X. Zhang, L. Xiang, J. Wang, and X. Gao, "Massive MIMO Multicasting with Finite Blocklength," accepted for publication in *IEEE Transactions on Wireless Communications*, Jul. 2024.

# Massive MIMO Multicasting with Finite Blocklength

Xuzhong Zhang, *Graduate Student Member, IEEE*, Lin Xiang, *Member, IEEE*, Jiaheng Wang, *Senior Member, IEEE*, and Xiqi Gao, *Fellow, IEEE*

***Abstract*—Massive multiple-input multiple-output (MIMO) multicasting is a promising approach for simultaneously delivering common messages to multiple users in next-generation wireless networks. However, existing studies have exclusively focused on multicast beamforming designs based on the Shannon capacity, assuming the infinite blocklength (IBL) for transmission. This assumption may lead to strictly suboptimal designs for practical multicast transmissions with finite blocklength (FBL), especially in ultra-reliable low-latency communications. In this paper, we explore the beamforming design for massive MIMO multi-group multicasting in the FBL regime. Our study considers both the max-min fairness and the weighted sum rate criteria for a comprehensive treatment. Due to the non-concave FBL rate function, the resulting optimization problems are known to be notoriously hard. We characterize the necessary and sufficient condition for the non-negative FBL rate to be a concave function of the received signal-to-interference-plus-noise ratio (SINR). Considering a finite number of transmit antennas, we propose low-complexity majorization-minimization (MM) type algorithms, which update variables in either closed or semi-closed form, to achieve locally optimal solutions of the formulated optimization problems. We further show that, as the number of transmit antennas becomes large, the optimal beamformer of each group aligns asymptotically with a linear combination of the channel vectors of that group of users, where the optimal normalized combining coefficients are derived in closed form. Subsequently, we obtain the globally optimal multicast beamformers by optimizing the power allocation using low-complexity iterative algorithms. Simulation results show that the proposed schemes outperform several existing methods, especially those employing the Shannon capacity as the performance metric. Moreover, the proposed algorithms exhibit complexities that only slightly grow with the number of transmit antennas and they can notably reduce the computation time by up to two orders of magnitude over the benchmarks, making them highly beneficial for massive MIMO applications.**

***Index Terms*—Finite blocklength transmission, massive multiple-input multiple-output (MIMO), multicast beamforming, max-min fairness, weighted sum rate.**

This work was supported in part by the National Key R&D Program of China under Grant 2021YFB2900303, the National Natural Science Foundation of China under Grants 62331024 and U22B2006, the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grants BK20212001, BK20222001, and BK20192002, the Key Technologies R&D Program of Jiangsu (Prospective and Key Technologies for Industry) under Grants BE2022067-5 and BE2022068-3, the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant BK20233002, the Fundamental Research Funds for the Central Universities under Grants 2242023K5003, 2242022K60002 and 2242022K60001. The work of L. Xiang has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center under grant LOEWE/1/12/519/03/05.001(0016)/72 and the BMBF project Open6GHub under grant 16KISKO14. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ozlem Demir. *(Corresponding author: Jiaheng Wang; Xiqi Gao.)*

Xuzhong Zhang, Jiaheng Wang, and Xiqi Gao are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 210023, China. Jiaheng Wang is also with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: xzzhang@seu.edu.cn, jhwang@seu.edu.cn, and xqgao@seu.edu.cn).

Lin Xiang is with the Communications Engineering Lab, Technische Universität Darmstadt, 64283 Darmstadt, Germany (e-mail: l.xiang@nt.tu-darmstadt.de).

## I. Introduction

Massive multiple-input multiple-output (MIMO) multicasting is an advanced physical-layer transmission technique for delivery of common messages in the next-generation wireless networks. By exploiting the broadcast nature of the wireless medium, multicasting enables simultaneous signal transmission to multiple receivers over shared frequency spectrum [1]. As such, multicasting provides a more resource-efficient and scalable solution than conventional unicasting for communicating common messages to a large number of receivers such as in massive machine-type communications (mMTC) [2]–[7]. Meanwhile, unlike broadcasting, multicasting receivers can be flexibly grouped to enable multi-group multicasting communications. Note that unicasting and broadcasting are only special cases of multi-group multicasting. In this regard, multi-group multicasting can always achieve the best performance compared with unicasting and broadcasting. Furthermore, through employing a large antenna array at the base station (BS), massive MIMO can acquire abundant multiplexing and diversity gains to significantly improve the reliability and spectral efficiency of multicasting [8]. However, beamforming design for (massive) MIMO multicasting proves to be more intricate than for conventional unicasting applications and has defined a key research challenge.

To maximize the performance of (massive) MIMO multicasting, multicast beamforming design has been recently investigated in [7], [9]–[20]. For example, the authors in [7], [9]–[14] studied either (i) transmit power minimization subject to the users' signal-to-interference-plus-noise ratio (SINR) requirements, or (ii) maximization of the minimum users' SINR subject to a total power budget. Considering unlimited transmit antennas at the BS, the optimal solution of problem (i) for massive MIMO multicasting was obtained in [14]. In [16], pilot contamination was further considered for maximization of the minimum multicast rate. Additionally, the authors of [17], [18] investigated the multicast beamforming under the weighted sum rate (WSR) criteria using a heuristic approach. And a mixed WSR multicast beamforming with a common message is studied in [19] using the weighted minimum mean square error (WMMSE) method [21].

The aforementioned works [7], [9]–[20] have exclusively focused on designing (massive) MIMO multicast beamformers based on the Shannon capacity, by assuming the *infinite* blocklength (IBL) for transmission. Nevertheless, only *finite* blocklengths (FBLs) can be afforded in practical multicasting systems. Thus, these existing multicast beamforming designs

are bound to be suboptimal, except for some special cases such as when the FBL is sufficiently long that the performance gap between FBL and IBL transmission becomes negligible. Recently, multicasting has been increasingly applied in latency-sensitive communications for real-time dissemination of control commands, alerts, and updating messages in Internet-of-Things (IoT), mMTC, and digital twin [2]–[6]. These FBL applications[1] typically operate with short transmission blocklengths, in order to lower the communication latency. Compared with unicasting, employing multicasting in these FBL applications creates spectrum sharing opportunities among multiple receivers to mitigate the resource limitation inherent in the FBL regime. However, the associated multicasting designs deviate significantly from the infinite or close-to-infinite blocklength regimes. Therefore, it is imperative to rethink the optimal beamforming design for (massive) MIMO multicasting in the FBL regime, which motivates our work in this paper.

Unlike IBL transmission, FBL transmission is no longer error-free and cannot achieve the Shannon capacity [25]. Instead, the achievable rate in the FBL regime is highly non-concave. When extending the massive MIMO multicasting to support FBL communications (FBLC), the resulting beamforming optimization problem is notoriously hard, hindering the solution that is both optimal in the FBL regime and scalable to a large number of transmit antennas at the BS. To the best of our knowledge, results in this direction have not been reported in the literature yet, except that some recent works on FBL beamforming have investigated the unicasting scenarios [26]–[31]. To bridge the knowledge gap, in this paper, we investigate low-complexity beamforming designs for massive MIMO multicasting in the FBL regime, considering both the max-min fair (MMF) optimization and the WSR maximization. Our contributions are summarized as follows:

- We comprehensively characterize the properties of the FBL rate function and reveal the necessary and sufficient condition for the non-negative FBL rate to be a concave function of the received SINR.
- Considering a finite number of transmit antennas at the BS, we propose low-complexity majorization-minimization (MM)-type algorithms to find locally optimal solutions of the MMF and the WSR problems, where the variables are updated in closed or semi-closed form.
- Considering unlimited transmit antennas at the BS, we show that the asymptotically optimal beamformer of each group is a linear combination of users' channel vectors in the group and further derive the optimal normalized combining coefficients in closed form. This result enables us to obtain the globally optimal multicast beamformers by optimizing the power allocations using low-complexity iterative algorithms.
- Simulation results demonstrate that the proposed design outperforms several benchmark schemes, especially those employing the Shannon capacity as the performance metric, highlighting the importance of adopting the FBL rate for multicast beamforming design. Additionally, the proposed algorithms can reduce the computation time by orders of magnitude compared to the benchmarks, which is appealing for massive MIMO systems.

[1]Practical FBL applications may exhibit distinct requirements on latency, reliability, and data rate [6], [22]–[24]. For example, according to [6, Table I], a general automation process requires 99.99% reliability (or a block error rate (BLER) of $10^{-4}$) within a latency of 50–100 ms, whereas self-driving car applications require 99% reliability within a latency of only 1 ms. On the other hand, immersive virtual reliability services require a data rate for transmitting vision information from 10 Mbits/s to 1 Gbits/s with 99.9%-99.999% reliability [22], [23].

The remainder of this paper is organized as follows. In Section II, we introduce the system model and formulate the MMF and WSR multicast beamforming problems in the FBL regime. In Section III, we analyze the concavity of the FBL rate. In Sections IV and V, we propose efficient beamforming designs for the MMF and WSR problems for a finite and an unlimited number of transmit antennas at the BS, respectively. Simulation results are presented in Section VI, and finally, conclusions are drawn in Section VII.

*Notations:* Throughout this paper, vectors and matrices are denoted in bold lower-case and capital letters, respectively. $\mathbb{R}$, $\mathbb{C}$, $\mathbb{C}^{N\times 1}$, and $\mathbb{C}^{N\times M}$ denote the sets of real numbers, complex numbers, complex vectors of length $N$, and complex matrices of size $N\times M$, respectively. $\Re\{x\}$ denotes the real part of complex number $x$. $[\mathbf{A}]_{i,j}$ denotes the $(i,j)$-th entry of matrix $\mathbf{A}$. $\mathbf{I}_N$ denotes the $N\times N$ identity matrix. $(\cdot)^T$, $(\cdot)^\star$, $(\cdot)^H$, and $(\cdot)^{-1}$ denote transpose, complex conjugate, Hermitian transpose, and inverse of matrix, respectively. $|\cdot|$ and $\|\cdot\|_2$ denote the absolute value of a complex scalar and the Euclidean norm of a vector, respectively. $\mathbb{E}(\cdot)$ is the expectation operator. $Q(x)$ is the Q-function defined as $Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-t^2/2)dt$. $Q^{-1}(\cdot)$ is the inverse Q-function, i.e., $Q(Q^{-1}(x))=x$. $\mathbf{x}\sim\mathcal{CN}(\mathbf{a},\mathbf{R})$ means that $\mathbf{x}$ is a circular symmetric complex Gaussian random vector with mean $\mathbf{a}$ and covariance $\mathbf{R}$. Finally, $\mathbb{I}_{\mathcal{S}}(\mathbf{X})$ is the indicator function for variable $\mathbf{X}$ and set $\mathcal{S}$, defined as

$$\mathbb{I}_{\mathcal{S}}(\mathbf{X})=\begin{cases}0, & \text{if } \mathbf{X}\in\mathcal{S},\\ \infty, & \text{otherwise.}\end{cases} \tag{1}$$

## II. System Model And Problem Formulation

We consider the downlink (DL) of a single-cell multi-group multicasting system, where a BS equipped with $N_t$ transmit antennas serves $K$ single-antenna users over a bandwidth of $B$ Hz. The users, indexed by set $\mathcal{K}=\{1,\ldots,K\}$, are grouped into $M$ multicast groups, indexed by $\mathcal{M}=\{1,\ldots,M\}$. Group $m$ consists of $K_m$ users that are indexed by set $\mathcal{K}_m$ and receive a common data stream. Each user is assigned to only one group and we use $I(k)$ to denote the group index of user $k$. Let $\mathbf{h}_k\in\mathbb{C}^{N_t\times 1}$ be the channel vector between the BS and user $k$ and $\mathbf{w}_m\in\mathbb{C}^{N_t\times 1}$ be the beamforming vector for group $m$. The received signal of user $k$ is given by

$$y_k=\mathbf{h}_k^H\mathbf{w}_{I(k)}s_{I(k)}+\sum\nolimits_{m\neq I(k)}\mathbf{h}_k^H\mathbf{w}_m s_m+n_k, \tag{2}$$

where $s_m$ is the data symbol intended for the users in group $m$ with $\mathbb{E}(|s_m|^2)=1$ and $n_k\sim\mathcal{CN}(0,\sigma_k^2)$ is the additive white Gaussian noise (AWGN) at user $k$. Defining

