Non-Orthogonal Unicast and Broadcast Transmission via Joint Beamforming and LDM in Cellular Networks

Junlin Zhao, Deniz Gündüz, Osvaldo Simeone, and David Gómez-Barquero

***Abstract*—Limited bandwidth resources and higher energy efficiency requirements motivate incorporating multicast and broadcast transmission into the next-generation cellular network architectures, particularly for multimedia streaming applications. Layered division multiplexing (LDM), a form of non-orthogonal multiple access (NOMA), can potentially improve unicast throughput and broadcast coverage with respect to traditional orthogonal frequency division multiplexing (FDM) or time division multiplexing (TDM), by simultaneously using the same frequency and time resources for multiple unicast or broadcast transmissions. In this paper, the performance of LDM-based unicast and broadcast transmission in a cellular network is studied by assuming a single frequency network (SFN) operation for the broadcast layer, while allowing arbitrarily clustered cooperation among the base stations (BSs) for the transmission of unicast data streams. Beamforming and power allocation between unicast and broadcast layers, the so-called *injection level* in the LDM literature, are optimized with the aim of minimizing the sum-power under constraints on the user-specific unicast rates and on the common broadcast rate. The effects of imperfect channel coding and imperfect channel state information (CSI) are also studied to gain insights into robust implementation in practical systems. The non-convex optimization problem is tackled by means of successive convex approximation (SCA) techniques. Performance upper bounds are also presented by means of the $\mathcal{S}$-procedure followed by semidefinite relaxation (SDR). Finally, a dual decomposition-based solution is proposed to facilitate an efficient distributed implementation of LDM in each of the SCA subproblems, where the unicast beamforming vectors can be obtained locally by the cooperating BSs. Numerical results are presented, which show the tightness of the proposed bounds and hence the near-optimality of the proposed solutions.**

I. Introduction

With the growing demand for multimedia streaming applications, research efforts to incorporate multicast and broadcast transmission into the cellular network architecture have intensified in recent years. In 3G networks, multimedia broadcast multicast services (MBMS) was introduced to support new point-to-multipoint radio bearers and multicast capability in the core network [2]. However, due to its reduced capacity, which did not meet the requirement of mass media services, MBMS has never been deployed commercially. The broadcast extension of 4G LTE is named evolved MBMS (eMBMS), commercially known as LTE Broadcast [3].

Following many field trials worldwide, the first commercial deployment of eMBMS was launched in South Korea in 2014. eMBMS provides full integration and seamless transition between broadcast and unicast modes [4], and significant performance improvement with respect to MBMS, thanks to the higher and more flexible data rates provided by the LTE architecture. Furthermore, it also allows single frequency network (SFN) operation across different cells as in digital television broadcasting, since the LTE waveform is OFDM-based. While it is commonly accepted that eMBMS, in its current form, needs further enhancements to be adopted as a successful commercial platform for TV broadcasting [5], it has been proposed as a converged platform in the UHF band for TV and mobile broadband [6], [7]. For eMBMS TV services, a study has been carried out within 3GPP in 2015 for application scenarios and use cases, as well as for potential requirements and improvements [8]. In 2017, advances have been published in the 3GPP Release 14, including standardization of radio interfaces between mobile network operators and broadcasters and the possibility for free-to-air reception, which is an essential feature for broadcasting TV programs over mobile networks [9]. While the standardization and evolvement of point-to-multipoint transmission are primarily led by multimedia broadcasting services, point-to-multipoint transmission techniques have also been adopted in LTE-Advanced Pro for emerging use cases including vehicular to everything (V2X), Internet of things (IoT) and machine-type communication (MCC) [10]. In 2019, the Study Item on potential enhancements on the existing 5G architecture for 5G multicast-broadcast services has been approved by 3GPP, which opens the door to the standardization of MBMS in the 3GPP Release 17 for 5G [11].

LTE Broadcast entails a reduction in system capacity for unicast services, since eMBMS and unicast services are multiplexed in time in different sub-frames. Superposition coding, a form of non-orthogonal multiple access (NOMA), was proposed in [12] to improve unicast throughput and broadcast cov-

J. Zhao and D. Gündüz have received support from the European Research Council (ERC) through project BEACON (No. 677854), and from the European Unions Horizon 2020 Research and Innovation Programme through project SCAVENGE (No. 675891). O. Simeone has received funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation program (grant agreement 725731). Part of this work was presented at the IEEE Global Communications Conference (Globecom), Washington, D.C., Dec. 2016. [1]

J. Zhao and D. Gündüz are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, UK (e-mail: j.zhao15@imperial.ac.uk, d.gunduz@imperial.ac.uk).

O. Simeone is with the Department of Informatics, King's College London, London WC2R 2LS, UK (e-mail: osvaldo.simeone@kcl.ac.uk).

David Gómez-Barquero is with the Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, València 46022, Spain (e-mail: dagobar@iteam.upv.es).

erage with respect to traditional orthogonal frequency division multiplexing (FDM) or time division multiplexing (TDM), by simultaneously using the same frequency and time resources for multiple unicast or broadcast transmissions. Superposition coding has been adopted in the next-generation TV broadcasting US standard ATSC 3.0 [13] under the name layer division multiplexing (LDM) [14]. To improve the performance of the LTE-based eMBMS service specified in the 3GPP Release 14, LDM has been identified as a candidate technology in LTE-based 5G Terrestrial Broadcast, also known as 5G Broadcast, as part of the 3GPP Release 16 framework [15].

At the cost of an increased complexity at the receivers, which need to perform interference cancellation by decoding the generic broadcast content prior to decoding the unicast content, LDM may provide significant gains especially when the superposed signals exhibit large disparities in terms of signal-to-noise-plus-interference ratio (SINR). This is expected to be the case for multiplexing broadcast and unicast services. In fact, the unicast throughput is limited by intercell interference; and hence, increasing the transmit unicast power across the network does not necessarily improve the unicast SINR. In contrast, broadcast does not suffer from intercell interference in an SFN, and increasing the broadcast power results in an increased SINR. This not only helps improve the reliability of the broadcast layer, but it also reduces the interference on the unicast messages as the broadcast layer can be decoded and canceled more reliably. A performance comparison of LDM with TDM/FDM for unequal error protection in broadcast systems in the absence of multicell interference from an information theoretic perspective can be found in [16].

In this paper, we study the performance of non-orthogonal unicast and broadcast transmission in a cellular network via LDM, in order to demonstrate and quantify its benefits compared to orthogonal transmission methods, i.e., TDM and FDM. We assume an SFN operation for the broadcast layer, while allowing arbitrarily clustered cooperation for the unicast data streams. Cooperative transmission for broadcast traffic, and potentially also for unicast data streams, takes place by means of distributed beamforming at multi-antenna base stations. To better account for potential practical impairments, and to evaluate the robustness of LDM in real systems, we also consider imperfections in channel state information (CSI) through an additive error model. Beamforming and power allocation between unicast and broadcast layers, and the so-called injection level in the LDM literature (see, e.g., [16]), are optimized with the aim of minimizing the sum-power under constraints on the user-specific unicast rates and the common broadcast rate. The optimization of orthogonal transmission via TDM/FDM is also studied for comparison, and the corresponding nonconvex optimization problems are tackled by means of successive convex approximation (SCA) techniques [17], as well as through the calculation of performance upper bounds by means of the S-procedure followed by semidefinite relaxation (SDR) [18].

Finally, we also present an efficient distributed implementation of the proposed LDM system based on the dual decomposition method. The dual decomposition based-algorithm allows each cluster of BSs cooperating to transmit a unicast message to obtain their beamforming vector locally with limited information exchange. A completely distributed implementation is not viable due to the presence of the broadcast layer, whose beamforming vector needs to be determined centrally at one of the BSs or in the cloud; however, local computation of the unicast beamforming vectors allows exploiting the computation resources distributed across the network, which can help parallelize these computations.

With regards to previous work, the optimization of the beamforming vectors in multicell systems has been investigated in [19] and [20], where the base station in each cell multicasts one or more data streams to the specified given groups of in-cell users. The coexistence of broadcast and unicast traffic is studied in [21], where the surplus of degrees-of-freedom provided by massive MIMO systems is leveraged to broadcast data to a group of users whose CSI is not available, without creating interference to conventional unicast users. Recently, the rate splitting technique is considered in [22] to construct the unicast and multicast messages, which are then transmitted through joint beamforming. In [23], joint beamforming for multicast and unicast transmission and BS clustering is considered, which is shown to improve the system performance as compared to fixed BS clustering in [1]. Robust coordinated beamforming in a multicell network with imperfect CSI is studied in [24], where the optimization problem is solved by a second-order cone program after relaxing the worst-case SINR requirement. The same problem is also studied in [25], [26], and [27], where the infinitely many constraints introduced due to the imperfect channel estimation are tackled by the S-procedure.

Distributed implementations of multigroup multicast beamforming have also been a focus in the literature. A dual decomposition-based scheme has been proposed in [28] by creating consensus over inter-cell interference terms between all the BSs. In [20], a primal decomposition-based algorithm and an alternating direction method of multipliers (ADMM)-based algorithm have been proposed for the SDR version of the original problem. In [29], instead of directly dealing with the relaxed problem, the authors proposed to apply ADMM for each of the convexified SCA problems, obtaining a double-loop scheme. In [27], an ADMM-based algorithm is proposed for a distributed solution of the problem with imperfect CSI after relaxing the original problem with S-procedure.

The rest of this paper is organized as follows. Section II introduces the system model and the problem formulation. In Section III, the characterized problem is tackled by using the S-procedure and the SCA technique. Dual decomposition-based distributed algorithms for both TDM and LDM are developed in Section IV. Numerical results are presented in Section V, followed by the conclusions in Section VI.

II. System Model and Problem Formulation

In this section, we present the model of the joint unicast and broadcast transmission system under study, by highlighting orthogonal and non-orthogonal multiplexing schemes. For both schemes, we formulate a power minimization problem under user quality of service (QoS) constraints.

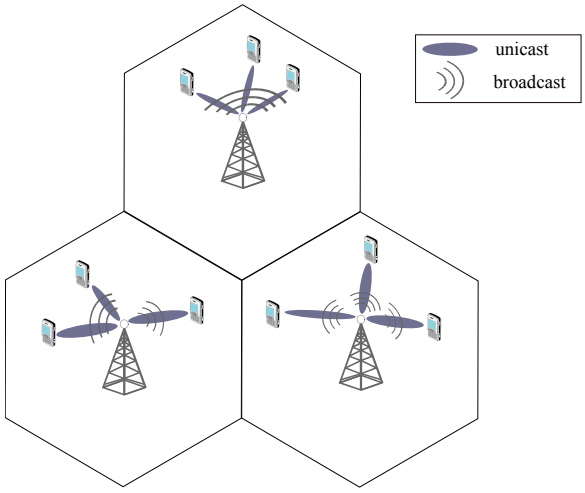

Fig. 1. Illustration of a multicell network with N=3 cells and $K = 3$ users in each cell with simultaneous unicast and broadcast transmission.

