boldsymbol{\eta})}. \tag{48}$$

APPENDIX B
PROOF OF PROPOSITION 2

Starting with the SINR constraint of (11) and replacing $\mathbf{c}_j$ with $\frac{1}{\sqrt{t^{\circ}_{app}(\boldsymbol{\eta},P)}}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)$, we have

$$\frac{1}{t^{\circ}_{app}(\boldsymbol{\eta},P)}|\overline{\mathbf{g}}^H_{jk}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)|^2 = \frac{\eta_{jk}}{t^{\circ}_{app}(\boldsymbol{\eta},P)}\frac{|\overline{\mathbf{g}}^H_{jk}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)|^2}{\eta_{jk}} \overset{(a)}{\geq} \eta_{jk} \tag{49}$$

where in (a) is due to the fact that $t^{\circ}_{app}(\boldsymbol{\eta},P)$ is the minimum weighted SINR among all UEs. Therefore $\{\frac{1}{\sqrt{t^{\circ}_{app}(\boldsymbol{\eta},P)}}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\}$ is a feasible answer of $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and $\mathcal{Q}(\boldsymbol{\eta})$. For the objective value we have

$$P^{\star}_{\mathrm{app}}(\boldsymbol{\eta}) = \sum_{j=1}^{G}\|\frac{1}{\sqrt{t^{\circ}_{app}(\boldsymbol{\eta},P)}}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\|^2 = \frac{1}{t^{\circ}_{app}(\boldsymbol{\eta},P)}\sum_{j=1}^{G}\|\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\|^2 \leq \frac{P}{t^{\circ}_{app}(\boldsymbol{\eta},P)}. \tag{50}$$

Denote $\lambda := \frac{1}{P}\sum_{j=1}^{G}\|\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\|^2$. As $\{\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\}$ is a set of precoding vectors of $\overline{\mathcal{F}}(\boldsymbol{\eta},P)$, $\lambda \leq 1$. Therefore,

$$P^{\star}_{\mathrm{app}}(\boldsymbol{\eta}) = \frac{P}{\frac{1}{\lambda}t^{\circ}_{app}(\boldsymbol{\eta},P)} \leq \frac{P}{t^{\circ}_{app}(\boldsymbol{\eta},P)}. \tag{51}$$

Since $\frac{1}{\lambda}t^{\circ}_{app}(\boldsymbol{\eta},P)$ is an objective value of $\overline{\mathcal{F}}(\boldsymbol{\eta},P)$ achieved by $\{\frac{1}{\sqrt{\lambda}}\mathbf{c}^{\circ}_{j,\mathrm{app}}(\boldsymbol{\eta},P)\}$, it is less than or equal to the optimal objective value of $\overline{\mathcal{F}}(\boldsymbol{\eta},P)$, i.e., $\frac{1}{\lambda}t^{\circ}_{app}(\boldsymbol{\eta},P) \leq t^{\circ}(\boldsymbol{\eta},P)$. Therefore we have

$$\frac{P}{t^{\circ}(\boldsymbol{\eta},P)} \leq P^{\star}_{\mathrm{app}}(\boldsymbol{\eta}) \leq \frac{P}{t^{\circ}_{app}(\boldsymbol{\eta},P)}. \tag{52}$$

APPENDIX C
PROOF OF PROPOSITION 3

As $\mathbf{z}^{(1)H}_j\mathbf{X}_{jk}\mathbf{z}^{(1)}_j \geq \eta_{jk}\ \forall k \in \mathcal{K}_j, j \in \mathcal{G}$, $\mathbf{z}^{(1)}_j$ is a feasible solution of $\widetilde{\mathcal{Q}}^{(1)}_j(\boldsymbol{\eta}_j,\mathbf{z}^{(1)}_j)$. Now consider the $(i+1)$th iteration of the problem $\forall i \in \{0,1,\ldots\}$. $\forall k \in \mathcal{K}_j, j \in \mathcal{G}$ we have

$$2\mathfrak{Re}(\mathbf{z}^{(i+1)H}_j\mathbf{X}_{jk}\mathbf{c}^{(i+1)}_j) - \mathbf{z}^{(i+1)H}_j\mathbf{X}_{jk}\mathbf{z}^{(i+1)}_j \overset{a}{=} 2\mathfrak{Re}(\mathbf{c}^{(i)H}_j\mathbf{X}_{jk}\mathbf{c}^{(i+1)}_j) - \mathbf{c}^{(i)H}_j\mathbf{X}_{jk}\mathbf{c}^{(i)}_j \tag{53}$$