$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$ as the beamforming matrix, the received SINR of user $k$ is expressed as

$$\gamma_k(\mathbf{W}) = \frac{|\mathbf{h}_k^H \mathbf{w}_{I(k)}|^2}{\sum_{m \neq I(k)} |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_k^2}. \quad (3)$$

To satisfy the transmission latency requirement $T_{\max}$ for FBL applications, the signals are transmitted using the FBL $N = BT_{\max}$[2]. Since for many FBL applications $T_{\max}$ is usually much smaller than the channel coherence time, the channel can be regarded as quasi-static over the packet transmission duration [6], [26]–[30]. By adopting the FBL transmission, the users cannot decode the messages in an error-free manner. Following [34], we define $\epsilon_m$ as the BLER of group $m$, where each user in group $m$ can decode the multicast message with an error probability not exceeding $\epsilon_m$. The resulting achievable rate of user $k$ with FBL to achieve the BLER $\epsilon_{I(k)}$ in nats/sec/Hz is given by[3] [25]

$$R(\gamma_k(\mathbf{W}), \vartheta_{I(k)}) = \ln(1+\gamma_k(\mathbf{W})) - \vartheta_{I(k)} \sqrt{V(\gamma_k(\mathbf{W}))}, \quad (4)$$

where $\vartheta_{I(k)} = Q^{-1}(\epsilon_{I(k)})/\sqrt{N} > 0$ and $V(\gamma_k(\mathbf{W})) = 1 - (1+\gamma_k(\mathbf{W}))^{-2}$ is the channel dispersion. In (4), the second term adds a penalty on the achievable rate in order to guarantee transmission reliability, i.e., satisfying the required BLER. The penalty term vanishes as $N \to \infty$, where the FBL rate (4) approaches the Shannon capacity $\ln(1+\gamma_k(\mathbf{W}))$. To ensure reliable transmission to all users within a group, the multicasting rate of group $m$ is determined as the minimum rate of the users in that group and is given by

$$R_m^G = \min_{k \in \mathcal{K}_m} R(\gamma_k(\mathbf{W}), \vartheta_m). \quad (5)$$

To satisfy the QoS requirements of FBL applications on reliability, latency, and data rate, we consider that at least $D_m$ nats of data should be transmitted to the users in group $m$ within the required BLER $\epsilon_m$ and blocklength $N$ (or transmission latency $T_{\max} = N/B$). This requires $R_m^G \geq \bar{R}_m^G \triangleq D_m/N, \ \forall m \in \mathcal{M}$, where $\bar{R}_m^G > 0$ is the minimum rate required by the users in group $m$. We assume that the users are in relatively low mobility such that the BS has perfect knowledge of the channel state information (CSI) [11], [13], [14], [26]–[31], [38].[4] Meanwhile, we consider both the MMF and the WSR criteria for fair and efficient multicast beamforming design, respectively. Both criteria have been considered in the literature for multicast beamforming design with the IBL [7], [10]–[12], [16]–[18]. Here, we aim to extend them to the FBL regime. The resulting beamforming optimization problems are formulated as:

$$\mathbb{P}_1: \max_{\mathbf{W}} \ \min_{m \in \mathcal{M}} R_m^G \quad (6a)$$

$$\text{s.t.} \ \sum\nolimits_{m=1}^{M} \|\mathbf{w}_m\|_2^2 \leq P, \quad (6b)$$

$$R\left(\gamma_k(\mathbf{W}), \vartheta_{I(k)}\right) \geq \bar{R}_{I(k)}^G, \ \forall k \in \mathcal{K}, \quad (6c)$$

and

$$\mathbb{P}_2: \max_{\mathbf{W}} \ \sum\nolimits_{m=1}^{M} \omega_m R_m^G \quad (7)$$

$$\text{s.t.} \ (6b), (6c),$$

respectively, where $P$ is the maximum transmit power of the BS, and $\omega_m > 0$ is the weight assigned for group $m$. In problems $\mathbb{P}_1$ and $\mathbb{P}_2$, we maximize the minimum multicast rate among the groups and the weighted sum of multicast rates, respectively, for given BLER $\{\epsilon_m\}_{m=1}^M$ and blocklength $N$, while fulfilling the rate requirements in (6c).

Note that while we focus on multicasting in this paper, the considered multi-group multicasting scheme is quite *general* to also encompass conventional IBL unicasting [21], broadcasting [10], and multicasting [7], [9]–[19], as well as FBL unicasting [26]–[29], for which our system model, problem formulation and solutions remain applicable. Particularly, it reduces to unicasting and broadcasting when $K_m = 1$, $\forall m$ and $M = 1$, respectively.

As the FBL rate (4) is a non-concave function of $\mathbf{W}$ (or $\gamma_k(\mathbf{W})$) for finite $N$, problems $\mathbb{P}_1$ and $\mathbb{P}_2$ involve non-smooth non-concave objective functions and non-convex constraints (6c), which render their solutions challenging. In the following, we start with analyzing the concavity of the FBL rate with respect to (w.r.t.) $\gamma_k(\mathbf{W})$ in Section III. Considering finite $N_t$ in Section IV, we then propose low-complexity iterative algorithms with variables updated in closed or semi-closed form to obtain the locally optimal solutions of $\mathbb{P}_1$ and $\mathbb{P}_2$. We further study the globally optimal solutions for $N_t \to \infty$ in Section V.

## III. Concavity of the Achievable Rate with Finite Blocklength

In this section, we delve into the FBL rate function

$$R(\gamma, \vartheta) = \ln(1+\gamma) - \vartheta\sqrt{V(\gamma)}, \ \gamma \geq 0, \quad (8)$$

particularly its concavity w.r.t. the SINR $\gamma$ for given BLER $\epsilon$, blocklength $N$, and channel dispersion $V(\gamma) = 1-(1+\gamma)^{-2}$, where $\vartheta = Q^{-1}(\epsilon)/\sqrt{N} > 0$. Such analysis is nontrivial for finite $N$ and plays an important role in the FBLC research

[2]Since we focus on beamforming design in the physical layer, we only consider the DL transmission latency, rather than the overall end-to-end (E2E) latency. The latter includes uplink and DL transmission latency, coding and processing latency, queueing delay, and routing delay in backhaul and core networks etc, which needs to be evaluated with the cross-layer design and optimization approach, while considering the intermittent and random characteristics of FBL communications [32], [33]. However, due to the limited space, this is left for the future work. Instead, we only assume in this work that the DL transmission latency is specified according to the E2E latency requirement of the considered FBL applications.

[3]In this work, we focus on FBL applications with mild requirements on latency or reliability, and the normal approximation [25] is adopted in (4) to strike a good balance between approximation accuracy and complexity. According to the recent literature, the random-coding union (RCU) bound [25], [35], the saddlepoint approximation [36], and the Laplace method [37] can be adopted to further improve the approximation accuracy in the FBL regime. However, these bounds are overly complex, making the performance analysis and optimization intractable. Beamforming optimization using these approximations is an interesting open problem for future research.

[4]Under the perfect CSI assumption, we aim to obtain a performance upper bound for massive MIMO multicasting in the FBL regime. In time-division duplex (TDD) systems, the BS can obtain the DL CSI via uplink training process by exploiting the channel reciprocity. Meanwhile, the users can obtain the DL CSI using the DL pilots as in [39]. Our previous works [20], [40]–[42] have studied several robust precoding designs with imperfect CSI for massive MIMO, assuming the IBL. We would like to further study the FBL beamforming design with imperfect CSI in the future work.

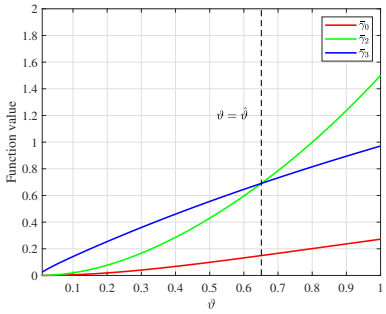


Figure 1. Illustration of $\bar{\gamma}_0$, $\bar{\gamma}_2$, $\bar{\gamma}_3$ versus $\vartheta$.

[26]. We first review in Lemmas 1 and 2 some properties of $R(\gamma, \vartheta)$ as given in [26]. Based on these, we present an important finding in Theorem 1 about the concavity of $R(\gamma, \vartheta)$ w.r.t. $\gamma$.

**Lemma 1** ( [26]). *Given $\vartheta > 0$, $R(\gamma, \vartheta)$ monotonically decreases (increases) with $\gamma$ for $0 \leq \gamma < \bar{\gamma}_0$ ($\gamma \geq \bar{\gamma}_0$), where*

$$\bar{\gamma}_0 = \sqrt{\frac{1+\sqrt{1+4\vartheta^2}}{2}} - 1. \tag{9}$$

*Moreover, $R(\bar{\gamma}_0, \vartheta) \leq 0$ for any $\vartheta > 0$, and $R(\gamma, \vartheta)$ has two different roots $\bar{\gamma}_1 = 0$ and $\bar{\gamma}_2 > 0$ such that $R(\bar{\gamma}_i, \vartheta) = 0$, $i = 1, 2$, where $\bar{\gamma}_1 \leq \bar{\gamma}_0 \leq \bar{\gamma}_2$.*

Based on Lemma 1, there exists a unique solution $\hat{\gamma}_{I(k)} > 0$ satisfying $R(\hat{\gamma}_{I(k)}, \vartheta_{I(k)}) = \bar{R}^G_{I(k)}$. Thus (6c) can be rewritten as

$$\gamma_k(\mathbf{W}) \geq \hat{\gamma}_{I(k)}, \forall k \in \mathcal{K}. \tag{10}$$

Then problems $\mathbb{P}_1$ and $\mathbb{P}_2$ can be equivalently reformulated as

$$\mathbb{P}_3 : \max_{\mathbf{W}} \min_{m \in \mathcal{M}} R^G_m \quad \text{s.t. (6b), (10),} \tag{11}$$

and

$$\mathbb{P}_4 : \max_{\mathbf{W}} \sum\nolimits_{m=1}^{M} \omega_m R^G_m \quad \text{s.t. (6b), (10).} \tag{12}$$

**Lemma 2** ( [26]). *Given $\vartheta > 0$, there exists an inflection point $\bar{\gamma}_3 \geq \bar{\gamma}_0$ such that $R(\gamma, \vartheta)$ is convex for $0 \leq \gamma \leq \bar{\gamma}_3$ and concave for $\gamma > \bar{\gamma}_3$.*

Although there are no closed-form expressions for $\bar{\gamma}_2$ and $\bar{\gamma}_3$ in the literature, they can be calculated using numerical approaches, e.g., bisection method. Lemma 1 reveals that for given $\vartheta > 0$, $\bar{\gamma}_2$ defines a cut-off SINR, below which no communication takes place, as $R(\gamma, \vartheta) \leq 0$ for $\gamma \leq \bar{\gamma}_2$. This suggests that one only needs to evaluate the concavity of the non-negative $R(\gamma, \vartheta)$ in the effective SINR regime $\gamma \geq \bar{\gamma}_2$. Moreover, according to Lemma 2, for given $\vartheta > 0$, $R(\gamma, \vartheta)$ is concave w.r.t. $\gamma$ in the effective SINR regime $\gamma \geq \bar{\gamma}_2$ if and only if $\bar{\gamma}_2 \geq \bar{\gamma}_3$. However, due to the lack of analytical expressions for $\bar{\gamma}_2$ and $\bar{\gamma}_3$, checking the concavity of $R(\gamma, \vartheta)$ remains a challenging task. To resolve this issue, we provide the following result.

**Theorem 1.** *$R(\gamma, \vartheta)$ is a concave function of $\gamma$ in the effective SINR regime $\gamma \geq \bar{\gamma}_2$ if and only if $\vartheta \geq \hat{\vartheta}$, where $\hat{\vartheta} \approx 0.651\,12$.*

*Proof.* Please refer to Appendix A. □

Note that for FBL systems under a reliability requirement of $1-\epsilon = 99.999\%$, we have $\vartheta \geq \hat{\vartheta}$ when $N \leq 42$, which can be easily satisfied even in URLLC. Fig. 1 plots the numerical values of $\bar{\gamma}_0$, $\bar{\gamma}_2$, and $\bar{\gamma}_3$ versus $\vartheta$, where the dashed line indicates $\vartheta = \hat{\vartheta}$. We see that $\bar{\gamma}_2 \geq \bar{\gamma}_3$ if and only if $\vartheta \geq \hat{\vartheta}$, whereby $R(\gamma, \vartheta)$ is concave in the effective SINR regime. This result is consistent with Theorem 1.