A. System Model

We investigate downlink transmission in a cellular network that serves both unicast and broadcast traffic. Specifically, we focus on a scenario in which a dedicated unicast data stream is to be delivered to each user, while there is a common broadcast data stream intended for all the users. A more general broadcast traffic model, in which distinct data streams are sent to different subsets of users, could be included in the analysis at the cost of a more cumbersome notation, but will not be further pursued in this paper.

As illustrated in Fig. 1, the network is comprised of N cells, each consisting of a base station (BS) with M antennas and K single-antenna mobile users. The notation (n, k) identifies the k-th user in cell n. All BSs cooperate via joint beamforming for the broadcast stream to all the users, while an arbitrary cluster $\mathcal{C}_{n,k}$ of BSs cooperate for the unicast transmission to user (n, k). Accordingly, all the BSs have access to the broadcast data stream, while only the BSs in cluster $\mathcal{C}_{n,k}$ are informed about the unicast data stream to be delivered to user (n, k). Note that, non-cooperative unicast transmission, whereby each BS serves only the users in its own cell, can be obtained as a special case when $\mathcal{C}_{n,k} = \{n\}$, for all users (n, k). Similarly, fully cooperative unicast transmission is obtained when $\mathcal{C}_{n,k} = \{1, \ldots, N\}$, for all users (n, k). We denote the set of users whose unicast messages are available at BS i as

$$\mathcal{U}_i = \{(n, k) \mid i \in \mathcal{C}_{n,k}\}. \tag{1}$$

We assume frequency-flat quasi-static complex channels, and define $\boldsymbol{h}_{i,n,k} \in \mathbb{C}^{M\times 1}$ as the channel vector from the BS in cell i to user (n, k). We use the notation $s^U_{n,k}$ to denote an encoded unicast symbol intended for user (n, k), and s^B to represent an encoded broadcast symbol. The signal received by user (n, k) at any given channel use can then be written as

$$y_{n,k} = \sum_{i=1}^{N} \boldsymbol{h}^H_{i,n,k}\boldsymbol{x}_i + n_{n,k}, \tag{2}$$

where $\boldsymbol{x}_i \in \mathbb{C}^{M\times 1}$ is the symbol transmitted by BS i, and $n_{n,k} \sim \mathcal{CN}(0, \sigma^2_{n,k})$ is the additive white Gaussian noise. We assume that both the intended and the interference signals at each user are in perfect synchronization without inter-symbol interference.

In practice, BSs have to operate with imperfect CSI. In Frequency Division Duplex (FDD) systems, it may arise from errors in downlink training-based CSI estimation, limited resolution in CSI feedback links, or from delays in CSI acquisition over fading channels, while in Time Division Duplex (TDD) systems, CSI errors are caused by impairments in channel estimation or imperfect channel reciprocity (see [26] and references therein). As common in the literature, we model the CSI uncertainty with an additive error by setting

$$\boldsymbol{h}_{i,n,k} = \hat{\boldsymbol{h}}_{i,n,k} + \boldsymbol{e}_{i,n,k}, \tag{3}$$

where $\hat{\boldsymbol{h}}_{i,n,k} \in \mathbb{C}^{M\times 1}$ is the estimated complex channel vector from cell i to user (n, k) available at the BSs, and $\boldsymbol{e}_{i,n,k} \in \mathbb{C}^{M\times 1}$ is the additive channel error. We consider a bounded error, which is typically used to model CSI imperfections resulting from quantization error due to feedback links of limited capacity. Hence, the set of channel vectors from BS i to user (n, k) can be defined as

$$\mathcal{H}_{i,n,k} = \{\boldsymbol{h}_{i,n,k} : \boldsymbol{h}_{i,n,k} = \hat{\boldsymbol{h}}_{i,n,k} + \boldsymbol{e}_{i,n,k}, \boldsymbol{e}^H_{i,n,k}\boldsymbol{Q}_{i,n,k}\boldsymbol{e}_{i,n,k} \le 1\},\ \forall i, n, k, \tag{4}$$

where $\boldsymbol{Q}_{i,n,k}$ is a known positive definite matrix. Accordingly, the structure of the uncertainty set of the quantization error vectors is known at the transmitters.

In what follows, we will consider two modes of transmission, namely orthogonal transmission via TDM and non-orthogonal transmission via LDM, where the former will serve as a benchmark to evaluate the potential performance gains from the LDM scheme.

1) TDM: We first consider the standard TDM approach based on the orthogonal transmission of unicast and broadcast signals. Note that orthogonalization can also be realized by means of other multiplexing schemes such as FDM, yielding the same mathematical formulation. With TDM, each transmission slot of duration T channel uses is divided into two subslots: a subslot of duration T_0 channel uses for unicast transmission, and a subslot of duration $T - T_0$ for broadcast transmission. Therefore, the signal $\boldsymbol{x}_i$ transmitted by cell i can be written as

$$\boldsymbol{x}_i = \begin{cases} \sum\limits_{(n,k)\in\mathcal{U}_i} \boldsymbol{w}^U_{i,n,k}s^U_{n,k} & \text{for } 0 \le t < T_0 \\ \boldsymbol{w}^B_i s^B & \text{for } T_0 \le t < T \end{cases}, \tag{5}$$

where $\boldsymbol{w}^U_{i,n,k} \in \mathbb{C}^{M\times 1}$ represents the unicast beamforming vector applied at the BS in cell i towards user (n, k), and $\boldsymbol{w}^B_i \in \mathbb{C}^{M\times 1}$ is the broadcast beamforming vector applied at the same BS.

The received signal $y_{n,k}$ at user (n, k) can be expressed as

$$y_{n,k} = \begin{cases} \Big(\sum\limits_{i\in\mathcal{C}_{n,k}} \boldsymbol{h}^H_{i,n,k}\boldsymbol{w}^U_{i,n,k}\Big)s^U_{n,k} + z_{n,k} + n_{n,k} & \text{for } 0 \le t < T_0 \\ \Big(\sum\limits_{i=1}^{N} \boldsymbol{h}^H_{i,n,k}\boldsymbol{w}^B_i\Big)s^B + n_{n,k} & \text{for } T_0 \le t < T \end{cases}, \tag{6}$$

where

$$z_{n,k} = \sum_{(p,q)\neq(n,k)} \Big(\sum_{i\in\mathcal{C}_{p,q}} \boldsymbol{h}_{i,n,k}^H \boldsymbol{w}_{i,p,q}^U \Big) s_{p,q}^U \tag{7}$$

denotes the interference at user (n,k).

2) LDM: In LDM, the transmitted signal $\boldsymbol{x}_i$ from the BS in cell i is the superposition of the broadcast and unicast signals for the entire time slot, which can be written as

$$\boldsymbol{x}_i = \boldsymbol{w}_i^B s^B + \sum_{(n,k)\in\mathcal{U}_i} \boldsymbol{w}_{i,n,k}^U s_{n,k}^U \quad \text{for } 0 \le t \le T, \tag{8}$$

for all channel uses in an entire time slot, i.e., for $0 \le t \le T$. We note that the power ratio between broadcast and unicast, which is referred to as the *injection level* (IL) in the literature (see, e.g., [16]), can be obtained as

$$\text{IL} = 10\,\log_{10}\frac{P^B}{P^U}, \tag{9}$$

where $P^B = \sum_{i=1}^N ||\boldsymbol{w}_i^B||^2$ is the total broadcast power, and $P^U = \sum_{i=1}^N \sum_{(n,k)\in\mathcal{U}_i} ||\boldsymbol{w}_{i,n,k}^U||^2$ is the total unicast power. The received signal at user (n,k) is given by

$$\begin{aligned} y_{n,k} \;=\; & \Big(\sum_{i=1}^N \boldsymbol{h}_{i,n,k}^H \boldsymbol{w}_i^B \Big) s^B + \Big(\sum_{i\in\mathcal{C}_{n,k}} \boldsymbol{h}_{i,n,k}^H \boldsymbol{w}_{i,n,k}^U \Big) s_{n,k}^U \\ & + z_{n,k} + n_{n,k}, \quad \text{for } 0 \le t \le T, \end{aligned} \tag{10}$$

where $z_{n,k}$ is the interference as defined in (7).

B. Problem Formulation

The power minimization problem for the above systems can be expressed in the following form:

$$\min_{\{\boldsymbol{w}_i^B\},\{\boldsymbol{w}_{i,n,k}^U\}} \quad \sum_{i=1}^N \Big(||\boldsymbol{w}_i^B||^2 + \sum_{(n,k)\in\mathcal{U}_i} ||\boldsymbol{w}_{i,n,k}^U||^2 \Big) \tag{11a}$$

$$\text{s.t.} \quad \min_{\mathcal{H}} \ \text{SINR}_{n,k}^B \ge \gamma^B, \ \forall n,k, \tag{11b}$$

$$\min_{\mathcal{H}} \ \text{SINR}_{n,k}^U \ge \gamma_{n,k}^U, \ \forall n,k, \tag{11c}$$

where the explicit expressions for the SINRs at user (n,k) for broadcast and unicast transmissions, namely $\text{SINR}_{n,k}^B$ and $\text{SINR}_{n,k}^U$ will be given below for TDM and LDM separately. The constraints in (11b) and (11c) are imposed on the worst-case SINRs for all possible channel realizations in the set $\mathcal{H} = \prod_{i,n,k} \mathcal{H}_{i,n,k}$. Note that, since all the users receive the same broadcast signal, we have enforced a common broadcast QoS requirement. In contrast, the unicast SINR requirements are allowed to be user-dependent.

1) TDM: From the expression of the received signal in (6), we derive the SINR for the broadcast layer in TDM for user (n,k) as

$$\text{SINR}_{n,k}^{B\text{-TDM}} = \frac{|\boldsymbol{h}_{n,k}^H \boldsymbol{w}^B|^2}{\sigma_{n,k}^2}, \tag{12}$$

where $\boldsymbol{h}_{n,k} = [\boldsymbol{h}_{1,n,k}^T, \ldots, \boldsymbol{h}_{N,n,k}^T]^T \in \mathbb{C}^{NM\times 1}$ is the aggregated channel vector from all the BSs to user (n,k). All broadcast beamforming vectors are similarly aggregated into the vector $\boldsymbol{w}^B = [\boldsymbol{w}_1^{B^T}, \ldots, \boldsymbol{w}_N^{B^T}]^T \in \mathbb{C}^{NM\times 1}$. The SINR for the unicast layer is instead given as

$$\text{SINR}_{n,k}^{U\text{-TDM}} = \frac{|\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U|^2}{\sum_{(p,q)\neq(n,k)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U|^2 + \sigma_{n,k}^2}, \tag{13}$$

where $\boldsymbol{h}_{n,k}^{(p,q)} = [\boldsymbol{h}_{i,n,k}^T]_{i\in\mathcal{C}_{p,q}}^T$ is the aggregated channel vector to user (n,k) from all the BSs in cluster $\mathcal{C}_{p,q}$ of BSs that serve user (p,q), and $\boldsymbol{w}_{n,k}^U = [\boldsymbol{w}_{i,n,k}^{U^T}]_{i\in\mathcal{C}_{n,k}}^T$ is similarly defined as the aggregate unicast beamforming vector for user (n,k) from all the BSs in cluster $\mathcal{C}_{n,k}$.