where (a) is due to our update rule, $\mathbf{z}^{(i+1)}_j \leftarrow \mathbf{c}^{(i)}_j$. Now if we set $\mathbf{c}^{(i+1)}_j = \mathbf{c}^{(i)}_j$, (53) reduces to $\mathbf{c}^{(i)H}_j\mathbf{X}_{jk}\mathbf{c}^{(i)}_j$ which is bigger than η_{jk} due to (26) and (27). Therefore $\mathbf{c}^{(i)}_j$ is a feasible solution of $\widetilde{\mathcal{Q}}^{(i+1)}_j(\boldsymbol{\eta}_j,\mathbf{z}^{(i+1)}_j)$. Hence the objective function of $(i+1)$th iteration is less than or equal to the objective function of (i)th iteration. As the objective function is bounded from below, by

successively solving the problem we achieve a non-increasing bounded sequence. Therefore the algorithm converges. Due to (26), any internal precoding vector $\mathbf{c}_j$ that satisfies (27), will also satisfy (11) and as a result, any answer to $\widetilde{\mathcal{Q}}(\boldsymbol{\eta}, \mathbf{z})$ is a feasible answer to $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ and therefore $\mathcal{Q}(\boldsymbol{\eta})$. Due to the update rule and the inner approximation in (27), the convergence point satisfies the KKT conditions for $\overline{\mathcal{Q}}(\boldsymbol{\eta})$ as detailed in [28].

APPENDIX D
SOLUTION TO (36)

Hereby we prove the solution of the following problem, i.e. (36),

$$\min_{\alpha_j^{(k)}} \quad |\alpha_j^{(k)}|^2 \quad \text{s.t.} \quad \left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k)}\right|^2 \geq \eta_{j\mu_{jk}} \tag{54}$$

is achieved by $\alpha_j^{(k)} = |\alpha_j^{(k)}| \exp(i \angle \alpha_j^{(k)})$ where:

$$\angle \alpha_j^{(k)} = -\angle \rho_j^{(k)} \tag{55}$$

$$|\alpha_j^{(k)}| = \frac{-\left|\rho_j^{(k)}\right| + \sqrt{\left|\rho_j^{(k)}\right|^2 - \left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)}\right|^2 \left(\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k-1)}\right|^2 - \eta_{j\mu_{jk}}\right)}}{\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)}\right|^2} \tag{56}$$

with $\rho_j^{(k)} = \overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)} \mathbf{c}_j^{(k-1)^H} \overline{\mathbf{g}}_{j\mu_{jk}}$. We start with the SNR constraint $\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k)}\right|^2 \geq \eta_{j\mu_{jk}}$, and replace $\mathbf{c}_j^{(k)}$ with $\mathbf{c}_j^{(k-1)} + \alpha_j^{(k)} \mathbf{d}_j^{(k)}$ using (34). Denote $|\bar{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)}|^2$, $2\,\mathrm{Re}(e^{j\angle\alpha_j^{(k)}}\ \bar{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)} \mathbf{c}_j^{(k-1)H} \bar{\mathbf{g}}_{j\mu_{jk}})$, and $|\bar{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k-1)}|^2 - \eta_{j\mu_{jk}}$ as A, B, and C, respectively. The SNR constraint can be represented as

$$\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k)}\right|^2 - \eta_{j\mu_{jk}} = A|\alpha_j^{(k)}|^2 + B|\alpha_j^{(k)}| + C \geq 0 \tag{57}$$

Notice that if $\left|\overline{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{c}_j^{(k)}\right|^2 \geq \eta_{j\mu_{jk}}$, to minimize the power, no transmission shall be arranged for user μ_{jk}, i.e., $\alpha_j^{(k)} = 0$, and the next user shall be served. Otherwise, $C < 0$. Now we transform (57) to an equality by introducing $\lambda \geq 0$ as follows

$$A|\alpha_j^{(k)}|^2 + B|\alpha_j^{(k)}| + C - \lambda = 0 \tag{58}$$

Therefore $|\alpha_j^{(k)}| = \frac{-B+\sqrt{B^2-4A(C-\lambda)}}{2A}$, as $\frac{-B-\sqrt{B^2-4A(C-\lambda)}}{2A} < 0$ and is not a valid answer for $|\alpha_j^{(k)}|$. Moreover, as $4A\lambda \geq 0$, to minimize the power λ should be equal to zero, i.e, the power should be used to meet the SNR constraint with equality. Hence $|\alpha_j^{(k)}| = \frac{-B+\sqrt{B^2-4AC}}{2A}$. Now as A

is fixed, to minimize $|\alpha_j^{(k)}|^2$ we should minimize $-B+\sqrt{B^2-4AC}$. Note $-B+\sqrt{B^2-4AC}$ always has a negative derivative with respect to B, hence its minimum is achieved for the maximum value of B. Denote $\Gamma_{jk}e^{j\angle\theta_{jk}} = \bar{\mathbf{g}}_{j\mu_{jk}}^H \mathbf{d}_j^{(k)} \mathbf{c}_j^{(k-1)H} \bar{\mathbf{g}}_{j\mu_{jk}}$, we have $B = 2\ \Gamma_{jk}\,\mathrm{Re}(e^{j(\angle\theta_{jk}+\angle\alpha_j^{(k)})})$, the maximum of which achieved if $\angle\alpha_j^{(k)} = -\angle\theta_{jk}$ and $|\alpha_j^{(k)}|$ is given as in (56).