*Remark* 1. To show the importance of Theorem 1 for resource allocation in the FBL regime, let us consider a DL unicasting using $L$ orthogonal AWGN channels with unit noise power, blocklength $N_i$, and BLER $\epsilon_i > 0, i = 1, \ldots, L$. The spectral efficiency maximization problem can be formulated as

$$\max_{\{p_i\}_{i=1}^L} \sum\nolimits_{i=1}^{L} R(h_i p_i, Q^{-1}(\epsilon_i)/\sqrt{N_i}) \quad \text{s.t.} \sum\nolimits_{i=1}^{L} p_i = P', \ h_i p_i \geq \bar{\gamma}_{2,i}, \ i = 1, \ldots, L, \tag{13}$$

where $P' > \sum_{i=1}^{L} \bar{\gamma}_{2,i}/h_i$ is the total transmit power, $\bar{\gamma}_{2,i} = \bar{\gamma}_2(Q^{-1}(\epsilon_i)/\sqrt{N_i}) > 0$ is given in Lemma 1 such that $R(\bar{\gamma}_{2,i}, Q^{-1}(\epsilon_i)/\sqrt{N_i}) = 0, i = 1, \ldots, L$, $h_l > 0$ and $p_l \geq 0$ are the channel gain and transmit signal power of channel $l$, respectively. Problem (13) is generally non-convex due to the non-concave objective function. However, Theorem 1 indicates that problem (13) becomes convex if and only if $Q^{-1}(\epsilon_i)/\sqrt{N_i} \geq \hat{\vartheta}, i = 1, \ldots, L$, whereby its globally optimally solution can be obtained by solving the Karush-Kuhn-Tucker (KKT) conditions. In Section V, we will further use Theorem 1 and Lemma 2 to obtain the globally optimal solution of the WSR problem $\mathbb{P}_4$ as $N_t \to \infty$, cf. Theorem 7.

## IV. Multicast Beamforming Design for Finite Number of Transmit Antennas

In this section, we propose low-complexity algorithms to solve the FBL multicast beamforming optimization problems $\mathbb{P}_3$ and $\mathbb{P}_4$ for finite number of transmit antennas at the BS. We use the MM method to transform the problems $\mathbb{P}_3$ and $\mathbb{P}_4$ into convex forms, which are further solved with efficient decompositions.

### A. MMF Multicast Beamforming Design

To facilitate a tractable solution for the non-convex problem $\mathbb{P}_3$, we first approximate $R\left(\gamma_k(\mathbf{W}), \vartheta_{I(k)}\right)$ by a concave function of $\mathbf{W}$.

**Theorem 2.** *For any feasible beamforming matrix $\mathbf{W}^{(t)}$ satisfying constraints* (6b) *and* (10)*, a lower bound of $R\left(\gamma_k(\mathbf{W}), \vartheta_{I(k)}\right)$ is given by:*

$$R\left(\gamma_k(\mathbf{W}), \vartheta_{I(k)}\right) \geq \underline{R}_k(\mathbf{W}, \mathbf{W}^{(t)}) \triangleq a_k^{(t)} + \sum\nolimits_{m=1}^{M} \Re\left\{(\mathbf{b}_{k,m}^{(t)})^H \mathbf{w}_m\right\} - c_k^{(t)} |\mathbf{w}_m^H \mathbf{h}_k|^2, \tag{14}$$

*where $\underline{R}_k(\mathbf{W}, \mathbf{W}^{(t)})$ is a concave function satisfying*

$$\left.\frac{\partial \underline{R}_k(\mathbf{W}, \mathbf{W}^{(t)})}{\partial \mathbf{w}_m}\right|_{\mathbf{w}_m = \mathbf{w}_m^{(t)}} = \left.\frac{\partial R(\gamma_k(\mathbf{W}), \vartheta_{I(k)})}{\partial \mathbf{w}_m}\right|_{\mathbf{w}_m = \mathbf{w}_m^{(t)}}, \tag{15}$$

$$a_k^{(t)} = \ln\left(1+\gamma_k(\mathbf{W}^{(t)})\right) - \gamma_k(\mathbf{W}^{(t)}) - \sigma_k^2\left(\frac{1}{\alpha_k(\mathbf{W}^{(t)})} - \frac{1}{\beta_k(\mathbf{W}^{(t)})}\right) - \frac{\vartheta_{I(k)}\sqrt{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}}{2}\left(1+\frac{1}{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}\right)$$
$$- \frac{\vartheta_{I(k)}}{2\sqrt{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}}\left(\frac{\alpha_k(\mathbf{W}^{(t)})}{\beta_k(\mathbf{W}^{(t)})}\right)^2 + \frac{\vartheta_{I(k)}\sigma_k^2\alpha_k(\mathbf{W}^{(t)})}{\beta_k(\mathbf{W}^{(t)})\sqrt{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}}\left(\frac{2}{\beta_k(\mathbf{W}^{(t)})} - \frac{\alpha_k(\mathbf{W}^{(t)})}{\beta_k^2(\mathbf{W}^{(t)})}\right). \quad (16)$$

*and* $R\left(\gamma_k(\mathbf{W}^{(t)}), \vartheta_{I(k)}\right) = \underline{R}_k(\mathbf{W}^{(t)}, \mathbf{W}^{(t)})$. *In* (14)*,* $a_k^{(t)}$ *is a constant whose value depends on* $\mathbf{W}^{(t)}$ *according to* (16) *at the top of this page,*

$$\mathbf{b}_{k,m}^{(t)} = \begin{cases} \frac{2\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_m^{(t)}}{\alpha_k(\mathbf{W}^{(t)})}, & \text{if } m = I(k), \\ \frac{2\vartheta_{I(k)}\alpha_k(\mathbf{W}^{(t)})\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_m^{(t)}}{\beta_k^2(\mathbf{W}^{(t)})\sqrt{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}}, & \text{otherwise}, \end{cases} \quad (17)$$

$$c_k^{(t)} = \frac{1}{\alpha_k(\mathbf{W}^{(t)})} - \frac{1}{\beta_k(\mathbf{W}^{(t)})} + \frac{\vartheta_{I(k)}}{\sqrt{V_k\left(\gamma_k(\mathbf{W}^{(t)})\right)}}\frac{\alpha_k^2(\mathbf{W}^{(t)})}{\beta_k^3(\mathbf{W}^{(t)})} > 0, \quad (18)$$

$$\alpha_k(\mathbf{W}) = \sum\nolimits_{m\in\mathcal{M}, m\neq I(k)} |\mathbf{w}_m^H\mathbf{h}_k|^2 + \sigma_k^2, \quad (19)$$

$$\beta_k(\mathbf{W}) = \alpha_k(\mathbf{W}) + |\mathbf{w}_{I(k)}^H\mathbf{h}_k|^2. \quad (20)$$

*Proof.* Please refer to Appendix B. □

Moreover, due to the convexity of $x^2$, we have $x^2 \geq 2x_0x - x_0^2$ for any $x_0 > 0$. Then, given any feasible solution $\mathbf{W}^{(t)}$, we approximate the non-convex constraints (10) as

$$\hat{\gamma}_{I(k)}\left(\sum\nolimits_{m\in\mathcal{M}, m\neq I(k)} |\mathbf{w}_m^H\mathbf{h}_k|^2 + \sigma_k^2\right) + \left|\mathbf{h}_k^H\mathbf{w}_{I(k)}^{(t)}\right|^2 - 2\Re\left\{(\mathbf{w}_{I(k)}^{(t)})^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_{I(k)}\right\} \leq 0, \forall k \in \mathcal{K}. \quad (21)$$

Employing (14) and (21), problem $\mathbb{P}_3$ can be rewritten as

$$\mathbb{P}_5: \max_{\mathbf{W}, r}\ r \quad (22a)$$
$$\text{s.t. (6b), (21)} \quad (22b)$$
$$r \leq \underline{R}_k(\mathbf{W}, \mathbf{W}^{(t)}), \forall k \in \mathcal{K}, \quad (22c)$$

where $r$ is an auxiliary variable. Problem $\mathbb{P}_5$ is convex and can be solved using standard approaches such as the interior-point method. However, the complexity of the interior-point method is still high for massive MIMO systems.

To facilitate efficient multicast beamforming designs, in the following, we develop a low-complexity block coordinate descent (BCD) algorithm to solve problem $\mathbb{P}_5$, where each block of variables can be updated in closed or semi-closed form. To this end, we first introduce the following auxiliary variables

$$\mathbf{v}_m = \mathbf{w}_m,\ \forall m \in \mathcal{M}, \quad (23)$$
$$q_{k,m} = \mathbf{h}_k^H\mathbf{w}_m,\ \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \quad (24)$$
$$\Gamma_{k,m} = \mathbf{h}_k^H\mathbf{w}_m,\ \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \quad (25)$$
$$d_k = r,\ \forall k \in \mathcal{K}, \quad (26)$$

where we define $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$, $[\mathbf{Q}]_{k,m} = q_{k,m}$, $[\mathbf{\Gamma}]_{k,m} = \Gamma_{k,m}$, and $\mathbf{d} = [d_1, \ldots, d_K]^T$. Letting

$$\delta_{k,m}^{(t)} \triangleq \frac{(\mathbf{b}_{k,m}^{(t)})^H\mathbf{h}_k}{\|\mathbf{h}_k\|^2},\ \forall k \in \mathcal{K}, \forall m \in \mathcal{M}. \quad (27)$$

$$\Theta_k(\mathbf{\Gamma}_k) \triangleq \hat{\gamma}_{I(k)}\left(\sum_{m\neq I(k)} |\Gamma_{k,m}|^2 + \sigma_k^2\right) + |\mathbf{h}_k^H\mathbf{w}_{I(k)}^{(t)}|^2 - 2\Re\left\{(\mathbf{w}_{I(k)}^{(t)})^H\mathbf{h}_k\Gamma_{k,I(k)}\right\}, \forall k \in \mathcal{K}, \quad (28)$$

$$\Upsilon_k(d_k, \mathbf{q}_k) \triangleq \sum\nolimits_{m=1}^M c_k^{(t)}|q_{k,m}|^2 - a_k^{(t)} + d_k - \sum\nolimits_{m=1}^M \Re\{\delta_{k,m}^{(t)}q_{k,m}\}, \forall k \in \mathcal{K}, \quad (29)$$

where $\mathbf{\Gamma}_k = [\Gamma_{k,1}, \ldots, \Gamma_{k,M}]^T$, and $\mathbf{q}_k = [q_{k,1}, \ldots, q_{k,M}]^T$, problem $\mathbb{P}_5$ can be equivalently reformulated as

$$\mathbb{P}_6: \max_{\mathbf{W}, r, \mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}}\ r \quad (30a)$$
$$\text{s.t.}\ \sum\nolimits_{m=1}^M \|\mathbf{v}_m\|_2^2 \leq P, \quad (30b)$$
$$\Theta_k(\mathbf{\Gamma}_k) \leq 0, \forall k \in \mathcal{K}, \quad (30c)$$
$$\Upsilon_k(d_k, \mathbf{q}_k) \leq 0, \forall k \in \mathcal{K}, \quad (30d)$$
$$(23), (24), (25), (26),$$