We observe that the SINR targets $\gamma_{n,k}^{U\text{-TDM}}$ and $\gamma^{B\text{-TDM}}$ for unicast and broadcast traffic can be obtained from the corresponding transmission rates $R_{n,k}^U$ and R^B, respectively, as

$$\frac{T_0}{T}\log_2(1+\gamma_{n,k}^{U\text{-TDM}}) = R_{n,k}^U, \tag{14}$$

and

$$\frac{T-T_0}{T}\log_2(1+\gamma^{B\text{-TDM}}) = R^B. \tag{15}$$

2) LDM: With LDM, the broadcast layer, which is intended for all the users and usually has a higher SINR, is decoded first by treating unicast signals as noise, as in [12]. The users decode their unicast data streams after canceling the decoded broadcast message. The broadcast SINR in LDM for user (n,k) is hence obtained from the received signal (10) as follows

$$\text{SINR}_{n,k}^{B\text{-LDM}} = \frac{|\boldsymbol{h}_{n,k}^H \boldsymbol{w}^B|^2}{\sum_{(p,q)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U|^2 + \sigma_{n,k}^2}, \tag{16}$$

while the unicast SINR is the same as TDM given in (13), i.e.,

$$\text{SINR}_{n,k}^{U\text{-LDM}} = \text{SINR}_{n,k}^{U\text{-TDM}}. \tag{17}$$

Similarly to TDM, SINR thresholds for unicast and broadcast can be obtained from the transmission rates $R_{n,k}^U$ and R^B, respectively, as

$$\log_2(1+\gamma_{n,k}^{U\text{-LDM}}) = R_{n,k}^U, \tag{18}$$

and

$$\log_2(1+\gamma^{B\text{-LDM}}) = R^B. \tag{19}$$

In [1], a performance lower bound on the power minimization problem is obtained by standard semidefinite relaxation (SDR), assuming that perfect CSI is available at all the BSs. In this paper, the problem formulation incorporates CSI uncertainty in (11b) and (11c) by imposing constraints on the worst-case performance over all possible channel realizations on the optimization problem. The formulated worst-case quadratically-constrained quadratic program (QCQP) is intractable due to the induced additional constraints on the CSI error vectors. Nevertheless, the uncertainty due to CSI errors can be tackled by applying the S-procedure as in [26], as a result of which SDR can be employed as in the perfect CSI case to obtain a lower bound on the optimal

solution. Furthermore, an achievable beamformer design under the worst-case SINR constraints will be obtained based on SCA, and its performance will be compared with the obtained lower bound.

III. Bounds on the minimum total power

The optimization problem formulated in (11) is nonconvex due to the QoS constraints in (11b) and (11c). Therefore, there are in general no numerical solution techniques with guaranteed convergence to a global optimal solution. In this section, we will present numerical tools to obtain lower and upper bounds on the minimum total transmit power.

A. Lower Bound via S-Procedure

The optimization problem in (11) contains an infinite number of constraints in (11b) and (11c), thus it is intractable. To address this issue, S-procedure [18] will be adopted to derive an equivalent but tractable problem formulation. Following the CSI error model in (4) we can form the aggregated CSI error vector $\boldsymbol{e}_{n,k}$ for user (n,k) consistent with the aggregated channel vector $\boldsymbol{h}_{n,k}$, and define the relaxed set of possible channel vectors to user (n,k) as:

$$\mathcal{H}_{n,k} \triangleq \{\boldsymbol{h}_{n,k} : \boldsymbol{h}_{n,k} = \hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k},\ \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1\}, \tag{20}$$

where

$$\boldsymbol{Q}_{n,k} \triangleq \frac{1}{N} \begin{bmatrix} \boldsymbol{Q}_{1,n,k} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{Q}_{N,n,k} \end{bmatrix}. \tag{21}$$

It is noted that the set of possible channel vectors in (20) is a relaxed version of the original set given in (4). For reference, we present the S-procedure in the following lemma for completeness.

Lemma 1 (S-procedure): Let $f_i(\boldsymbol{x}) \triangleq \boldsymbol{x}^H \boldsymbol{F}_i \boldsymbol{x} + \boldsymbol{g}_i^H \boldsymbol{x} + \boldsymbol{x}^H \boldsymbol{g}_i + c_i,$ for $i = 0, 1$, where $\boldsymbol{F}_i \in \mathbb{C}^{NM \times NM}$ is Hermitian semidefinite, $\boldsymbol{g} \in \mathbb{C}^{NM \times 1}$, and $c_i \in \mathbb{R}$, then $f_1(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x}$ satisfying $f_0(\boldsymbol{x}) \leq 0$ holds if and only if there exists a $\lambda \geq 0$ such that

$$\begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{g}_1 \\ \boldsymbol{g}_1^H & c_1 \end{bmatrix} \preceq \lambda \begin{bmatrix} \boldsymbol{F}_0 & \boldsymbol{g}_0 \\ \boldsymbol{g}_0^H & c_0 \end{bmatrix}. \tag{22}$$

1) TDM: The constraint for the broadcast layer in (11b) can be rewritten as

$$(\hat{\boldsymbol{h}}_{n,k}^H + \boldsymbol{e}_{n,k}^H) \boldsymbol{W}^B (\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k}) \geq \sigma_{n,k}^2 \gamma_{n,k}^B, \text{for} \, \forall \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1,$$

where $\boldsymbol{W}^B \triangleq \boldsymbol{w}^B \boldsymbol{w}^{B^H}$. By applying the S-procedure, the worst-case SINR constraint in (11b) can be recast as

$$\begin{bmatrix} \boldsymbol{W}^B & \boldsymbol{W}^B \hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{W}^B & \frac{1}{\gamma_{n,k}^B} \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{W}^B \hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^B \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \tag{23}$$

for some $\lambda_{n,k}^B \geq 0,\ \forall n, k$. Accordingly to Lemma 1, the constraints on the unicast transmissions in (11c) can be written as

$$(\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k})^H \Big(\frac{1}{\gamma_{n,k}^U} \boldsymbol{T}_{n,k}^T \boldsymbol{W}_{n,k}^U \boldsymbol{T}_{n,k} - \sum_{(p,q) \neq (n,k)} \boldsymbol{T}_{p,q}^T \boldsymbol{W}_{p,q}^U \boldsymbol{T}_{p,q} \Big) \cdot \cdot (\hat{\boldsymbol{h}}_{n,k} + \boldsymbol{e}_{n,k}) \geq \sigma_{n,k}^2, \ \text{for} \ \forall \boldsymbol{e}_{n,k}^H \boldsymbol{Q}_{n,k} \boldsymbol{e}_{n,k} \leq 1, \tag{24}$$

where $\boldsymbol{W}_{n,k}^U \triangleq \boldsymbol{w}_{n,k}^U \boldsymbol{w}_{n,k}^{U^H}$, and $\boldsymbol{T}_{p,q}$ is a constructed block matrix of dimension $|\mathcal{C}_{p,q}| \times N$ such that $\boldsymbol{h}_{n,k}^{(p,q)} = \boldsymbol{T}_{p,q} \boldsymbol{h}_{n,k}$. Following the S-procedure, the worst-case SINR constraint for the unicast layer can be recast as

$$\begin{bmatrix} \boldsymbol{V}_{n,k} & \boldsymbol{V}_{n,k} \hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{V}_{n,k} & \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{V}_{n,k} \hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^U \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \forall n, k, \tag{25}$$

for some $\lambda_{n,k}^U \geq 0,\ \forall n, k$, where $\boldsymbol{V}_{n,k}$ is defined as

$$\boldsymbol{V}_{n,k} \triangleq \frac{1}{\gamma_{n,k}^U} \boldsymbol{T}_{n,k}^T \boldsymbol{W}_{n,k}^U \boldsymbol{T}_{n,k} - \sum_{(p,q) \neq (n,k)} \boldsymbol{T}_{p,q}^T \boldsymbol{W}_{p,q}^U \boldsymbol{T}_{p,q}. \tag{26}$$

Following these transforms and definitions, the problem in (11) can be relaxed to a tractable semidefinite program by dropping the rank constraints on matrices $\boldsymbol{W}^B$ and $\boldsymbol{W}_{n,k}^U$. Specifically, for TDM, the relaxed problem after SDR is given by

$$\min_{\boldsymbol{W}^B, \{\boldsymbol{W}_{n,k}^U\}, \{\lambda_{n,k}^B\}, \{\lambda_{n,k}^U\}} \ \text{tr}(\boldsymbol{W}^B) + \sum_{n=1}^N \sum_{k=1}^K \text{tr}(\boldsymbol{W}_{n,k}^U) \tag{27a}$$

$$\text{s.t.} \ \ (23) \text{ and } (25), \tag{27b}$$

$$\lambda_{n,k}^B \geq 0, \lambda_{n,k}^U \geq 0, \ \forall n, k. \tag{27c}$$

2) LDM: Similar to the analysis in TDM, the constraint on the broadcast transmission in LDM can be equivalently written as

$$\begin{bmatrix} \boldsymbol{U} & \boldsymbol{U} \hat{\boldsymbol{h}}_{n,k} \\ \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{U} & \hat{\boldsymbol{h}}_{n,k}^H \boldsymbol{U} \hat{\boldsymbol{h}}_{n,k} - \sigma_{n,k}^2 \end{bmatrix} + \lambda_{n,k}^B \begin{bmatrix} \boldsymbol{Q}_{n,k} & \boldsymbol{0} \\ \boldsymbol{0}^T & -1 \end{bmatrix} \succeq 0, \tag{28}$$

where $\lambda_{n,k}^U \geq 0,\ \forall n, k$, and $\boldsymbol{U}$ is defined as

$$\boldsymbol{U} \triangleq \frac{1}{\gamma_{n,k}^B} \boldsymbol{W}^B - \sum_{(p,q)} \boldsymbol{T}_{p,q}^T \boldsymbol{W}_{p,q}^U \boldsymbol{T}_{p,q}. \tag{29}$$

The unicast constraint in LDM can be reformulated as in (25), hence the relaxed problem after dropping the rank-1 constraints on matrices $\boldsymbol{W}^B$ and $\boldsymbol{W}_{n,k}^U$ is obtained as follows:

$$\min_{\boldsymbol{W}^B, \{\boldsymbol{W}_{n,k}^U\}, \{\lambda_{n,k}^B\}, \{\lambda_{n,k}^U\}} \ \text{tr}(\boldsymbol{W}^B) + \sum_{n=1}^N \sum_{k=1}^K \text{tr}(\boldsymbol{W}_{n,k}^U) \tag{30a}$$

$$\text{s.t.} \ \ (25) \text{ and } (28), \tag{30b}$$

$$\lambda_{n,k}^B \geq 0, \lambda_{n,k}^U \geq 0, \ \forall n, k. \tag{30c}$$

As the rank-1 constraint has been dropped in (27) and (30), the corresponding optimal solutions provide lower bounds on the optimal solutions of the original problems in (11). Note that,

under perfect CSI, i.e., $\boldsymbol{e}_{i,n,k}=\mathbf{0}$, the problem formulation in (11) boils down to the one presented in [1], and the solution obtained by first applying the S-procedure is equal to that obtained directly by SDR.