REFERENCES

[1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2015 – 2020," San Jose, CA, Tech. Rep., Feb. 2016.

[2] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and Rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[3] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[4] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[5] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Weighted fair multicast multigroup beamforming under per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.

[6] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–21, Jan. 2013.

[7] S. X. Wu, W. K. Ma, and A. M. C. So, "Physical-layer multicasting by stochastic transmit beamforming and Alamouti space-time coding," *IEEE Trans. Signal Process.*, vol. 61, no. 17, pp. 4230–4245, Sep. 2013.

[8] I. H. Kim, D. J. Love, and S. Y. Park, "Optimal and successive approaches to signal design for multiple antenna physical layer multicasting," *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2316–2327, Aug. 2011.

[9] M. Li, S. Kundu, D. A. Pados, and S. N. Batalama, "Waveform design for secure SISO transmissions and multicasting," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1864–1874, Sep. 2013.

[10] B. Zhu, J. Ge, Y. Huang, Y. Yang, and M. Lin, "Rank-two beamformed secure multicasting for wireless information and power transfer," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 199–203, Feb. 2014.

[11] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Multicast multigroup beamforming for per-antenna power constrained large-scale arrays," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 271–275.

[12] R. Hunger, D. A. Schmidt, M. Joham, A. Schwing, and W. Utschick, "Design of single-group multicasting-beamformers," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Glasgow, Jun. 2007, pp. 2499–2505.

[13] M. C. Yue, S. X. Wu, and A. M. C. So, "A robust design for MISO physical-layer multicasting over line-of-sight channels," *IEEE Signal Process. Lett.*, vol. 23, no. 7, pp. 939–943, Jul. 2016.

[14] D. Chen and V. Kuehn, "Weighted max-min fairness oriented load-balancing and clustering for multicast cache-enabled F-RAN," in *Int. Symp. Turbo Codes and Iterative Inform. Process. (ISTC)*, Brest, France, Sept 2016, pp. 395–399.

[15] Z.-Q. Luo, W. K. Ma, A. M. C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[16] 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.

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

[18] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[19] T. L. Marzetta, "How much training is required for multiuser MIMO?" in *Proc. IEEE Asilomar Conference on Signals, Systems and Computers (ACSSC'06)*, Pacific Grove, CA, USA, Oct. 2006, pp. 359–363.

[20] A. Arvola, A. Tolli, and D. Gesbert, "Two-layer precoding for dimensionality reduction in massive MIMO," in *24th European Signal Process. Conf. (EUSIPCO)*, budapest, Hungary, Aug. 2016, pp. 2000–2004.

[21] Z. Xiang, M. Tao, and X. Wang, "Massive MIMO multicasting in noncooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1180–1193, Jun. 2014.

[22] H. Zhou and M. Tao, "Joint multicast beamforming and user grouping in massive MIMO systems," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, London, Jun. 2015, pp. 1770–1775.

[23] M. Sadeghi and C. Yuen, "Multi-cell multi-group massive MIMO multicasting: An asymptotic analysis," in *Proc. IEEE Global Commun. Conf. (Globecom)*, San Diego, CA, Dec. 2015, pp. 1–6.

[24] 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.

[25] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 804–808, Jul. 2015.

[26] Z. Xiang, M. Tao, and X. Wang, "Massive MIMO multicasting in noncooperative multicell networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, Jun. 2014, pp. 4777–4782.

[27] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, no. 4, pp. 681–683, 1978.

[28] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, 2010.

[29] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[30] L.-U. Choi and R. D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 20–24, Jan. 2004.

[31] M. Arakawa, "Computational workloads for commonly used signal processing kernels," MIT Lincoln Lab., Lexington, MA, Tech. Rep., 2006.

[32] R. Chen, R. W. Heath, and J. G. Andrews, "Transmit selection diversity for unitary precoded multiuser spatial multiplexing systems with linear receivers," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1159–1171, Mar. 2007.

[33] M. Sadegi, "meysamsadeghi/reducing-the-computational-complexity-of-multicasting-in-large-scale-antenna-systems." [Online]. Available: https://github.com/meysamsadeghi/Reducing-the-Computational-Complexity-of-Multicasting-in-Large-Scale-Antenna-Systems.

[34] Technical Specification Group RAN, "Physical layer aspects for e-utra," 3rd Generation Partnership Project (3GPP), Tech. Rep., TS 25.814, 2006.

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

[36] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.