Using the indicator function $\mathbb{I}_{(\cdot)}(\cdot)$, an equivalent reformulation of $\mathbb{P}_6$ is further obtained as

$$\mathbb{P}_7: \max_{\mathbf{W}, r, \mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}}\ r + \mathbb{I}_{\mathcal{C}}(\mathbf{V}) + \mathbb{I}_{\mathcal{D}}(\mathbf{\Gamma}) + \mathbb{I}_{\mathcal{E}}(\mathbf{Q}, \mathbf{d}) \quad (31)$$
$$\text{s.t. (23), (24), (25), (26),}$$

where sets $\mathcal{C}$, $\mathcal{D}$, and $\mathcal{E}$ are defined by (30b), (30c), and (30d), respectively. Note that $\mathbb{P}_7$ is a convex problem, whose optimization variables can be split into two blocks, $\{\mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}\}$ and $\{\mathbf{W}, r\}$, and optimized alternatively using the BCD approach. The augmented Lagrangian of $\mathbb{P}_7$ is given by

$$\begin{aligned} &\mathcal{L}_{\mathrm{M}}(\mathbf{W}, r, \mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}) \\ &= r + \mathbb{I}_{\mathcal{C}}(\mathbf{V}) + \mathbb{I}_{\mathcal{D}}(\mathbf{\Gamma}) + \mathbb{I}_{\mathcal{E}}(\mathbf{Q}, \mathbf{d}) \\ &- \frac{\rho}{2}\sum\nolimits_{k=1}^K\sum\nolimits_{m=1}^M \left|\Gamma_{k,m} - \mathbf{h}_k^H\mathbf{w}_m + \Lambda_{k,m}\right|^2 \\ &- \frac{\rho}{2}\left(\sum\nolimits_{k=1}^K |d_k - r + u_k|^2 + \sum\nolimits_{m=1}^M \|\mathbf{v}_m - \mathbf{w}_m + \mathbf{z}_m\|_2^2\right) \\ &- \frac{\rho}{2}\sum\nolimits_{k=1}^K\sum\nolimits_{m=1}^M \left|q_{k,m} - \mathbf{h}_k^H\mathbf{w}_m + \Psi_{k,m}\right|^2, \end{aligned} \quad (32)$$

where $\rho > 0$ is the penalty parameter, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_M]$, $[\mathbf{\Psi}]_{k,m} = \Psi_{k,m}$, $[\mathbf{\Lambda}]_{k,m} = \Lambda_{k,m}$, and $\mathbf{u} = [u_1, \ldots, u_K]^T$ are the scaled dual variables for constraints (23), (24), (25), and (26), respectively. Our proposed algorithm for solving $\mathbb{P}_5$ is described in Algorithm 1, which alternates between updating the two blocks of the primal variables, $\{\mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}\}$ and

$\{\mathbf{W}, r\}$, and the dual variables, $\{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \mathbf{u}, \mathbf{Z}\}$, till convergence. In Algorithm 1, $\epsilon^{\text{pri,M}}$ and $\epsilon^{\text{dual,M}}$ are the primal and dual residuals defined similar to [43]. Interestingly, the primal variables can be updated in closed or semi-closed form in steps 3 and 4, as elaborated in the following. On the other hand, the dual variables are updated using the subgradient method in step 5.

*1) Update of* $\{\mathbf{V}, \boldsymbol{\Gamma}, \mathbf{Q}, \mathbf{d}\}$*:* Let $n$ be the iteration index. The update at each iteration can be further decomposed into three independent subproblems of $\mathbf{V}$, $\boldsymbol{\Gamma}$, and $\{\mathbf{Q}, \mathbf{d}\}$, respectively, which can be solved in parallel.

*Updating* $\mathbf{V}$*:* The update of $\mathbf{V}$ at iteration $n$ can be expressed as

$$\min_{\mathbf{V}} \ \sum\nolimits_{m=1}^{M} \left\| \mathbf{v}_m - \mathbf{w}_m^{(n)} + \mathbf{z}_m^{(n)} \right\|_2^2 \quad \text{s.t. (30b)}, \tag{33}$$

whose optimal solution is given by [44]

$$\mathbf{v}_m^* = \min\left\{1, \sqrt{\frac{P}{\sum_{m=1}^{M} \left\| \mathbf{w}_m^{(n)} - \mathbf{z}_m^{(n)} \right\|_2^2}}\right\} * (\mathbf{w}_m^{(n)} - \mathbf{z}_m^{(n)}). \tag{34}$$

*Updating* $\boldsymbol{\Gamma}$*:* The update of $\boldsymbol{\Gamma}$ at iteration $n$ further divides into $K$ independent subproblems and the subproblem with index $k$ is expressed as

$$\min_{\boldsymbol{\Gamma}_k} \ \sum\nolimits_{m=1}^{M} \left| \Gamma_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Lambda_{k,m}^{(n)} \right|^2 \tag{35a}$$

$$\text{s.t. } \Theta_k(\boldsymbol{\Gamma}_k) \le 0. \tag{35b}$$

Problem (35) is convex and its optimal solution is characterized as follows.

**Theorem 3.** *The optimal solution of problem (35) is given by*

$$\Gamma_{k,m}^* = \begin{cases} \mathbf{h}_k^H \mathbf{w}_m^{(n)} - \Lambda_{k,m}^{(n)} + \mu^* \mathbf{h}_k^H \mathbf{w}_m^{(t)}, & \text{if } m = I(k), \\ \frac{\mathbf{h}_k^H \mathbf{w}_m^{(n)} - \Lambda_{k,m}^{(n)}}{1 + \mu^* \hat{\gamma}_{I(k)}}, & \text{otherwise}, \end{cases} \tag{36}$$

*where* $\mu^* = 0$ *if* $\psi(0) < 0$*,* $\psi(\mu^*) = \Theta_k(\boldsymbol{\Gamma}_k^*)$*; otherwise,* $\mu^* > 0$ *is the unique root of equation* $\psi(\mu^*) = 0$*.*

*Proof.* Please refer to Appendix C. □

*Updating* $\{\mathbf{Q}, \mathbf{d}\}$*:* The update of $\{\mathbf{Q}, \mathbf{d}\}$ at iteration $n$ also divides into $K$ independent subproblems and the subproblem with index $k$ is expressed as

$$\min_{\mathbf{q}_k, d_k} \ \left| d_k - r^{(n)} + u_k^{(n)} \right|^2 + \sum\nolimits_{m=1}^{M} \left| q_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Psi_{k,m}^{(n)} \right|^2 \tag{37a}$$

$$\text{s.t. } \Upsilon_k(d_k, \mathbf{q}_k) \le 0. \tag{37b}$$

Problem (37) is also convex and its optimal solution is given as follows.

**Theorem 4.** *The optimal solution of problem (37) is given by*

$$d_k^* = r^{(n)} - u_k^{(n)} - \varrho^*/2, \tag{38a}$$

$$q_{k,m}^* = \frac{2\mathbf{h}_k^H \mathbf{w}_m^{(n)} - 2\Psi_{k,m}^{(n)} + \varrho^* \left(\delta_{k,m}^{(t)}\right)^\star}{2 + 2\varrho^* c_k^{(t)}}, \tag{38b}$$

**Algorithm 1** Proposed algorithm for solving $\mathbb{P}_5$

1: Initialize $n = 0$, $\mathbf{W}^{(n)}$, $r^{(n)}$, $\boldsymbol{\Lambda}^{(n)}, \boldsymbol{\Psi}^{(n)}, \mathbf{u}^{(n)}, \mathbf{Z}^{(n)}$, set $\rho$, $\varepsilon_1$ and $N_1^{\max}$;
2: **repeat**
3: Update $\{\mathbf{V}^{(n+1)}, \boldsymbol{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}\}$:
$$\{\mathbf{V}^{(n+1)}, \boldsymbol{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}\} = \arg\max_{\mathbf{V},\mathbf{Q},\boldsymbol{\Gamma},\mathbf{d}} \mathcal{L}_{\text{M}}(\mathbf{W}^{(n)}, r^{(n)}, \mathbf{V}, \boldsymbol{\Gamma}, \mathbf{Q}, \mathbf{d})$$
4: Update $\{\mathbf{W}^{(n+1)}, r^{(n+1)}\}$:
$$\{\mathbf{W}^{(n+1)}, r^{(n+1)}\} = \arg\max_{\mathbf{W},r} \mathcal{L}_{\text{M}}(\mathbf{W}, r, \mathbf{V}^{(n+1)}, \boldsymbol{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}),$$
5: Update $\{\boldsymbol{\Lambda}^{(n+1)}, \boldsymbol{\Psi}^{(n+1)}, \mathbf{u}^{(n+1)}, \mathbf{Z}^{(n+1)}\}$:
$$\begin{aligned} \Lambda_{k,m}^{(n+1)} &= \Lambda_{k,m}^{(n)} + \Gamma_{k,m}^{(n+1)} - \mathbf{h}_k^H \mathbf{w}_m^{(n+1)}, \\ \Psi_{k,m}^{(n+1)} &= \Psi_{k,m}^{(n)} + q_{k,m}^{(n+1)} - \mathbf{h}_k^H \mathbf{w}_m^{(n+1)}, \\ u_k^{(n+1)} &= u_k^{(n)} + d_k^{(n+1)} - r^{(n+1)}, \\ \mathbf{z}_m^{(n+1)} &= \mathbf{z}_m^{(n)} + \mathbf{v}_m^{(n+1)} - \mathbf{w}_m^{(n+1)}, \end{aligned}$$
6: $n = n + 1$;
7: **until** $\max\{\varepsilon^{\text{pri,M}}, \varepsilon^{\text{dual,M}}\} \le \varepsilon_1$ or $n \ge N_1^{\max}$.

*where* $\varrho^* = 0$ *if* $\phi(0) < 0$*,* $\phi(\varrho^*) = \Upsilon_k(d_k^*, \mathbf{q}_k^*)$*; otherwise,* $\varrho^* > 0$ *is the unique root of* $\phi(\varrho^*) = 0$*.*

*Proof.* Please refer to Appendix D. □

*2) Update of* $\{\mathbf{W}, r\}$*:* The update of $\{\mathbf{W}, r\}$ at iteration $n$ involves solving an unconstrained optimization problem

$$\begin{aligned} \min_{\mathbf{W},r} \ & \frac{\rho}{2} \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{M} \left| \Gamma_{k,m}^{(n+1)} - \mathbf{h}_k^H \mathbf{w}_m + \Lambda_{k,m}^{(n)} \right|^2 \\ & + \frac{\rho}{2} \sum\nolimits_{m=1}^{M} \left\| \mathbf{v}_m^{(n+1)} - \mathbf{w}_m + \mathbf{z}_m^{(n)} \right\|_2^2 \\ & + \frac{\rho}{2} \sum\nolimits_{k=1}^{K} \sum\nolimits_{m=1}^{M} \left| q_{k,m}^{(n+1)} - \mathbf{h}_k^H \mathbf{w}_m + \Psi_{k,m}^{(n)} \right|^2 \\ & - r + \frac{\rho}{2} \sum\nolimits_{k=1}^{K} \left| d_k^{(n+1)} - r + u_k^{(n)} \right|^2. \end{aligned} \tag{39}$$

By setting the first-order derivative of the objective function to zero, the optimal solution of (39) is given by

$$\begin{aligned} \mathbf{w}_m^* = & \left( \mathbf{I}_{N_t} + 2 \sum\nolimits_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^H \right)^{-1} \left\{ \mathbf{v}_m^{(n+1)} + \mathbf{z}_m^{(n)} \right. \\ & \left. + \sum\nolimits_{k=1}^{K} \mathbf{h}_k \left( \Gamma_{k,m}^{(n+1)} + \Lambda_{k,m}^{(n)} + q_{k,m}^{(n+1)} + \Psi_{k,m}^{(n)} \right) \right\}, \end{aligned} \tag{40a}$$

$$r^* = \frac{1 + \rho \sum_{k=1}^{K} \left( d_k^{(n+1)} + u_k^{(n)} \right)}{K\rho}. \tag{40b}$$

*3) Overall Solution of Problem* $\mathbb{P}_3$*:* The proposed algorithm for solving $\mathbb{P}_3$ is described in Algorithm 2. According to [43], Algorithm 1 is guaranteed to converge to the optimal solution of the convex problem $\mathbb{P}_5$. Moreover, Algorithm 2 is an MM-type solution [45], which is guaranteed to converge to a locally optimal solution of $\mathbb{P}_3$.

**Algorithm 2** Proposed algorithm for solving $\mathbb{P}_3$

1: Initialize $\mathbf{W}^{(0)}$, $r^{(0)} = \min_{k\in\mathcal{K}}\left\{R(\gamma_k(\mathbf{W}^{(0)}),\vartheta_{I(k)})\right\}$, $t=0$, set $\varepsilon_2$, and $T_1^{\max}$;
2: **repeat**
3: Update $(\mathbf{W}^{(t+1)}, r^{(t+1)})$ by solving $\mathbb{P}_5$ with Algorithm 1;
4: $t = t+1$;
5: **until** $\left|r^{(t)} - r^{(t-1)}\right| \le \varepsilon_2$ or $t \ge T_1^{\max}$.