B. Upper Bound via SCA

Instead of adopting Gaussian randomization [30] to obtain a feasible (achievable) beamforming scheme, we leverage the SCA method [17] to obtain an achievable beamformer, which yields an upper bound on the minimum required power. In particular, by rewriting the nonconvex QoS constraints as the difference of convex (DC) functions, the SCA algorithm reduces to the conventional convex-concave procedure [31]. We remark that the SCA scheme is known to converge to a stationary point of the original problem [17].

In order to apply the SCA approach, each nonconvex constraint in (11) will be expressed as

$$g(\boldsymbol{w})=g^{+}(\boldsymbol{w})-g^{-}(\boldsymbol{w})\leq 0, \tag{31}$$

where $g^{+}(\boldsymbol{w})$ and $g^{-}(\boldsymbol{w})$ are both convex functions on the set of all beamforming vectors $\boldsymbol{w}$. Then a convex upper bound is obtained by linearizing the nonconvex part around any given vector $\boldsymbol{u}$, yielding the stricter constraint on the solution $\boldsymbol{w}$ as

$$\tilde{g}(\boldsymbol{w};\boldsymbol{u})\triangleq g^{+}(\boldsymbol{w})-g^{-}(\boldsymbol{u})-\bigtriangledown_{\boldsymbol{w}}g^{-}(\boldsymbol{u})^{T}(\boldsymbol{w}-\boldsymbol{u})\leq 0. \tag{32}$$

1) TDM: The constraint in (11b) on the broadcast layer can be approximated and replaced by the following tighter constraint:

$$|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}|-|\boldsymbol{e}_{n,k}^{H}\boldsymbol{w}^{B}|\geq\sqrt{\gamma^{B}}\sigma_{n,k}\ \text{for}\ \forall\ \boldsymbol{e}_{n,k}^{H}\boldsymbol{Q}_{n,k}\boldsymbol{e}_{n,k}\leq 1, \tag{33}$$

which can be further tightened as:

$$|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}|-\|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\|\geq\sqrt{\gamma^{B}}\sigma_{n,k}, \tag{34}$$

since $|\boldsymbol{e}_{n,k}^{H}\boldsymbol{w}^{B}|\leq\|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\|$ holds for the CSI error vectors $\boldsymbol{e}_{n,k}$ as we have $\boldsymbol{e}_{n,k}\in\{\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{u}\mid\|\boldsymbol{u}\|\leq 1\}$.

The constraint in (34) is in the DC form, for which SCA can be adopted to obtain an iterative algorithm which converges to a stationary point of the original problem. The constraint at the ν-th iteration of the SCA algorithm is given by

$$\sqrt{\gamma^{B}}\sigma_{n,k}+\|\boldsymbol{Q}_{n,k}^{-\frac{1}{2}}\boldsymbol{w}^{B}\|+|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)| -2\frac{\hat{\boldsymbol{h}}_{n,k}^{H}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^{H}}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{H}\boldsymbol{w}^{B}(\nu)|}\boldsymbol{w}^{B}\leq 0,\ \forall n,k. \tag{35}$$

Also, the constraint in (11c) for the unicast transmission can be tightened by considering the worst-case SINR, i.e.,

$$\frac{\min\limits_{\mathcal{H}}\ |\boldsymbol{h}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}|^{2}}{\max\limits_{\mathcal{H}}\sum\limits_{(p,q)\neq(n,k)}|\boldsymbol{h}_{n,k}^{(p,q)^{H}}\boldsymbol{w}_{p,q}^{U}|^{2}+\sigma_{n,k}^{2}}\geq\gamma_{n,k}^{U},\ \text{for}\ \forall n,k, \tag{36}$$

TABLE I
SCA ALGORITHM

STEP 0: Set $\nu=1$. Set a step size μ. Initialize $\boldsymbol{w}^{B}(1)$ and $\boldsymbol{w}_{n,k}^{U}(1)$ with feasible values
STEP 1: If a stopping criterion is satisfied, then STOP
STEP 2: Set $\boldsymbol{w}^{B}(\nu+1)=\boldsymbol{w}^{B}(\nu)+\mu(\boldsymbol{w}^{B}-\boldsymbol{w}^{B}(\nu))$, $\boldsymbol{w}_{n,k}^{U}(\nu+1)=\boldsymbol{w}_{n,k}^{U}(\nu)+\mu(\boldsymbol{w}_{n,k}^{U}-\boldsymbol{w}_{n,k}^{U}(\nu))$, where $\{\boldsymbol{w}^{B}\}$ and $\{\boldsymbol{w}_{n,k}^{U}\}$ are obtained as solutions of problems (42) for TDM and (49) for LDM
STEP 3: Set $\nu=\nu+1$, and go to STEP 1

which can then be replaced equivalently by the following set of constraints:

$$\max_{\mathcal{H}}\ |\boldsymbol{h}_{n,k}^{(p,q)^{H}}\boldsymbol{w}_{p,q}^{U}|\leq\beta_{n,k}^{(p,q)},\ \forall n,k,\forall(p,q)\neq(n,k), \tag{37a}$$

$$\min_{\mathcal{H}}\ |\boldsymbol{h}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}|\geq t_{n,k}^{U}, \tag{37b}$$

$$\gamma_{n,k}^{U}\Big(\sum_{(p,q)\neq(n,k)}\beta_{n,k}^{(p,q)^{2}}+\sigma_{n,k}^{2}\Big)-t_{n,k}^{U^{2}}\leq 0, \tag{37c}$$

where $\{t_{n,k}^{U}\}$ and $\{\beta_{n,k}^{(p,q)}\}$ are auxiliary variables. Note that $\beta_{n,k}^{(p,q)}$ indicates the interference power from BSs in the cluster $\mathcal{C}_{p,q}$ to user (n,k), and $t_{n,k}^{U}$ indicates the received unicast power at user (n,k). The constraint in (37a) and (37b) can be further relaxed by

$$|\hat{\boldsymbol{h}}_{n,k}^{(p,q)^{H}}\boldsymbol{w}_{p,q}^{U}|+|\boldsymbol{Q}_{n,k}^{(p,q)^{-1/2}}\boldsymbol{w}_{p,q}^{U}|\leq\beta_{n,k}^{(p,q)},\ \forall n,k,\forall(p,q)\neq(n,k), \tag{38}$$

and

$$t_{n,k}^{U}+\|\boldsymbol{Q}_{n,k}^{(n,k)^{-1/2}}\boldsymbol{w}_{n,k}^{U}\|-|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}|\leq 0, \tag{39}$$

respectively, where $\boldsymbol{Q}_{n,k}^{(p,q)^{-1/2}}=\boldsymbol{Q}_{n,k}^{-1/2}\boldsymbol{T}_{p,q}$. According to (31) and (32), in the SCA algorithm, the corresponding constraints in the ν-th iteration for (37c) and (39) can be written as

$$\gamma_{n,k}^{U}\Big(\sum_{(p,q)\neq(n,k)}\beta_{n,k}^{(p,q)^{2}}+\sigma_{n,k}^{2}\Big)+t_{n,k}^{U^{2}}(\nu)-2t_{n,k}^{U}(\nu)t_{n,k}^{U}\leq 0,\forall n,k, \tag{40}$$

and

$$t_{n,k}^{U}+\|\boldsymbol{Q}_{n,k}^{(n,k)^{-\frac{1}{2}}}\boldsymbol{w}_{n,k}^{U}\|+|\boldsymbol{h}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}(\nu)| -2\frac{\hat{\boldsymbol{h}}_{n,k}^{(n,k)^{H}}\hat{\boldsymbol{h}}_{n,k}^{(n,k)}\boldsymbol{w}_{n,k}^{U^{H}}(\nu)}{|\hat{\boldsymbol{h}}_{n,k}^{(n,k)^{H}}\boldsymbol{w}_{n,k}^{U}(\nu)|}\boldsymbol{w}_{n,k}^{U}\leq 0,\ \forall n,k, \tag{41}$$

respectively.

Due to the fact that the feasible convexified constraints in (35), (38), (40) and (41) are stricter than the original constraints in (11), the solution obtained at each iteration is feasible for the original problem (11) as long as a feasible initial point is available. When the stopping criterion is satisfied, we take the last iteration as the solution of the SCA algorithm. Please refer to Table I for an algorithmic description of the SCA approach.

When obtaining the numerical results in the next section, initialization of the SCA algorithm is carried out based on the

solution $\{\boldsymbol{W}^B\}$ and $\{\boldsymbol{W}^U_{n,k}\}$ obtained from the S-procedure. Specifically, we perform a rank-1 reduction of matrices $\{\boldsymbol{W}^B\}$ and $\{\boldsymbol{W}^U_{n,k}\}$, obtaining vectors $\{\boldsymbol{w}^B\}$ and $\{\boldsymbol{w}^U_{n,k}\}$, respectively, as the largest principal component. These vectors are then scaled with the smallest common factor t, which is evaluated through line search, to satisfy constraints (11b) and (11c), yielding the initial points $\{\boldsymbol{w}^B(1)\}$ and $\{\boldsymbol{w}^U_{n,k}(1)\}$ for SCA. If a feasible value for t is not found through a line search, then the SCA method is considered to be infeasible. Further discussion on this point can be found in Section V.

As a summary, the relaxed version of the problem for (11) in TDM in the SCA form is given as

$$\min_{\boldsymbol{w}^B,\{\boldsymbol{w}^U_{n,k}\},\{\beta^{(p,q)}_{n,k}\},\{t^U_{n,k}\}} \|\boldsymbol{w}^B\|^2+\sum_{(n,k)}\|\boldsymbol{w}^U_{n,k}\|^2 \quad (42a)$$

$$\text{s.t.} \quad (35),\ (38),\ (40),\ \text{and}\ (41). \quad (42b)$$

2) LDM: Similarly to the TDM approach, the constraint in (11b) can be relaxed as the worst-case SINR constraint, i.e.,

$$\frac{\min\limits_{\mathcal{H}}\ |\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B|^2}{\max\limits_{\mathcal{H}}\sum\limits_{(p,q)}|\boldsymbol{h}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}|^2+\sigma^2_{n,k}}\geq\gamma^B, \quad (43)$$

which is then replaced by the following equivalent constraints:

$$\max_{\mathcal{H}}\ |\boldsymbol{h}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}|\leq\beta^{(p,q)}_{n,k}, \quad (44a)$$

$$\min_{\mathcal{H}}\ |\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B|\geq t^B_{n,k}, \quad (44b)$$

$$\gamma^U_{n,k}\Big(\sum_{(p,q)}\beta^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\Big)-t^{B^2}_{n,k}\leq 0 \quad (44c)$$

for all n,k, where $\{t^B_{n,k}\}$ are auxiliary variables indicating the received broadcast power at user (n,k). Similarly to the relaxation we adopt for the TDM case, the constraint in (44a) can be relaxed as

$$|\hat{\boldsymbol{h}}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}|+|\boldsymbol{Q}^{(p,q)^{-1/2}}_{n,k}\boldsymbol{w}^U_{p,q}|\leq\beta^{(p,q)}_{n,k},\ \forall n,k,p,q \quad (45)$$

for all n,k. The constraint in (44b) can be relaxed as

$$t^B_{n,k}+\|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\|-|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B|\leq 0, \quad (46)$$

which is in the convex-concave form. According to (31) and (32), in the SCA algorithm, the corresponding constraints in the ν-th iteration for (44c) and (46) can be written as