### *B. WSR Multicasting Beamforming Design*

Employing the approximations in (14) and (21), and introducing auxiliary variables $\mathbf{r} = [r_1, \ldots, r_M]^T$, we can rewrite problem $\mathbb{P}_4$ in the following convex form

$$\mathbb{P}_8: \max_{\mathbf{W},\mathbf{r}} \ F(\mathbf{r}) \triangleq \sum\nolimits_{m=1}^{M} \omega_m r_m \tag{41a}$$

$$\text{s.t. } (6b), (21), \tag{41b}$$

$$r_{I(k)} \le \underline{R}_k(\mathbf{W}, \mathbf{W}^{(t)}), \forall k \in \mathcal{K}. \tag{41c}$$

where $\mathbf{W}^{(t)}$ is a given feasible beamforming solution. Problem $\mathbb{P}_8$ can be solved using the same method as $\mathbb{P}_5$, for which we only provide a sketch of the derivations below.

Particularly, by introducing auxiliary variables as in (23), (24), (25) and additionally

$$d_k = r_{I(k)}, \ \forall k \in \mathcal{K}, \tag{42}$$

problem $\mathbb{P}_8$ can be equivalently reformulated as

$$\mathbb{P}_9: \max_{\mathbf{W},\mathbf{r},\mathbf{V},\mathbf{\Gamma},\mathbf{Q},\mathbf{d}} \ F(\mathbf{r}) + \mathbb{I}_{\mathcal{C}}(\mathbf{V}) + \mathbb{I}_{\mathcal{D}}(\mathbf{\Gamma}) + \mathbb{I}_{\mathcal{E}}(\mathbf{Q},\mathbf{d}) \tag{43}$$

$$\text{s.t.} \quad (23), (24), (25), (42).$$

Similar to $\mathbb{P}_7$, the variables of $\mathbb{P}_9$ can be split into two blocks, $\{\mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}\}$ and $\{\mathbf{W}, \mathbf{r}\}$, which are then optimized using the BCD approach. The proposed solutions for $\mathbb{P}_8$ and $\mathbb{P}_4$ are summarized in Algorithms 3 and 4, respectively. In Algorithm 3, the augmented Lagrangian of $\mathbb{P}_9$ is given by

$$\begin{aligned}
&\mathcal{L}_{\mathrm{W}}(\mathbf{W},\mathbf{r},\mathbf{V},\mathbf{\Gamma},\mathbf{Q},\mathbf{d}) \\
&= F(\mathbf{r}) + \mathbb{I}_{\mathcal{C}}(\mathbf{V}) + \mathbb{I}_{\mathcal{D}}(\mathbf{\Gamma}) + \mathbb{I}_{\mathcal{E}}(\mathbf{Q},\mathbf{d}) \\
&- \frac{\rho}{2}\sum\nolimits_{k=1}^{K}\sum\nolimits_{m=1}^{M}\left|\Gamma_{k,m} - \mathbf{h}_k^H\mathbf{w}_m + \mathit{\Lambda}_{k,m}\right|^2 \\
&- \frac{\rho}{2}\left(\sum_{k=1}^{K}\left|d_k - r_{I(k)} + u_k\right|^2 + \sum_{m=1}^{M}\left\|\mathbf{v}_m - \mathbf{w}_m + \mathbf{z}_m\right\|_2^2\right) \\
&- \frac{\rho}{2}\sum\nolimits_{k=1}^{K}\sum\nolimits_{m=1}^{M}\left|q_{k,m} - \mathbf{h}_k^H\mathbf{w}_m + \Psi_{k,m}\right|^2.
\end{aligned} \tag{44}$$

Based on (44), $\mathbf{W}$, $\mathbf{\Gamma}$, $\mathbf{V}$ can be updated at iteration $n$ in the same manner as in the MMF problem. Meanwhile, $\mathbf{Q}$ and $\mathbf{d}$ can be updated similarly by replacing $r^{(n)}$ in (37) with $r_{I(k)}^{(n)}$. The details are omitted here for saving space. On the other hand, $\mathbf{r}$ can be updated in iteration $n$ as

$$\max_{\mathbf{r}} \ F(\mathbf{r}) - \frac{\rho}{2}\sum\nolimits_{k=1}^{K}\left|d_k^{(n+1)} - r_{I(k)} + u_k^{(n)}\right|^2, \tag{45}$$

whose optimal solution is given by

$$r_m^* = \frac{\omega_m + \rho\sum_{k\in\mathcal{K}_{\mathrm{m}}}\left(d_k^{(n+1)} + u_k^{(n)}\right)}{K_m\rho}, \ m \in \mathcal{M}. \tag{46}$$

**Algorithm 3** Proposed algorithm for solving $\mathbb{P}_8$

1: Initialize $n=0$, $\mathbf{W}^{(n)}$, $\mathbf{r}^{(n)}$, $\mathit{\Lambda}^{(n)}$, $\mathbf{\Psi}^{(n)}$, $\mathbf{u}^{(n)}$, $\mathbf{Z}^{(n)}$, set $\rho$, $\varepsilon_3$, and $N_2^{\max}$;
2: **repeat**
3: Update $\{\mathbf{V}^{(n+1)}, \mathbf{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}\}$:

$$\begin{aligned}
&\{\mathbf{V}^{(n+1)}, \mathbf{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}\} \\
&= \arg\max_{\mathbf{V},\mathbf{\Gamma},\mathbf{Q},\mathbf{d}} \mathcal{L}_{\mathrm{W}}(\mathbf{W}^{(n)}, \mathbf{r}^{(n)}, \mathbf{V}, \mathbf{\Gamma}, \mathbf{Q}, \mathbf{d}),
\end{aligned}$$

4: Update $\{\mathbf{W}^{(n+1)}, \mathbf{r}^{(n+1)}\}$:

$$\begin{aligned}
&\{\mathbf{W}^{(n+1)}, \mathbf{r}^{(n+1)}\} \\
&= \arg\max_{\mathbf{W},\mathbf{r}} \mathcal{L}_{\mathrm{W}}(\mathbf{W}, \mathbf{r}, \mathbf{V}^{(n+1)}, \mathbf{\Gamma}^{(n+1)}, \mathbf{Q}^{(n+1)}, \mathbf{d}^{(n+1)}),
\end{aligned}$$

5: Update $\{\mathit{\Lambda}^{(n+1)}, \mathbf{\Psi}^{(n+1)}, \mathbf{u}^{(n+1)}, \mathbf{Z}^{(n+1)}\}$

$$\begin{aligned}
\mathit{\Lambda}_{k,m}^{(n+1)} &= \mathit{\Lambda}_{k,m}^{(n)} + \Gamma_{k,m}^{(n+1)} - \mathbf{h}_k^H\mathbf{w}_m^{(n+1)}, \\
\Psi_{k,m}^{(n+1)} &= \Psi_{k,m}^{(n)} + q_{k,m}^{(n+1)} - \mathbf{h}_k^H\mathbf{w}_m^{(n+1)}, \\
u_k^{(n+1)} &= u_k^{(n)} + d_k^{(n+1)} - r_{I(k)}^{(n+1)}, \\
\mathbf{z}_m^{(n+1)} &= \mathbf{z}_m^{(n)} + \mathbf{v}_m^{(n+1)} - \mathbf{w}_m^{(n+1)},
\end{aligned}$$

6: $n = n+1$;
7: **until** $\max\{\varepsilon^{\text{pri,W}}, \varepsilon^{\text{dual,W}}\} \le \varepsilon_3$ or $n \ge N_2^{\max}$.

**Algorithm 4** Proposed algorithm for solving $\mathbb{P}_4$

1: Initialize $\mathbf{W}^{(0)}$, $r_m^{(0)} = \min_{k\in\mathcal{K}_m}\left\{R(\gamma_k(\mathbf{W}^{(0)}),\vartheta_m)\right\}$, $\forall m \in \mathcal{M}$, $t=0$, set $\varepsilon_4$ and $T_2^{\max}$;
2: **repeat**
3: Update $(\mathbf{W}^{(t+1)}, \mathbf{r}^{(t+1)})$ by solving $\mathbb{P}_8$ with Algorithm 3;
4: $t = t+1$;
5: **until** $\left|F(\mathbf{r}^{(t)}) - F(\mathbf{r}^{(t-1)})\right| \le \varepsilon_4$ or $t \ge T_2^{\max}$.

Algorithm 3 is guaranteed to converge to the optimal solution of $\mathbb{P}_8$ [43]. Following [45], the MM-type Algorithm 4 is guaranteed to converge to a locally optimal solution of $\mathbb{P}_4$.

*Remark* 2. Although problems $\mathbb{P}_3$ and $\mathbb{P}_4$ are non-convex, they can be efficiently solved using our proposed low-complexity Algorithms 2 and 4, respectively, with the variables parallelly updated in closed or semi-closed form. Note that although the matrix inversion step in updating the beamforming matrix (40a) is computationally intensive, it is only incurred when the channels change. Therefore, the overall computation time of our proposed multicast beamforming design remains low, as evidenced by the results in Tables I and II of Section VI. Due to the significant reduction in the processing delay, our proposed low-complexity multicast beamforming designs for massive MIMO further help lower the E2E latency, which is desired for practical systems.

## V. Asymptotic Multicast Beamforming Design

In this section, we study the multicast beamforming design in the FBL regime when the BS has a large number of transmit antennas, i.e., $N_t \to \infty$. Throughout this section,

we consider the independent Rayleigh fading channel[5] $\mathbf{h}_k = \sqrt{\Gamma_k}\widetilde{\mathbf{h}}_k$, $k \in \mathcal{K}$, where $\Gamma_k \in \mathbb{R}$ and $\widetilde{\mathbf{h}}_k \sim \mathcal{CN}(\mathbf{0}_{N_t}, \mathbf{I}_{N_t})$ denote the large-scale channel attenuation and the small-scale fading coefficients, respectively. Moreover, similar to [14], we consider that the BS transmit power $P$ is inversely proportional to $N_t$, i.e., satisfying $P = E/N_t$ for a constant power $E$. Our derivation is based on the following lemma about random vectors.

**Lemma 3** ( [49]). *Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{L\times 1}$ be mutually independent random vectors whose elements are independent and identically distributed zero-mean random variables with variances $\sigma_x^2$ and $\sigma_y^2$, respectively. The law of large numbers yields*

$$\lim_{L\to\infty} \frac{\mathbf{x}^H\mathbf{y}}{L} = 0, \text{ and } \lim_{L\to\infty} \frac{\mathbf{x}^H\mathbf{x}}{L} = \sigma_x^2. \tag{47}$$

Lemma 3 implies that the channel vectors of different users become pairwisely orthogonal as $N_t \to \infty$. Subsequently, one may conjecture that, as $N_t \to \infty$, the optimal beamformer of group $m$ lies in the subspace spanned by the channel vectors $\mathbf{h}_k, \forall k \in \mathcal{K}_m$, for maximizing the users' received SINR. This intuition is manifested by the following result.

**Lemma 4.** *When $N_t \to \infty$, the optimal beamforming solutions for $\mathbb{P}_3$ and $\mathbb{P}_4$ have the following structure*[6]*,*

$$\mathbf{w}_m^* = \sum\nolimits_{k\in\mathcal{K}_m} \xi_k \mathbf{h}_k, \ \forall m \in \mathcal{M}, \tag{48}$$

*where $\{\xi_k\}_{k=1}^K$ are the complex combining coefficients.*

*Proof.* Please refer to Appendix E. □

According to Lemma 4, to facilitate the optimal multicast beamforming design for large $N_t$, it remains to find the optimal combining coefficients $\xi_k$. However, directly solving the problems $\mathbb{P}_3$ and $\mathbb{P}_4$ for large $N_t$ by substituting (48) is still challenging due to their non-smooth and non-concave objective functions. To tackle this obstacle, we let $p_m = \|\mathbf{w}_m^*\|_2^2/P$ and $\kappa_k = \xi_k\sqrt{\Gamma_k N_t/p_m}$ be the normalized power of group $m$ and the normalized combining coefficients, respectively. Then, (6b) reduces to $\sum_{m\in M} p_m \le 1$. Substituting $\xi_k$ in (48) by $p_m$ and $\kappa_k$, we have

$$\mathbf{w}_m^* = \sum\nolimits_{k\in\mathcal{K}_m} \sqrt{Pp_m/N_t}\kappa_k\widetilde{\mathbf{h}}_k, \ \forall m \in \mathcal{M}, \tag{49}$$

and as $N_t \to \infty$,

$$\frac{\|\mathbf{w}_m^*\|_2^2}{p_m P} = \sum_{i,j\in\mathcal{K}_m} \frac{|\kappa_i^\star\kappa_j|\,\widetilde{\mathbf{h}}_i^H\widetilde{\mathbf{h}}_j}{N_t} \overset{(a)}{=} \sum_{k\in\mathcal{K}_m} |\kappa_k|^2 = 1. \tag{50}$$

[5]The independent Rayleigh fading channel is an accurate and tractable channel model for communication scenarios with rich scattering [8], [46], [47]. For this reason, it has been widely considered in the massive MIMO literature [8], [11], [14], [16], [29], [47], [48] to analyze or derive the optimal beamforming solutions for unicasting with IBL/FBL or multicasting with IBL. This channel model is also considered in our paper for insights into the optimal beamforming structure for massive MIMO multicasting in the FBL regime. Extending the FBL multicast beamforming design to general correlated fading channels is an interesting topic left for the future work.

[6]With a slight abuse of notation in (48), we don't distinguish the combining coefficients nor the beamforming vectors for $\mathbb{P}_3$ and $\mathbb{P}_4$. Because we will show in Sec. V-A that $\mathbb{P}_3$ and $\mathbb{P}_4$ have the same optimal normalized combining coefficients.

Here, $(a)$ follows from Lemma 3. Moreover, the asymptotic SINR of user $k$ is given by

$$\lim_{N_t\to\infty} \gamma_k = \lim_{N_t\to\infty} \frac{p_{I(k)}E\left|\sum\limits_{i\in\mathcal{K}_{I(k)}} \sqrt{\Gamma_k}\kappa_i^\star\widetilde{\mathbf{h}}_i^H\widetilde{\mathbf{h}}_k/N_t\right|^2}{\sum\limits_{m\neq I(k)} p_m E\left|\sum\limits_{j\in\mathcal{K}_{\mathrm{m}}} \sqrt{\Gamma_k}\kappa_j^\star\widetilde{\mathbf{h}}_j^H\widetilde{\mathbf{h}}_k/N_t\right|^2 + \sigma_k^2} \overset{(b)}{=} g_k|\kappa_k|^2, \tag{51}$$

where $g_k = \Gamma_k p_{I(k)}E/\sigma_k^2$ and $(b)$ follows from Lemma 3.

*Remark* 3. In (51), the interference from inter-group users vanishes as their fading channel vectors are orthogonal to $\widetilde{\mathbf{h}}_k$. Meanwhile, thanks to the law of large numbers, the effect of small-scale fading disappears such that the users' SINR depends only on the large-scale attenuation. Hence, massive MIMO can enhance the reliability for communications over multiuser fading channels.

Based on (51), the optimal multicast beamformers for large $N_t$ are determined by both the power allocation $\{p_m\}_{m=1}^M$ and the normalized combining coefficients $\{\kappa_k\}_{k=1}^K$. We decompose the optimal solution into two steps. First, given any feasible power allocation, we optimize the combing coefficients. In Section V-A, we show that the resulting optimal combining coefficients $\{\kappa_k^*\}_{k=1}^K$ of problems $\mathbb{P}_3$ and $\mathbb{P}_4$ can be derived in closed form for large $N_t$. Next, we optimize the power allocations to obtain the globally optimal multicast beamformers for large $N_t$. This is considered in Section V-B and V-C using iterative algorithms.

### A. Optimal Combining Coefficients

Following (51), let $\{p_m\}_{m=1}^M$ be any feasible power allocation such that there exists combining coefficients $\{\kappa_k\}_{k=1}^K$ satisfying (6b) and (10) as $N_t \to \infty$. Since the users' SINRs do not change with the phases of $\{\kappa_k\}_{k=1}^K$ in (51), we set $\{\kappa_k\}_{k=1}^K$ to be real positive numbers without loss of the optimality. Consequently, problems $\mathbb{P}_3$ and $\mathbb{P}_4$ with given feasible power allocation $\{p_m\}_{m=1}^M$ reduce to

$$\mathbb{P}_{10}: \max_{\{\kappa_k\}_{k\in\mathcal{K}}} \min_{k\in\mathcal{K}} R(g_k\kappa_k^2, \vartheta_{I(k)}) \tag{52a}$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{K}_m} \kappa_k^2 = 1, \forall m \in \mathcal{M}, \tag{52b}$$

$$g_k\kappa_k^2 \ge \hat{\gamma}_{I(k)}, \forall k \in \mathcal{K}, \tag{52c}$$

and

$$\mathbb{P}_{11}: \max_{\{\kappa_k\}_{k\in\mathcal{K}}} \sum\nolimits_{m=1}^M \min_{k\in\mathcal{K}_m} R(g_k\kappa_k^2, \vartheta_m) \quad \text{s.t.} \quad (52b), (52c), \tag{53}$$

where (52b) follows from (50). Note that there exists no coupling between different groups of multicasting users $\{\kappa_k\}_{k\in\mathcal{K}_m}$ and $\{\kappa_k\}_{k\in\mathcal{K}_{m'}}$ for any $m, m' \in \mathcal{M}$ and $m \neq m'$ in problems $\mathbb{P}_{10}$ and $\mathbb{P}_{11}$. Hence, despite of their different objective functions, $\mathbb{P}_{10}$ and $\mathbb{P}_{11}$ can be divided into exactly

**Algorithm 5** Proposed algorithm for solving $\mathbb{P}_{13}$

Initialize $r^l = \min_{m\in\mathcal{M}} \bar{R}_m^G$, $r^u = \max_{m\in\mathcal{M}} R(\widehat{\Gamma}_m, \vartheta_m)$, set $\varepsilon_5$;
**repeat**
  $\bar{r} = (r^l + r^u)/2$;
  Calculate $\{\bar{p}_m\}$ using (58);
  **if** $\sum_{m\in\mathcal{M}} \bar{p}_m \geq 1$; **then**
    $r^u = \bar{r}$;
  **else**
    $r^l = \bar{r}$;
  **end if**
**until** $|r^l - r^u| \leq \varepsilon_5$.

**Algorithm 6** Proposed algorithm for solving $\mathbb{P}_{15}$

Initialize $\{p_m^{(0)}\}$, $t = 0$, set $\varepsilon_6$ and $T_3^{\max}$;
**repeat**
  Update $\{p_m^{(t+1)}\}$ based on Theorem 6;
  $t = t + 1$;
**until** $\left|\Xi\left(\{p_m^{(t)}\}\right) - \Xi\left(\{p_m^{(t-1)}\}\right)\right| \leq \varepsilon_6$ or $t \geq T_3^{\max}$.

the same set of $M$ independent subproblems, where the subproblem $m$ is given as

$$\mathbb{P}_{12}: \max_{\{\kappa_k\}_{k\in\mathcal{K}_m}} \min_{k\in\mathcal{K}_m} R(g_k\kappa_k^2, \vartheta_m) \tag{54a}$$

$$\text{s.t.} \quad \sum\nolimits_{k\in\mathcal{K}_m} \kappa_k^2 = 1, \tag{54b}$$

$$g_k\kappa_k^2 \geq \hat{\gamma}_m, \forall k \in \mathcal{K}_m. \tag{54c}$$

Meanwhile, although problem $\mathbb{P}_{12}$ is non-convex due to the non-convex objective function and constraints (54b) and (54c), we are able to analytically characterize its optimal solution in the following theorem.

**Theorem 5.** *Problem $\mathbb{P}_{12}$ is feasible if and only if $\widehat{\Gamma}_m p_m \geq \hat{\gamma}_m,\ \forall m \in \mathcal{M}$, where $\widehat{\Gamma}_m = E/\left(\sum_{i\in\mathcal{K}_m} \sigma_i^2/\Gamma_i\right)$. When $\mathbb{P}_{12}$ is feasible, its optimal solution is given by*

$$\kappa_k^* = \frac{1}{\sqrt{\sum_{i\in\mathcal{K}_m} (\Gamma_k\sigma_i^2)/(\Gamma_i\sigma_k^2)}},\ \forall k \in \mathcal{K}_m, \tag{55}$$

*the resulting optimal objective value is $R(\widehat{\Gamma}_m p_m, \vartheta_m) = R(g_k(\kappa_k^*)^2, \vartheta_m), \forall k \in \mathcal{K}_m$.*

*Proof.* Please refer to Appendix F. □

Theorem 5 reveals that when $N_t \to \infty$, the optimal normalized combining coefficients $\{\kappa_k^*\}_{k=1}^K$ for $\mathbb{P}_3$ and $\mathbb{P}_4$ are *independent* of the given power allocation $\{p_m\}_{m=1}^M$. Therefore, the asymptotic multicast beamforming optimization problems $\mathbb{P}_3$ and $\mathbb{P}_4$ are feasible if and only if there exists power allocation $\{p_m\}_{m=1}^M$ such that $\widehat{\Gamma}_m p_m \geq \hat{\gamma}_m,\ \forall m \in \mathcal{M}$. Moreover, for solving $\mathbb{P}_3$ and $\mathbb{P}_4$, we only need to optimize the power allocation with the number of variables being reduced from $MN_t$ to $M$. This significantly lowers the computation time for beamforming design in the FBL regime and improves the system performance.

### *B. Asymptotic MMF Multicast Beamforming Design*

Following Theorem 5 and introducing an auxiliary variable $r$, when $N_t \to \infty$, problem $\mathbb{P}_3$ simplifies into

$$\mathbb{P}_{13}: \max_{\{p_m\},r} \quad r \tag{56a}$$

$$\text{s.t.} \quad \sum\nolimits_{m=1}^M p_m \leq 1, \tag{56b}$$

$$\widehat{\Gamma}_m p_m \geq \hat{\gamma}_m, \forall m \in \mathcal{M}, \tag{56c}$$

$$r \leq R(\widehat{\Gamma}_m p_m, \vartheta_m), \forall m \in \mathcal{M}. \tag{56d}$$

Note that when $\mathbb{P}_{13}$ is feasible, the constraint (56b) must be satisfied with equality at the optimality, i.e., $\sum_{m=1}^M p_m = 1$. To solve problem $\mathbb{P}_{13}$, we first consider the following feasibility problem for given $\bar{r} \in \mathbb{R}$:

$$\mathbb{P}_{14}: \text{find} \ \{p_m\} \tag{57}$$

$$\text{s.t. (56b), (56c)},$$

$$\bar{r} \leq R(\widehat{\Gamma}_m p_m, \vartheta_m), \forall m \in \mathcal{M}.$$

Let $r^*$ denote the optimal objective value of $\mathbb{P}_{13}$. Problem $\mathbb{P}_{14}$ is feasible if and only if $r^* \geq \bar{r}$; otherwise, $r^* < \bar{r}$. This implies that the globally optimal solution of $\mathbb{P}_{13}$ can be obtained using a bisection procedure as given in Algorithm 5, which solves a sequence of problems $\mathbb{P}_{14}$ to tighten the gap between $\bar{r}$ and $r^*$. Recall that $R(\hat{\gamma}_m, \vartheta_m) = \bar{R}_m^G,\ \forall m \in \mathcal{M}$. Based on Lemma 1, if $\bar{r} \geq \bar{R}_m^G$, there must be a unique solution $\widetilde{p}_m > \hat{\gamma}_m/\widehat{\Gamma}_m$ such that $R(\widehat{\Gamma}_m\widetilde{p}_m, \vartheta_m) = \bar{r}$; otherwise, $p_m = \hat{\gamma}_m/\widehat{\Gamma}_m$ holds in (56c). Thus $\mathbb{P}_{14}$ is feasible if and only if

$$\bar{p}_m = \max\left\{\hat{\gamma}_m/\widehat{\Gamma}_m, \widetilde{p}_m\right\},\ \forall m \in \mathcal{M}, \tag{58}$$

satisfies $\sum_{m=1}^M \bar{p}_m \leq 1$.