$$\gamma^B_{n,k}\Big(\sum_{(p,q)}\beta^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\Big)+t^{B^2}_{n,k}(\nu)-2t^B_{n,k}(\nu)t^B_{n,k}\leq 0,\forall n,k, \quad (47)$$

and

$$t^B_{n,k}+\|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\|+|\boldsymbol{h}^H_{n,k}\boldsymbol{w}^B(\nu)| -2\frac{\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)|}\boldsymbol{w}^B\leq 0,\ \forall n,k, \quad (48)$$

respectively. As a summary, the relaxed version of the (11) for LDM in the SCA form is given as

$$\min_{\boldsymbol{w}^B,\{\boldsymbol{w}^U_{n,k}\},\{\beta^{(p,q)}_{n,k}\},\{t^B_{n,k}\},\{t^U_{n,k}\}} \|\boldsymbol{w}^B\|^2+\sum_{(n,k)}\|\boldsymbol{w}^U_{n,k}\|^2 \quad (49a)$$

$$\text{s.t.}\ (40),\ (41),\ (45),(47),\ \text{and}\ (48). \quad (49b)$$

IV. Dual Decomposition-based Distributed Optimization

In this section, we propose a distributed algorithm to solve the SCA problem in (49) using dual decomposition as in [17]. In particular, while the broadcast beamforming vector $\boldsymbol{w}^B$ is designed at a central node that gathers full CSI between all the BSs and the users, the optimization of unicast beamforming vectors $\{\boldsymbol{w}^U_{n,k}\}$ is offloaded to the processing unit of the corresponding cluster $\mathcal{C}_{n,k}$, which can be located at one of the BSs within the cluster. This distributed implementation is made possible by the fact that the optimization of $\{\boldsymbol{w}^U_{n,k}\}$ can be decomposed into NK independent subproblems, and the processing unit of each cluster $\mathcal{C}_{n,k}$ can calculate $\boldsymbol{w}^U_{n,k}$ locally, but still optimally, based only on local CSI, in addition to certain limited information exchange with other clusters.

The benefits of this distributed implementation are as follows. First, it reduces the computational requirements on the central processing unit as compared to the centralized approach. This is done by parallelizing the computation by distributing it across many nodes in the network. Second, transmitting all the CSI back to a central unit may lead to increased CSI uncertainty, as the CSI could need to be further compressed to be communicated to a single node. In our formulation here, for simplicity, we consider the same CSI error variance for both the broadcast and unicast beamforming optimization problems. Finally, in the absence of a broadcast message destined for the whole network, all computations can be carried out locally at the cluster heads.

For clarity, to start, we reproduce the problem in (49):

$$\min\ \|\boldsymbol{w}^B\|^2+\sum_{(n,k)}\|\boldsymbol{w}^U_{n,k}\|^2 \quad (50a)$$

$$\text{s.t.}\ |\hat{\boldsymbol{h}}^{(p,q)^H}_{n,k}\boldsymbol{w}^U_{p,q}|+\|\boldsymbol{Q}^{(p,q)^{-1/2}}_{n,k}\boldsymbol{w}^U_{p,q}\|\leq\beta^{(p,q)}_{n,k},\forall n,k,p,q, \quad (50b)$$

$$\gamma^U_{n,k}\left(\sum_{(p,q)\neq(n,k)}\beta^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\right)+t^{U^2}_{n,k}(\nu) -2\,t^U_{n,k}(\nu)t^U_{n,k}\leq 0,\ \forall n,k, \quad (50c)$$

$$t^U_{n,k}+\|\boldsymbol{Q}^{(n,k)^{-\frac{1}{2}}}_{n,k}\boldsymbol{w}^U_{n,k}\|+|\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\boldsymbol{w}^U_{n,k}(\nu)| -2\frac{\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\hat{\boldsymbol{h}}^{(n,k)}_{n,k}\boldsymbol{w}^{U^H}_{n,k}(\nu)}{|\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\boldsymbol{w}^U_{n,k}(\nu)|}\boldsymbol{w}^U_{n,k}\leq 0,\ \forall n,k, \quad (50d)$$

$$\gamma^B_{n,k}\left(\sum_{(p,q)}\beta^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\right)+t^{B^2}_{n,k}(\nu) -2t^B_{n,k}(\nu)t^B_{n,k}\leq 0,\ \forall n,k, \quad (50e)$$

$$t^B_{n,k}+\|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\|+|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)| -2\frac{\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)|}\boldsymbol{w}^B\leq 0,\ \forall n,k. \quad (50f)$$

We now introduce Lagrangian multipliers $\boldsymbol{\lambda}\triangleq\{\lambda^{(p,q)}_{n,k}\},\boldsymbol{\mu}\triangleq\{\mu_{n,k}\},\boldsymbol{\kappa}\triangleq\{\kappa_{n,k}\},\boldsymbol{\xi}\triangleq\{\xi_{n,k}\},\boldsymbol{\rho}\triangleq\{\rho_{n,k}\}$ for the constraints in (50b)-(50f), respectively, and define $\boldsymbol{z}\triangleq$

$\left(\boldsymbol{w}^B, \{\boldsymbol{w}^U_{n,k}\}, \{\beta^{(p,q)}_{n,k}\}, \{t^B_{n,k}\}, \{t^U_{n,k}\}\right)$. Then the Lagrangian of (50) can then be obtained as

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho},\boldsymbol{z};\boldsymbol{z}(\nu)\right) &= \mathcal{L}_{\boldsymbol{w}^B}\left(\boldsymbol{\rho},\boldsymbol{w}^B;\boldsymbol{w}^B(\nu)\right)\\ &+\sum_{n,k}\mathcal{L}_{\boldsymbol{w}^U_{n,k}}\left(\boldsymbol{\lambda}_{n,k},\kappa_{n,k},\boldsymbol{w}^U_{n,k};\boldsymbol{w}^U_{n,k}(\nu)\right)\\ &+\sum_{n,k}\mathcal{L}_{\boldsymbol{\beta}_{n,k}}\left(\boldsymbol{\lambda}_{n,k},\mu_{n,k},\xi_{n,k},\boldsymbol{\beta}_{n,k}\right)\\ &+\sum_{n,k}\mathcal{L}_{t^U_{n,k}}\left(\mu_{n,k},\kappa_{n,k},t^U_{n,k};t^U_{n,k}(\nu)\right)\\ &+\sum_{n,k}\mathcal{L}_{t^B_{n,k}}\left(\xi_{n,k},\rho_{n,k},t^B_{n,k};t^B_{n,k}(\nu)\right),\end{aligned} \tag{51}$$

where

$$\mathcal{L}_{\boldsymbol{w}^B}\left(\boldsymbol{\rho},\boldsymbol{w}^B;\boldsymbol{w}^B(\nu)\right) \triangleq \|\boldsymbol{w}^B\|^2+\sum_{n,k}\rho_{n,k}\|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\boldsymbol{w}^B\| -2\sum_{n,k}\rho_{n,k}\frac{\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)|}\boldsymbol{w}^B, \tag{52a}$$

$$\mathcal{L}_{\boldsymbol{w}^U_{n,k}}\left(\boldsymbol{\lambda}^{(n,k)},\kappa_{n,k},\boldsymbol{w}^U_{n,k};\boldsymbol{w}^U_{n,k}(\nu)\right) \triangleq \|\boldsymbol{w}^U_{n,k}\|^2 +\sum_{p,q}\lambda^{(n,k)}_{p,q}\left(|\hat{\boldsymbol{h}}^{(n,k)^H}_{p,q}\boldsymbol{w}^U_{n,k}|+|\boldsymbol{Q}^{(n,k)^{-1/2}}_{p,q}\boldsymbol{w}^U_{n,k}|\right) +\kappa_{n,k}\|\boldsymbol{Q}^{(n,k)^{-\frac{1}{2}}}_{n,k}\boldsymbol{w}^U_{n,k}\|-2\kappa_{n,k}\frac{\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\hat{\boldsymbol{h}}^{(n,k)}_{n,k}\boldsymbol{w}^{U^H}_{n,k}(\nu)}{|\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\boldsymbol{w}^U_{n,k}(\nu)|}\boldsymbol{w}^U_{n,k}, \tag{52b}$$

$$\mathcal{L}_{\boldsymbol{\beta}_{n,k}}\left(\boldsymbol{\lambda}_{n,k},\mu_{n,k},\xi_{n,k},\boldsymbol{\beta}_{n,k}\right) \triangleq -\sum_{p,q}\lambda^{(p,q)}_{n,k}\beta^{(p,q)}_{n,k} +\mu_{n,k}\gamma^U_{n,k}\sum_{(p,q)\neq(n,k)}\beta^{(p,q)^2}_{n,k}+\xi_{n,k}\gamma^B_{n,k}\sum_{(p,q)}\beta^{(p,q)^2}_{n,k}, \tag{52c}$$

$$\mathcal{L}_{t^U_{n,k}}\left(\mu_{n,k},\kappa_{n,k},t^U_{n,k};t^U_{n,k}(\nu)\right) \triangleq -2\mu_{n,k}t^U_{n,k}(\nu)t^U_{n,k}+\kappa_{n,k}t^U_{n,k}, \tag{52d}$$

$$\mathcal{L}_{t^B_{n,k}}\left(\xi_{n,k},\rho_{n,k},t^B_{n,k};t^B_{n,k}(\nu)\right) \triangleq -2\xi_{n,k}t^B_{n,k}(\nu)t^B_{n,k}+\rho_{n,k}t^B_{n,k}. \tag{52e}$$

The optimization problem in (50) is strongly convex and satisfies Slater's condition, thus strong duality holds. Therefore, the optimal solution can be obtained by solving its dual problem, which is given by