### *C. Asymptotic WSR Multicasting Beamforming Design*

Similarly, following Theorem 5, when $N_t \to \infty$, problem $\mathbb{P}_4$ can be simplified as

$$\mathbb{P}_{15}: \max_{\{p_m\}} \ \Xi(\{p_m\}) = \sum\nolimits_{m=1}^M \omega_m R(\widehat{\Gamma}_m p_m, \vartheta_m) \tag{59a}$$

$$\text{s.t.} \quad \sum\nolimits_{m=1}^M p_m = 1, \tag{59b}$$

$$\text{(56c)}.$$

From Theorem 1, the objective function of $\mathbb{P}_{15}$ is non-concave in general, making problem $\mathbb{P}_{15}$ non-convex. We first rewrite $\mathbb{P}_{15}$ in a convex form and then solve it using an iterative algorithm. To this end, we note that $\Omega(x) = \sqrt{1-(1+x)^{-2}}$ is a concave function with second-order derivative

$$\frac{d^2\Omega(x)}{dx^2} = \frac{-3x^2 - 6x - 1}{\left(1-(1+x)^{-2}\right)^{3/2}(1+x)^6} < 0,\ \forall x > 0. \tag{60}$$

This implies that $\Omega(x)$ is upper bounded by its first-order Taylor expansion, i.e., $\Omega(x) \leq \Omega_1(x_0)x + \Omega_2(x_0),\ \forall x_0 > 0$, where $\Omega_1(x_0) = (1+x_0)^{-3}/\sqrt{1-(1+x_0)^{-2}} > 0$ and $\Omega_2(x_0) = \Omega(x_0) - \Omega_1(x_0)x_0$. Then $\mathbb{P}_{15}$ can be approximated as:

$$\mathbb{P}_{16}: \max_{\{p_m\}} \ \sum\nolimits_{m=1}^M \omega_m \left(\ln(1+\widehat{\Gamma}_m p_m) - \chi_m p_m - \nu_m\right) \tag{61}$$

$$\text{s.t. (56c), (59b)}.$$

where $\chi_m = \vartheta_m\widehat{\Gamma}_m\Omega_1(\widehat{\Gamma}_m p_m^{(t)})$, $\nu_m = \vartheta_m\Omega_2(\widehat{\Gamma}_m p_m^{(t)})$, and $\{p_m^{(t)}\}_{m=1}^M$ is a feasible solution satisfying (56c) and (59b).

Problem $\mathbb{P}_{16}$ is convex and admits an optimal solution as given in the following theorem.

**Theorem 6.** *The optimal solution of $\mathbb{P}_{16}$ is given by*

$$p_m^* = \max\left\{\frac{\omega_m}{\omega_m\chi_m + \pi^*} - \frac{1}{\widehat{\Gamma}_m}, \hat{\gamma}_m/\widehat{\Gamma}_m\right\},\ \forall m \in \mathcal{M}, \tag{62}$$

*where $\pi^*$ is chosen such that $\sum_{m=1}^M p_m^* = 1$.*

*Proof.* Problem $\mathbb{P}_{16}$ has a similar formulation as conventional power allocation problem for sum-rate maximization. Hence it also adopts a waterfilling-like solution in (62) [44]. □

The algorithm for solving $\mathbb{P}_{15}$ provides a solution of the MM-type and is summarized in Algorithm 6, which converges to a locally optimal solution of $\mathbb{P}_{15}$ [50]. Moreover, we have the following result on the globally optimal solution of $\mathbb{P}_{15}$.

**Theorem 7.** *When $\mathbb{P}_{15}$ is feasible and additionally, $\hat{\gamma}_m \geq \bar{\gamma}_{3,m}, \forall m \in \mathcal{M}$, problem $\mathbb{P}_{15}$ is convex and its optimal solution is given by*

$$p_m^* = \begin{cases} \breve{p}_m, & \text{if } \omega_m\widehat{\Gamma}_m\Delta(\hat{\gamma}_m, \vartheta_m) \geq \tau^*, \\ \hat{\gamma}_m/\widehat{\Gamma}_m, & \text{otherwise}, \end{cases} \tag{63}$$

*where $\bar{\gamma}_{3,m} = \bar{\gamma}_3(\vartheta_m)$ is given in Lemma 2, $\tau^*$ is chosen such that $\sum_{m=1}^M p_m^* = 1$, $\breve{p}_m \geq 0$ is the unique solution to equation $\omega_m\widehat{\Gamma}_m\Delta(\widehat{\Gamma}_m\breve{p}_m, \vartheta_m) = \tau^*$ if $\omega_m\widehat{\Gamma}_m\Delta(\hat{\gamma}_m, \vartheta_m) \geq \tau^*$, and*

$$\Delta(x, y) = \frac{1}{1+x}\left(1 - y\frac{1}{(1+x)\sqrt{(1+x)^2 - 1}}\right). \tag{64}$$

*Proof.* Please refer to Appendix G. □

## VI. Simulation Results

In this section, we evaluate the performance of the proposed algorithms by simulation. We consider a multicasting from a BS with $N_t$ antennas to $K$ single-antenna users over a system bandwidth of $B = 1$ MHz and a transmission blocklength of $N$. The users are uniformly and randomly located in a disk with an inner radius of 50 meters and an outer radius of 200 meters. The users are divided into $M$ groups of equal size. We consider the independent Rayleigh fading channel $\widetilde{\mathbf{h}}_k \sim \mathcal{CN}(\mathbf{0}_{N_t}, \mathbf{I}_{N_t})$ and calculate the large-scale path loss as $\Gamma_k = -35.3 - 37.6\log_{10}L_k$ dB, where $L_k$ is the distance between the BS and user $k$ in meters [8], [12], [14], [16], [26]. The noise power spectral density is $\sigma_k^2 = -174$ dBm/Hz. Unless otherwise specified, we set $M = 3, K = 12, N_t = 64$, $T_{\max} = 0.1$ ms, $\omega_m = 1/M$, and $N = BT_{\max} = 100$; meanwhile, each multicast group requires a BLER of $\epsilon_m = 10^{-e(m)}$ according to $e(m) = \min\{5 + (m-1) \times \lceil 5/(M-1) \rceil, 10\}$, a minimum data volume $D_m = D = 50$ nats, and a minimum rate $\bar{R}_m^G = \bar{R} = 0.5$ nats/s/Hz, $\forall m \in \mathcal{M}$. The results of finite $N_t$ are averaged over 100 random realizations of channel $\mathbf{h}_k, \forall k \in \mathcal{K}$. Similar to [27], the value of the objective function is set to zero if the obtained solution violates any constraint. We set $\rho = 0.1$ for both Algorithms 1 and 3. For the multicast beamforming with finite $N_t$, semi-definite relaxation (SDR) in [7] is used for initialization. [7]

### A. MMF Beamforming

First, we evaluate the MMF performance of the following schemes: (i) 'Proposed', namely the proposed Algorithm 2; (ii) 'BFwMOSEK', which solves the subproblem $\mathbb{P}_5$ in the proposed Algorithm 2 using Mosek [51]; (iii) 'SDR-IBL', which assumes the infinite blocklength in problem $\mathbb{P}_1$ and solves it with SDR [7]; (iv) 'SDR-FBL', which evaluates the solution obtained from 'SDR-IBL' directly with the FBL rate; and (v) 'SDR-GauRan-FBL', which differs from 'SDR-FBL' in approximating the solution from 'SDR-IBL' using the Gaussian randomization before employing it for FBL transmission. We set $P = 10$ mW.

Fig. 2 and Table I show the average minimum rate and the computation time of different considered schemes versus the number of transmit antennas $N_t$, respectively. We observe that, as expected, 'SDR-IBL' provides a performance upper bound for the minimum rate since it employs the infinite blocklength. On the other hand, the 'SDR-FBL' and 'SDR-GauRan-FBL' approaches lead to not only poor MMF performance but also high computation time. This result implies that it is imperative to consider the FBL rate and design low-complexity algorithms tailored for optimizing massive MIMO multicast beamforming in the FBL regime. Meanwhile, the proposed and the 'BFwMOSEK' schemes achieve similar performance, which validates the effectiveness of the proposed Algorithm 1. Interesting, the proposed scheme always achieves both the best performance and the lowest computation time among the considered FBL transmission schemes, irrespective of the number of transmit antennas $N_t$. The latter is due to the parallel updating of beamforming optimization in closed or semi-closed form, whereby the computation requirement of our proposed scheme increases much slower with $N_t$ than the other schemes. Thus, our proposed scheme is particularly attractive for massive MIMO systems in the FBL regime. For example, the proposed design is about 10~350 times faster than the SDR methods and about 3~160 times faster than 'BFwMOSEK'.

Fig. 3 illustrates the minimum rate versus the transmission blocklength $N$. We observe that the minimum rates of the considered FBL transmission schemes increase monotonically with the blocklength and only coincide with that of the 'SDR-IBL' scheme for very large blocklength $N$, before the FBL rate (4) approaches the Shannon rate $\ln(1 + \gamma_k(\mathbf{W}))$. However, large performance gaps exist for small $N$, e.g., $N = 100$, where the impact of blocklength needs to be considered for multicast beamforming optimization to guarantee reliable communications.

We further evaluate the minimum multicast rate of different groups over the considered feasible channel realizations for the MMF beamforming design in Fig. 4, where we set $N = 100$, and $P = 0.8$ mW. As can be seen from Fig. 4, the QoS

[7] Since the considered schemes use the same initial solutions, the time for initialization is not included in calculating the computation time. For fast initialization, the alternating direction method of multipliers (ADMM) algorithm in [11] can be used.

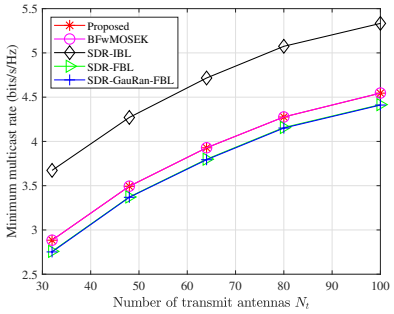


Figure 2. Minimum multicast rate vs. number of transmit antennas $N_t$ ($P = 10$ mW, $M = 3$, $K = 12$, and $N = 100$).

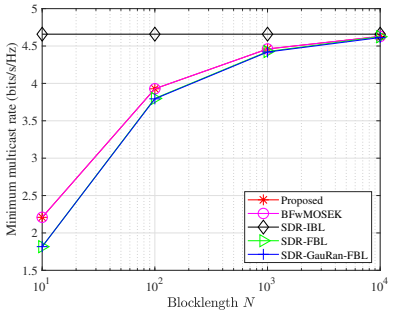


Figure 3. Minimum multicast rate vs. blocklength $N$ ($P = 10$ mW, $M = 3$, $K = 12$, $N_t = 64$).

Table I
AVERAGE COMPUTATION TIME OF DIFFERENT MMF BEAMFORMING DESIGNS OVER $N_t$ IN SECONDS ($P = 10$ MW, $M = 3$, $K = 12$, AND $N = 100$).

| $N_t$ | 32 | 48 | 64 | 80 | 100 |
|---|---|---|---|---|---|
| Proposed | 2.93 | 3.59 | 4.43 | 4.83 | 5.49 |
| BFwMOSEK | 9.3 | 58.3 | 101.2 | 321.41 | 912.1 |
| SDR[8] | 32.59 | 53.11 | 343.74 | 631.77 | 1922.5 |

requirement (6c) can be strictly fulfilled with our proposed design. On the contrary, the 'SDR-GauRan-FBL' scheme using the Shannon rate cannot meet the QoS requirement. This result indicates that it is necessary to consider the impact of blocklength to ensure the QoS requirements of FBL communications.

### *B. WSR Beamforming*

We now evaluate the WSR performance of the following schemes: (i) 'Proposed', namely the proposed Algorithm 4; (ii) 'BFwMOSEK', which solves problem $\mathbb{P}_8$ in the proposed Algorithm 4 using Mosek [51]; (iii) 'Proposed-IBL', which assumes the infinite blocklength in problem $\mathbb{P}_2$ and solves it via the proposed Algorithm 4; (iv) 'Conventional alg.', which employs the solution from 'Proposed-IBL' directly for FBL transmission; and (v) 'WMMSE', which differs from (iv) in solving problem $\mathbb{P}_2$ with the infinite blocklength via the WMMSE method [21] before employing the obtained solution for FBL transmission. The FBL rate is evaluated for both 'Conventional alg.' and 'WMMSE' schemes.

Fig. 5 and Table II present the WSR and the computation time of the considered schemes versus the number of transmit antennas $N_t$, respectively, where $P = 1.5$ mW. As expected, since 'Proposed-IBL' employs the infinite blocklength, it provides a performance upper bound. On the other hand, the WMMSE method not only has a poor performance similar to 'Conventional alg.', but also leads to a higher computation time, highlighting the importance of low-complexity beamforming solutions for FBL multicast. Moreover, we observe that the proposed and 'BFwMOSEK' schemes achieve similar performance, which proves the effectiveness of Algorithm 3. Due to the parallel updating for beamforming optimization in closed or semi-closed forms, our proposed scheme always achieves both the best performance and lowest computation time among the considered FBL transmission schemes. For example, the proposed scheme is about 14~220 times faster than 'BFwMOSEK' and about 29~2200 times faster than the WMMSE method. Due to the significant reductions in the processing delay (cf. Table I as well), our proposed low-complexity solutions can also contribute to lower the E2E latency for latency-sensitive applications. Interestingly, the 'Proposed', 'BFwMOSEK', and 'Conventional alg.' schemes have similar performance for large $N_t$, e.g., $N_t = 100$. This is because the users' SINR are large enough such that the channel dispersion $V(\gamma_k(\mathbf{W})) \approx 1$ in (4) and its penalty on the achievable rate in the objective function of problem $\mathbb{P}_2$ can be ignored.