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho}} \quad D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right) \tag{53a}$$

$$\text{s.t.} \quad \boldsymbol{\lambda}\geq\boldsymbol{0},\boldsymbol{\mu}\geq\boldsymbol{0},\boldsymbol{\kappa}\geq\boldsymbol{0},\boldsymbol{\xi}\geq\boldsymbol{0},\boldsymbol{\rho}\geq\boldsymbol{0}, \tag{53b}$$

where the dual function $D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right)$ is obtained by minimizing the Lagrangian over the primal variables as

$$\begin{aligned}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right) &= \min_{\boldsymbol{w}^B}\mathcal{L}_{\boldsymbol{w}^B}\left(\boldsymbol{\rho},\boldsymbol{w}^B;\boldsymbol{w}^B(\nu)\right) && \text{(54a)}\\ &+\sum_{n,k}\min_{\boldsymbol{w}^U_{n,k}}\mathcal{L}_{\boldsymbol{w}^U_{n,k}}\left(\boldsymbol{\lambda}_{n,k},\kappa_{n,k},\boldsymbol{w}^U_{n,k};\boldsymbol{w}^U_{n,k}(\nu)\right) && \text{(54b)}\\ &+\sum_{n,k}\min_{\beta^{(p,q)}_{n,k}}\mathcal{L}_{\{\beta^{(p,q)}_{n,k}\}}\left(\boldsymbol{\lambda}_{n,k},\mu_{n,k},\xi_{n,k},\beta^{(p,q)}_{n,k}\right) && \text{(54c)}\\ &+\sum_{n,k}\min_{t^U_{n,k}}\mathcal{L}_{t^U_{n,k}}\left(\mu_{n,k},\kappa_{n,k},t^U_{n,k};t^U_{n,k}(\nu)\right) && \text{(54d)}\\ &+\sum_{n,k}\min_{t^B_{n,k}}\mathcal{L}_{t^B_{n,k}}\left(\xi_{n,k},\rho_{n,k},t^B_{n,k};t^B_{n,k}(\nu)\right), && \text{(54e)}\end{aligned}$$

yielding the optimal solutions $\hat{\boldsymbol{z}}\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho}\right) = \left(\hat{\boldsymbol{w}}^B_{n,k},\{\hat{\boldsymbol{w}}^U_{n,k}\},\{\hat{\beta}^{(p,q)}_{n,k}\},\{\hat{t}^U_{n,k}\},\{\hat{t}^B_{n,k}\}\right)$. The optimization over $\boldsymbol{w}^U_{n,k},\beta^{(p,q)}_{n,k},t^U_{n,k},t^B_{n,k}$ in (54) can be decomposed into NK separable subproblems. The dual function $D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\boldsymbol{z}(\nu)\right)$ is differentiable with its gradient given by

$$\nabla_{\lambda^{n,k}_{p,q}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right) = |\hat{\boldsymbol{h}}^{(n,k)^H}_{p,q}\hat{\boldsymbol{w}}^U_{n,k}| +\|\boldsymbol{Q}^{(n,k)^{-1/2}}_{p,q}\hat{\boldsymbol{w}}^U_{n,k}\|-\hat{\beta}^{(n,k)}_{p,q},\ \forall p,q, \tag{55a}$$

$$\nabla_{\mu_{n,k}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right) = \gamma^U_{n,k}\left(\sum_{(p,q)\neq(n,k)}\hat{\beta}^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\right) +\hat{t}^{U^2}_{n,k}(\nu)-2t^U_{n,k}(\nu)\hat{t}^U_{n,k}, \tag{55b}$$

$$\nabla_{\kappa_{n,k}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right) = \hat{t}^U_{n,k}+\|\boldsymbol{Q}^{(n,k)^{-\frac{1}{2}}}_{n,k}\hat{\boldsymbol{w}}^U_{n,k}\| +|\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\boldsymbol{w}^U_{n,k}(\nu)|-2\frac{\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\hat{\boldsymbol{h}}^{(n,k)}_{n,k}\boldsymbol{w}^{U^H}_{n,k}(\nu)}{|\hat{\boldsymbol{h}}^{(n,k)^H}_{n,k}\boldsymbol{w}^U_{n,k}(\nu)|}\hat{\boldsymbol{w}}^U_{n,k}, \tag{55c}$$

$$\nabla_{\xi_{n,k}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right) = \gamma^B_{n,k}\left(\sum_{(p,q)}\hat{\beta}^{(p,q)^2}_{n,k}+\sigma^2_{n,k}\right) +t^{B^2}_{n,k}(\nu)-2t^B_{n,k}(\nu)\hat{t}^B_{n,k}, \tag{55d}$$

$$\nabla_{\rho_{n,k}}D\left(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\xi},\boldsymbol{\rho};\hat{\boldsymbol{z}}(\nu)\right) = \hat{t}^B_{n,k}+\|\boldsymbol{Q}^{-\frac{1}{2}}_{n,k}\hat{\boldsymbol{w}}^B\| +|\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{w}}^B(\nu)|-2\frac{\hat{\boldsymbol{h}}^H_{n,k}\hat{\boldsymbol{h}}_{n,k}\boldsymbol{w}^{B^H}(\nu)}{|\hat{\boldsymbol{h}}^H_{n,k}\boldsymbol{w}^B(\nu)|}\hat{\boldsymbol{w}}^B, \tag{55e}$$

all of which can be computed efficiently in a distributed manner.

Overall, the obtained algorithm is a double-loop scheme. The outer loop consists of the SCA iterations as described in Table I. In each of the SCA iteration, gradient descent based dual ascent algorithm is adopted. First, the primal variable $\boldsymbol{z}^j$ is updated by solving the optimization problems outlined in (54a)-(54e), each of which is solved by solving NK subproblems. Specifically, the update of $\boldsymbol{w}^{U^j}_{n,k}$ only requires local CSI, i.e., $\hat{\boldsymbol{h}}^{(n,k)}_{p,q}$ for $\forall p,q$, and other local information such as $\boldsymbol{\lambda}^{(n,k)}$ and $\kappa_{n,k}$. Similarly, the updates of $t^U_{n,k}$ and $t^B_{n,k}$ only require local information. On the other hand, the update of the networkwide beamforming vector $\boldsymbol{w}^{B^j}$ needs full CSI across the network, as well as gathered information $\rho_{n,k}$ from all the clusters. The update of $\boldsymbol{\beta}_{n,k}$, which measures the received interference powers at user (n,k) from BSs outside the cluster

TABLE II
DISTRIBUTED ALGORITHM WITHIN THE v-TH SCA ITERATION IN LDM

STEP 0: Set $j=1$. Initialize dual variables $\boldsymbol{\lambda}^0, \boldsymbol{\mu}^0, \boldsymbol{\kappa}^0, \boldsymbol{\xi}^0, \boldsymbol{\rho}^0$.
STEP 1: If the stopping criterion is satisfied, then STOP
STEP 2: At the central node:
solve (54a) to obtain $\boldsymbol{w}^{B^j}$
At each cluster $\mathcal{C}_{n,k}$:
update $\boldsymbol{w}_{n,k}^{U^j}, t_{n,k}^{U^j}, t_{n,k}^{B^j}$ with only local information
update $\boldsymbol{\beta}_{n,k}^j$ with $\lambda_{n,k}^{(p,q)^{j-1}}$ from $\mathcal{C}_{p,q}$ where $(p,q) \neq (n,k)$
STEP3: The central node broadcasts $\boldsymbol{w}^{B^j}$ to all the clusters
Each cluster $\mathcal{C}_{n,k}$ sends $\beta_{n,k}^{(p,q)^j}$ to $\mathcal{C}_{p,q}$
STEP 4: At each cluster $\mathcal{C}_{n,k}$:
update $\boldsymbol{\lambda}^{n,k^j}, \mu_{n,k}^j, \kappa_{n,k}^j, \xi_{n,k}^j, \rho_{n,k}^j$ according to (55a)-(55e)
Each cluster $\mathcal{C}_{n,k}$ sends $\lambda_{p,q}^{(n,k)^j}$ to $\mathcal{C}_{p,q}$
STEP 4: Set $j=j+1$, and go to STEP 1

$\mathcal{C}_{n,k}$, involves the exchange of $\{\lambda_{n,k}^{(p,q)}\}$ from all p,q. Once the primal variable is updated, dual variable updates can be executed with the gradient descent method, with gradient given in (55a)-(55e), respectively. Note that the update of dual variables can be performed locally with the message $\boldsymbol{w}^{B^j}$ from the central processing unit. The detailed algorithm description can be found in Table II. We finally remark that, while the computation of the broadcast beamforming vector is performed at a processing unit with full CSI, the proposed implementation is more efficient when compared to the centralized approach thanks to the distributed optimization of unicast transmissions. Specifically, the optimization problems in (54b)-(54e) can be solved in parallel using distributed computing resources, and each of the problems is for a single scalar variable or for a vector of dimension M or NK.

V. SIMULATION RESULTS

In this section, simulation results are presented to obtain insights into the performance comparison between LDM and TDM for the purpose of transmission of unicast and broadcast services in cellular systems. Unless stated otherwise, we consider a network comprised of macro-cells, each with $K = 3$ single-antenna active users. The radius of each cell is 500 m, and the users are located uniformly around the BS at a distance of 400 m. Each BS is equipped with $M = 3$ antennas. All channel vectors $\boldsymbol{h}_{i,n,k}$ are written as $\boldsymbol{h}_{i,n,k} = \left(10^{-\mathrm{PL}/10}\right)^{1/2} \tilde{\boldsymbol{h}}_{i,n,k}$, where the path loss exponent is modeled as $\mathrm{PL} = 148.1 + 37.6\log_{10}(d_{i,n,k})$, with $d_{i,n,k}$ denoting the distance (in kilometers) between the i-th BS and user (n,k), and $\tilde{\boldsymbol{h}}_{i,n,k}$ denoting an i.i.d. vector accounting for Rayleigh fading of unitary power. The noise variance is set to $\sigma_{n,k}^2 = -134$ dBW for all users (n,k). Unless stated otherwise, we assume non-cooperative unicast transmission, i.e., each BS is informed only about the unicast data streams of its own users.

A. Perfect CSI

Initially, we assume perfect CSI at all the BSs in the network. We plot the cumulative distribution function (CDF) of the transmission power per BS for LDM and TDM with $N = 3$ cells in Fig. 2. For the latter, we consider different values for the fraction of time T_0/T devoted to unicast traffic. Other values of T_0/T were seen not to improve the performance. The transmission power per BS is defined as the sum-power divided by the number of BSs. We observe that the curves may represent improper CDFs in the sense that their asymptotic values may be below 1. This gap accounts for the probability of the set of channel realizations in which the problem is found to be infeasible. We refer to the previous section for the assumed definition of infeasibility for SCA, whereas the standard definition is used for the convex problems in (27) and (30) solved using the S-procedure. Henceforth, we refer to the probability of an infeasible channel realization as the *outage probability*.

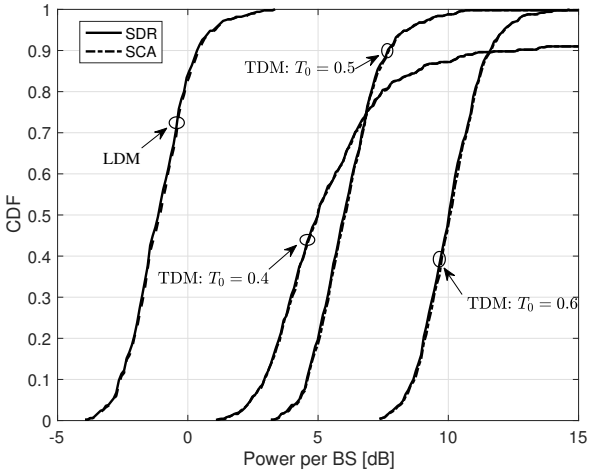

Fig. 2. The CDF of power consumption per BS with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

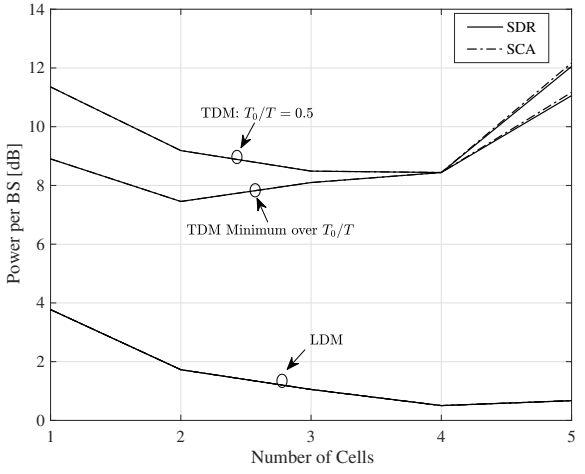

Fig. 3. Power consumption per BS as a function of the number of cells with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

We can observe from Fig. 2 that LDM enables a significant reduction in the transmission power per BS as compared with TDM. In fact, even with an optimized choice of T_0/T, LDM can improve the 95th percentile of the transmitted power

per BS by around 7 dB. Another observation is that SCA operates close to the lower bound set by SDR. Note also that LDM has a significantly lower outage probability than TDM. Finally, we remark that a large value of T_0/T is beneficial to obtain a lower outage probability in TDM, suggesting that the unicast constraints have more significant impact on the feasibility of the problem due to the need to cope with the mutual interference among unicast data streams. For the rest of this section, the displayed power values correspond to the 95th percentile of the corresponding CDF.

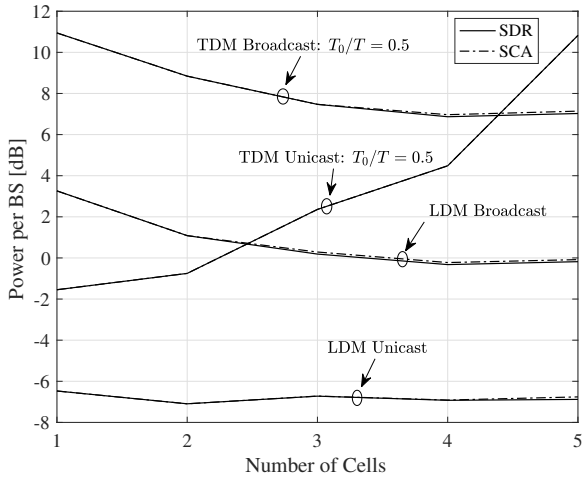

Fig. 4. Power consumption per BS as a function of the number of cells with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

Next we study the impact of the number of cells on the performance of the system. To this end, Fig. 3 and Fig. 4 show the power per BS as a function of the number of cells. Specifically, Fig. 3 shows the overall power per BS, while Fig. 4 illustrates separately the power per BS used for the broadcast and unicast layers. Note that in Fig. 4 we fixed $T_0/T = 0.5$, while in Fig. 3 we also show the power obtained by selecting, for any number of cells, the value of T_0/T that minimizes the overall sum-power consumption (obtained by a line search). A key observation from Fig. 3 is that the power saving afforded by LDM increases with the number of cells. This gain can be attributed to the following two facts: (*i*) the optimal injection level is high (see Fig. 4), and hence the broadcast layer requires more power than unicast; and (*ii*) the performance of LDM is enhanced by the presence of more cells broadcasting the same message in the SFN, which increases the broadcast SINR and the broadcast layer can be more easily canceled by the users. The latter fact can be seen from Fig. 4, in which the required unicast power decreases with the number of cells when using LDM, unlike in TDM. Furthermore, the optimal IL of TDM decreases significantly, also suggesting that TDM is more sensitive to the mutual interference introduced by unicast data streams.

Fig. 5 compares the required power per BS for non-cooperative unicast transmission and for fully cooperative unicast transmission, i.e., clusters $\mathcal{C}_{n,k} = \{1, \ldots, N\}$ for all users (n, k). Here we consider a network comprised of $N = 3$ cells, and set $T_0/T = 0.8$ for TDM. From Fig. 5, it can be concluded that a higher unicast rate entails larger power savings by means of cooperative unicast transmission, especially for TDM. It is also worth mentioning that the LDM approach without BS cooperation in unicast transmission can even outperform the fully cooperative TDM approach in certain scenarios, e.g., when the rate for unicast messages is considerably lower than the broadcast rate.

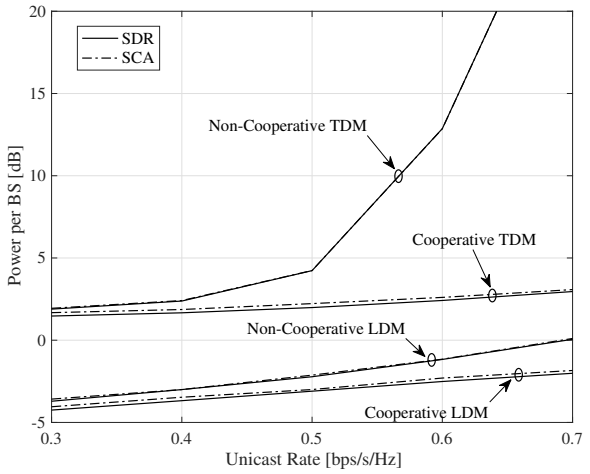

Fig. 5. Power consumption per BS, separately for the unicast and broadcast signals, for values of unicast rate with R^B=2 bps/Hz for non-cooperative and fully cooperative schemes.

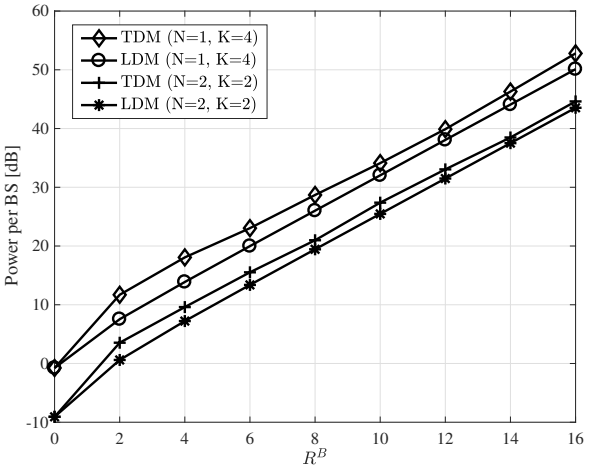

Fig. 6. Power consumption per BS as a function of the broadcast rate with R^U=0.5 bps/Hz

We present the required power per BS of LDM and TDM as a function of the broadcast rate in Fig. 6. The unicast rate is set to $R^U = 0.5$ bps/Hz for all the users. The optimal time allocation T_0 in TDM is found by a line search with step size 0.05. When only unicast transmissions exist, i.e., $R^B = 0$, both LDM and TDM problems boil down to the multigroup multicast beamforming problem, and have the same performance in terms of power consumption. When the broadcast message and unicast messages are jointly transmitted, LDM always outperforms TDM in the considered range of broadcast

rates. It is also concluded that the performance gain of LDM is larger with a higher user density.

Finally, we show the impact of the distance between users and the BS on the performance of TDM and LDM in Fig. 7. Here we consider the network consisting of $N = 5$ cells, each with a BS of $M = 5$ antennas. The scenarios with $K = 1$ and $K = 5$ users in each cell are simulated to observe the impact of user density on the performance of the system. It can be seen that LDM always outperforms TDM and has a power gain of around 5 dB in the considered range of distances. It is also observed that LDM can provide the same level of performance for cell-edge users as cell-center users in TDM.

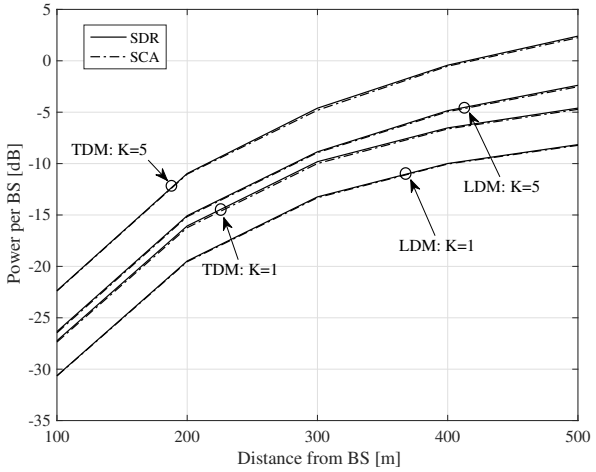

Fig. 7. Power consumption per BS as a function of the distance between users and BSs with R^B=1 bps/Hz, R^U=0.5 bps/Hz, $N = 5$, and $M = 5$.

Next, we present the performance comparison between TDM and LDM considering two practical impairments, namely, imperfect channel coding, and imperfect CSI.

B. Imperfect Channel Coding

To account for the channel coding suboptimality, the SNR gap to capacity for broadcast and unicast layers is introduced as in [16]. Then, the SINR expressions of the broadcast signal are modified as follows:

$$\text{SINR}_{n,k}^{B\text{-TDM}} = \lambda^B \frac{|\boldsymbol{h}_{n,k}^H \boldsymbol{w}^B|^2}{\sigma_{n,k}^2} \tag{56}$$

and

$$\text{SINR}_{n,k}^{B\text{-LDM}} = \lambda^B \frac{|\boldsymbol{h}_{n,k}^H \boldsymbol{w}^B|^2}{\sum\limits_{(p,q)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U|^2 + \sigma_{n,k}^2}, \tag{57}$$

as opposed to (12) and (16) for TDM and LDM, respectively, where λ^B is the SNR gap to capacity of the broadcast layer. Similarly, the SINR expressions for the unicast transmission in (13) and (17) are modified to

$$\begin{aligned}\text{SINR}_{n,k}^{U\text{-LDM}} &= \text{SINR}_{n,k}^{U\text{-TDM}} \\ &= \lambda^U \frac{|\boldsymbol{h}_{n,k}^{(n,k)^H} \boldsymbol{w}_{n,k}^U|^2}{\sum\limits_{(p,q)\neq(n,k)} |\boldsymbol{h}_{n,k}^{(p,q)^H} \boldsymbol{w}_{p,q}^U|^2 + \sigma_{n,k}^2},\end{aligned} \tag{58}$$

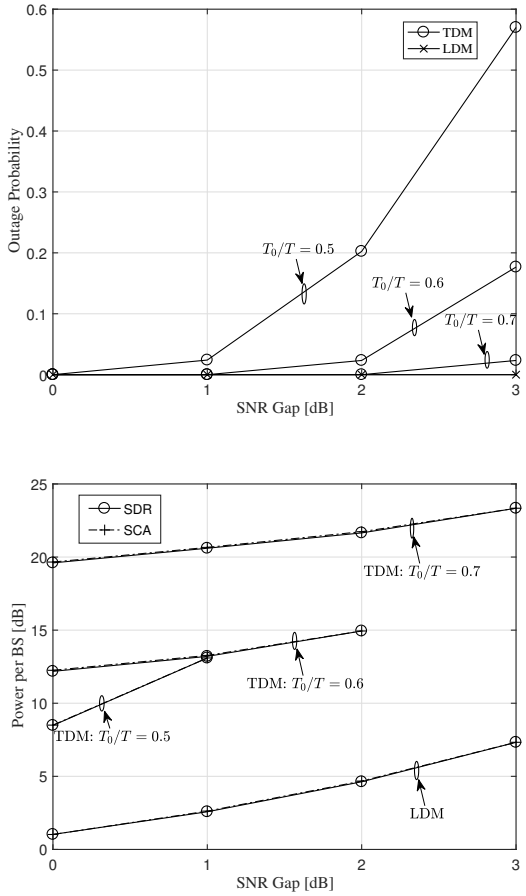

Fig. 8. Outage probability and power consumption per BS for various values of SNR gap from ideal channel coding with target rates R^B=3 bps/Hz and R^U=0.5 bps/Hz.

where λ^U is the SNR gap to capacity for the unicast layer.

The outage probability versus the SNR gap, measured in dB, is presented in Fig. 8(a), while the corresponding transmission power per BS for LDM and TDM are depicted in Fig. 8(b). It can be observed that the outage probability of TDM significantly increases with the increased SNR gap from perfect channel coding, while the outage probability of LDM remains zero in our setting. In the state-of-the-art terrestrial broadcasting system where $\lambda^U = \lambda^B = -1$ dB are considered as the realistic values for the SNR gaps of the two layers [16], although TDM provides acceptable system service availability, the power consumption is found to be about 10 dB higher than LDM, as shown in Fig. 8(b). It can be further noticed that even when the SNR gap is 3 dB in LDM, the power consumption is still lower than TDM with ideal channel coding.

C. Imperfect CSI

We then demonstrate the effect of imperfect CSI on the performance. The channel error covariance matrix is set as $\boldsymbol{Q}_{i,n,k} = 1/\epsilon^2 \boldsymbol{I}_M$, where ϵ^2 is the common CSI error variance for all $\boldsymbol{e}_{i,n,k}$'s. It is observed in Fig. 9 that the power consumption per BS increases for both TDM and LDM systems, with the increase in CSI error variance ϵ^2. It is interesting to note

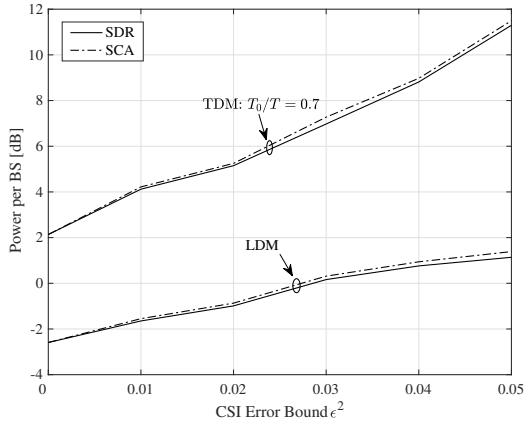

Fig. 9. Power consumption per BS as a function of CSI error bound ϵ^2 with target rates R^B=1 bps/Hz and R^U=1 bps/Hz.

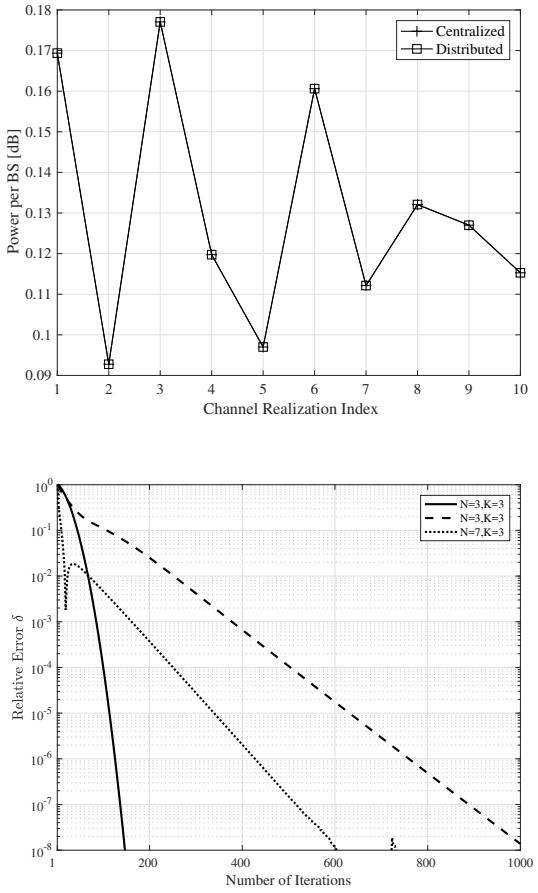

Fig. 10. Convergence of the dual decomposition-based algorithm and relative error within dual ascent iterations for a given SCA subproblem.

that the minimum required power of TDM increases faster than that of LDM, indicating that TDM is more sensitive to CSI errors compared to LDM. This effect resembles the results encountered with higher unicast rate requirement and more users. In general, LDM outperforms TDM in terms not only of power consumption, but also of robustness against flexible system QoS targets and CSI imperfections.

D. Distributed Implementation

We first demonstrate that the distributed algorithm can converge to the same optimal solution as the centralized scheme, as shown in Fig. 10(a). The centralized solution was obtained by solving the optimization problem in (49) by CVX. The performance of the proposed dual decomposition-based distributed algorithm is studied in Fig. 10(b). The relative error at the j-th iteration of the algorithm in the ν-th SCA loop is computed by $\delta = |p^j - p^*|/p^*$, where p^j denotes the dual ascent solution at iteration j, and p^* denotes the optimal solution obtained by CVX in the best precision mode. The appropriate penalty parameters ρ are found empirically to observe fast convergence. Accordingly, Fig. 10(b) shows the convergence behavior of the distributed solution as a function of the number of iterations. It can be seen that for LDM, the algorithm converges fast to achieve an acceptable relative value, say $\delta = 10^{-4}$, within 500 iterations for a $N = 7$ cell network.

VI. Conclusions

In this paper, we have analyzed the performance gain of LDM over TDM/FDM as a potential NOMA approach for simultaneous transmission of broadcast and unicast messages over cellular networks. Joint beamforming design and power allocation was formulated as a sum-power minimization problem under distinct QoS constraints for the individual unicast messages and the common broadcast message. The resulting non-convex problem has been tackled by means of SCA and S-procedure, which provide upper and lower bounds on the optimal solution, respectively. Our numerical results have shown that the upper and lower bounds are tight, which indicates the near-optimality of the proposed solutions. We have also observed that LDM significantly improves the performance as compared to orthogonal transmission, and that it provides power savings for both the unicast and broadcast transmissions thanks to the larger bandwidth available. We have seen that the benefit of the increased bandwidth available for the broadcast layer outweights the interference caused by unicast transmissions. In the case of imperfect CSI, we have noted that, while increased CSI error adversely affects both LDM and TDM, the increase in minimum required power as a function of the CSI error variance is much faster with TDM compared to LDM, indicating that LDM also provides better robustness against CSI uncertainties commonly experienced in real systems. A dual decomposition-based distributed solution has also been presented, which facilitates efficient distributed implementation for the LDM technique.

References

[1] J. Zhao, O. Simeone, D. Gunduz, and D. Gómez-Barquero, "Non-orthogonal unicast and broadcast transmission via joint beamforming and ldm in cellular networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.

[2] F. Hartung, U. Horn, J. Huschke, M. Kampmann, T. Lohmar, and M. Lundevall, "Delivery of broadcast services in 3G networks," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 188–199, Mar. 2007.

[3] Qualcomm, "LTE broadcast," https://www.qualcomm.com/documents/lte-broadcast-white-paper-idc, Sep. 2014.

[4] J. F. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gómez-Barquero, "Joint delivery of unicast and E-MBMS services in LTE networks," *IEEE Trans. Broadcast.*, vol. 58, no. 2, pp. 157–167, Jun. 2012.

[5] G. K. Walker, J. Wang, C. Lo, X. Zhang, and G. Bao, "Relationship between LTE broadcast/eMBMS and next generation broadcast television," *IEEE Trans. Broadcast.*, vol. 60, no. 2, pp. 185–192, Jun. 2014.

[6] L. Shi, E. Obregon, K. W. Sung, J. Zander, and J. Bostrom, "CellTV–on the benefit of TV distribution over cellular networks: A case study," *IEEE Trans. Broadcast.*, vol. 60, no. 1, pp. 73–84, Mar. 2014.

[7] L. Shi, K. W. Sung, and J. Zander, "Future TV content delivery over cellular networks from urban to rural environments," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6177–6187, Nov. 2015.

[8] 3GPP TR 22.816 V2.0.0, "3GPP enhancement for TV service," http://www.3gpp.org/DynaReport/22816.htm, Dec. 2015.

[9] 3GPP, "Enhanced television services over 3GPP eMBMS," http://www.3gpp.org/news-events/3gpp-news/1905-embms_r14, Oct. 2017.

[10] D. Gomez-Barquero, D. Navratil, S. Appleby, and M. Stagg, "Point-to-multipoint communication enablers for the fifth generation of wireless systems," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 53–59, Mar 2018.

[11] 3GPP SP-190253, "New SID: Architectural enhancements for 5G multicast-broadcast services," https://portal.3gpp.org/ngppapp/CreateTDoc.aspx?mode=view&contributionUid=SP-190253, Mar. 2019.

[12] D. Kim, F. Khan, C. V. Rensburg, Z. Pi, and S. Yoon, "Superposition of broadcast and unicast in wireless cellular systems," *IEEE Commun. Mag.*, vol. 46, no. 7, pp. 110–117, Jul. 2008.

[13] L. Fay, L. Michael, D. Gómez-Barquero, N. Ammar, and M. W. Caldwell, "An overview of the ATSC 3.0 physical layer specification," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 159–171, Mar. 2016.

[14] S. I. Park *et al.*, "Low complexity layered division multiplexing for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 233–243, Mar. 2016.

[15] 3GPP TR 36.776, "Evolved Universal Terrestrial radio access (E-UTRA); study on LTE-based 5G terrestrial broadcast," hhttps://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3500, Mar. 2019.

[16] D. Gómez-Barquero and O. Simeone, "LDM versus FDM/TDM for unequal error protection in terrestrial broadcasting systems: An information-theoretic view," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 571–579, Dec. 2015.

[17] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[19] 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.

[20] O. Tervo, H. Pennanen, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Distributed optimization for coordinated beamforming in multicell multigroup multicast systems: Power minimization and SINR balancing," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 171–185, Jan. 2018.

[21] E. G. Larsson and H. V. Poor, "Joint beamforming and broadcasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 3058–3070, Apr. 2016.

[22] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *CoRR*, vol. abs/1808.08325, 2018. [Online]. Available: http://arxiv.org/abs/1808.08325

[23] E. Chen, M. Tao, and Y. Liu, "Joint base station clustering and beamforming for non-orthogonal multicast and unicast transmission with backhaul constraints," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6265–6279, Sep. 2018.

[24] A. Tajer, N. Prasad, and X. Wang, "Robust linear precoder design for multi-cell downlink transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 235–251, Jan. 2011.

[25] G. Zheng, K.-K. Wong, and T.-S. Ng, "Robust linear MIMO in the downlink: A worst-case optimization with ellipsoidal uncertainty regions," *EURASIP J. Adv. Signal Process*, vol. 2008, no. 1, p. 609028, Jul. 2008.

[26] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inf. Theory*, vol. 9, no. 23, pp. 113–381, 2013.

[27] C. Shen, T. Chang, K. Wang, Z. Qiu, and C. Chi, "Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, Jun. 2012.

[28] A. Tolli, H. Pennanen, and P. Komulainen, "Decentralized minimum power multi-cell beamforming with limited backhaul signaling," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 570–580, Feb 2011.

[29] E. Chen and M. Tao, "ADMM-based fast algorithm for multi-group multicast beamforming in large-scale wireless systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2685–2698, Jun. 2017.

[30] Z. Luo, W. Ma, A. M. 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.

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