[8]The computation time of 'SDR-IBL' or 'SDR-FBL' is almost the same as that of 'SDR-GauRan-FBL' because the time for Gaussian randomization is negligible. Therefore, we only consider the computation time for SDR.

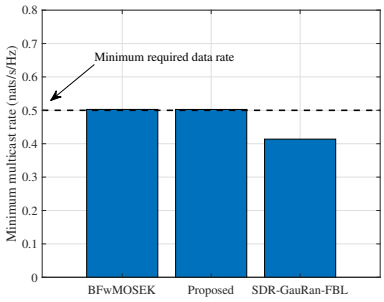


Figure 4. Minimum multicast rate of different groups for the MMF multicast beamforming ($P = 0.8$ mW, $M = 3$, $K = 12$, $N_t = 64$).

Fig. 6 shows the WSR versus the transmit power $P$. We observe that a large performance gap between the proposed and 'Conventional alg.' schemes also exists when the transmit power is low, e.g., $P = 1$ mW. This is because the 'Conventional alg.' scheme based on the Shannon rate has a high probability of violating the QoS constraints (6c). Thus, using the FBL rate is crucial for multicast beamforming design with low transmit power.

Fig. 7 illustrates the minimum multicast rate of different groups over the considered feasible channel realizations for the WSR beamforming design, where we set $N = 100$, and $P = 0.8$ mW. We observe that our proposed design can successfully meet the QoS requirement (6c). On the contrary, the minimum multicast rate of 'WMMSE' and 'Proposed-IBL' are only 7% of the required data rate, for which the QoS constraint (6c) cannot be guaranteed. This result again demonstrates the necessity of considering the impact of FBL in latency-sensitive and reliable communications.

Table II
AVERAGE COMPUTATION TIME OF DIFFERENT WSR BEAMFORMING DESIGNS OVER $N_t$ IN SECONDS ($P = 1.5$ MW, $M = 3$, $K = 12$, AND $N = 100$).

| $N_t$ | 32 | 48 | 64 | 80 | 100 |
|---|---|---|---|---|---|
| Proposed | 0.891 | 0.988 | 1.138 | 1.146 | 1.58 |
| BFwMOSEK | 13.214 | 28.157 | 80.106 | 243.18 | 353.529 |
| WMMSE | 26.153 | 128.65 | 576.71 | 2211.9 | N/A(>3600) |

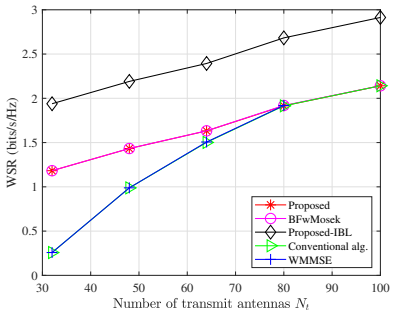


Figure 5. WSR vs. number of transmit antennas $N_t$ ($P = 1.5$ mW, $M = 3$, $K = 12$, and $N = 100$).

### C. Asymptotic Beamforming

We further evaluate the performance of the asymptotic beamforming designs. The asymptotic results are averaged over 500 random realizations of user positions. For each realization of user positions, 500 independent scenarios of small-scale fading, $\tilde{\mathbf{h}}_k, \forall k \in \mathcal{K}$, are generated.

First, we compare the MMF performance of the following schemes: (i) 'Proposed', namely the proposed Algorithm 5; (ii) 'Proposed-IBL', namely the proposed Algorithm 5 with the infinite blocklength; and (iii) 'Conventional alg.', where the solution from 'Proposed-IBL' is employed for FBL transmission. We set $\bar{R}_m^G = \bar{R} = 1$ nats/s/Hz, $\forall m \in \mathcal{M}$. Fig. 8 shows the average minimum rate versus $E$. We use color and marker to distinguish different schemes and use solid, dashed, and dotted lines to denote the asymptotic result, simulation results with $N_t = 1000$ and $N_t = 10000$, respectively. We observe that in both asymptotic analysis and simulation, our proposed solution achieves higher performance than 'Conventional alg.' scheme.

Next, we compare the WSR performance of the following schemes: (i) 'Proposed', namely the proposed Algorithm 6; (ii) 'Proposed-IBL', namely the proposed Algorithm 6 with the infinite blocklength; (iii) 'Conventional alg.', where the solution from 'Proposed-IBL' is employed for FBL transmission; and (iv) 'Optimal solution', which is obtained from Theorem 7. Fig. 9 shows the average WSR versus $E$. For simplicity, we set $\epsilon_m = 10^{-8}, \forall m \in \mathcal{M}$ and consider two cases, i.e., $\bar{R} = 1$ nats/s/Hz and $\bar{R} = 0.01$ nats/s/Hz. For $\bar{R} = 1$ nats/s/Hz, the globally optimal solution of problem $\mathbb{P}_{15}$ is given by Theorem 7, which is also plotted. From Fig. 9, we observe that the optimal solution and the proposed Algorithm 6 achieve similar performance. This is expected because, according to Theorem 7, problem $\mathbb{P}_{15}$ is convex such that its locally and the globally optimal solutions coincide with each other. Meanwhile, both the optimal and the proposed schemes significantly outperform 'Conventional alg.', particularly when $E$ is small. On the other hand, when $\bar{R} = 0.01$ nats/s/Hz, Theorem 7 cannot be used to obtain the globally optimal solution of problem $\mathbb{P}_{15}$. Nevertheless, the proposed Algorithm 6 can still be used to obtain the locally optimal solution of problem $\mathbb{P}_{15}$ , which is shown to outperform 'Conventional alg.' scheme. Moreover, as can be seen from Fig. 9, the performance gap between our proposed scheme and the conventional design narrows as the transmit power increases. This is because the users' SINR will be large enough such that the channel dispersion $V\left(\widehat{\Gamma}_m p_m\right) \approx 1$ and its impact on the objective function of problem $\mathbb{P}_{15}$ becomes negligible.

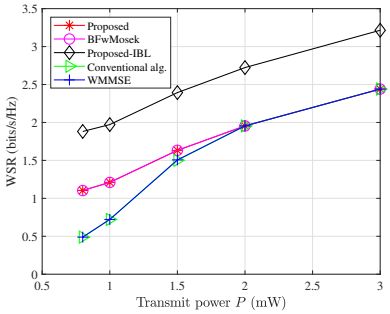


Figure 6. WSR vs. transmit power $P$ in mW ($N_t = 64$, $M = 3$, $K = 12$, and $N = 100$).

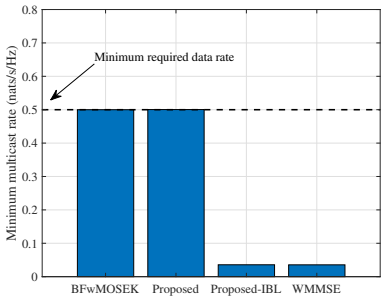


Figure 7. Minimum multicast rate of different groups for WSR multicast beamforming ($P = 0.8$ mW, $M = 3$, $K = 12$, $N_t = 64$).

## VII. CONCLUSIONS

In this paper, we studied the beamforming designs for massive MIMO multicast in the FBL regime under the MMF and WSR criteria. We revealed that the non-negative FBL rate is a concave function of the received SINR if and only if the function parameter exceeds a constant threshold. Considering finite number of transmit antennas at the BS, we proposed low-complexity algorithms to obtain the locally optimal solutions of the formulated problems, where variables are updated in closed or semi-closed form. For an unlimited number of transmit antennas at the BS, we showed that the asymptotic optimal beamformer of each group is a linear combination of the channel vectors of users in the group and derived the optimal normalized combining coefficients in closed form. Based on this fact, we obtain the globally optimal multicast beamformers by reducing the problems to power allocation optimizations and further solve them via iterative algorithms. Simulation results showed that our proposed designs outperform several

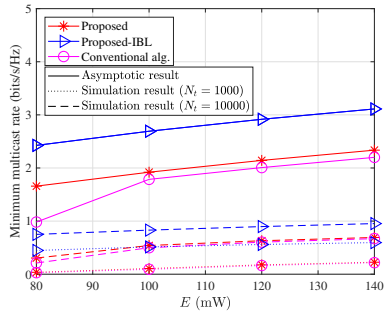


Figure 8. Minimum multicast rate vs. $E$ in mW ($M = 3$, $K = 12$, and $N = 100$).

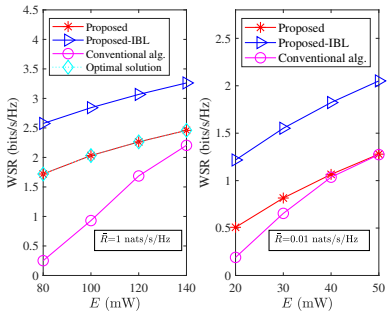


Figure 9. WSR vs. $E$ in mW ($M = 3$, $K = 12$, and $N = 100$).

baseline schemes and have a substantial computational reductions of up to two orders of magnitude over the benchmarks, which are appealing for massive MIMO systems.

## Appendix A
## Proof of Theorem 1

The first- and second-order derivatives of $R(\gamma, \vartheta)$ w.r.t. $\gamma$ are given by

$$\frac{\partial R(\gamma,\vartheta)}{\partial \gamma} = \frac{1}{1+\gamma}\left(1 - \vartheta \frac{1}{(1+\gamma)\sqrt{(1+\gamma)^2-1}}\right), \quad (65)$$

$$\frac{\partial^2 R(\gamma,\vartheta)}{\partial \gamma^2} = -\frac{1-\vartheta\varphi(\gamma)}{(1+\gamma)^2}, \quad (66)$$

where

$$\varphi(\gamma) = \frac{3(1+\gamma)^2 - 2}{(1+\gamma)\left((1+\gamma)^2-1\right)^{\frac{3}{2}}}. \quad (67)$$

Since $\bar{\gamma}_2 > 0$ and $R(\bar{\gamma}_2, \vartheta) = 0$, we have

$$\frac{\vartheta}{1+\bar{\gamma}_2} = \frac{\ln(1+\bar{\gamma}_2)}{\sqrt{(\bar{\gamma}_2+2)\bar{\gamma}_2}}. \quad (68)$$

From Lemma 2 and (66), we have $\vartheta\varphi(\gamma) < 1$ for $\gamma > \bar{\gamma}_3$ and $\vartheta\varphi(\gamma) \geq 1$ for $\gamma \leq \bar{\gamma}_3$. Based on this, now we prove that $\bar{\gamma}_2 \leq \bar{\gamma}_3$ if and only if $\bar{\gamma}_2 \in (0, x_h]$, where $x_h > 0$ is a constant. Plugging (68) into (67) and assuming $\bar{\gamma}_2 \leq \bar{\gamma}_3$, we have $h(\bar{\gamma}_2) \geq 0$, where

$$h(x) = \ln(1+x)[3(1+x)^2 - 2] - ((1+x)^2 - 1)^2. \quad (69)$$

Moreover, we have

$$h(0) = 0, \ \lim_{x\to\infty} h(x) = -\infty. \quad (70)$$

$$h'(x) = \frac{3x^2+6x+1}{1+x} - 4x(x+2)(1+x) + 6(1+x)\ln(1+x), \quad (71)$$

$$h''(x) = \frac{1}{(1+x)^2}\left(-12x^4 - 48x^3 - 59x^2 - 22x, + 6(1+x)^2\ln(1+x) + 3\right) \quad (72)$$

$$h'''(x) = -\frac{2[12x^4 + 48x^3 + 69x^2 + 42x + 11]}{(1+x)^3}. \quad (73)$$

One can verify that (i) $h'(0) > 0$, and $\lim_{x\to\infty} h'(x) = -\infty$; (ii) $h''(0) > 0$, $\lim_{x\to\infty} h''(x) = -\infty$; and (iii) $h'''(x) < 0$ for $x \geq 0$. Here (ii) and (iii) further imply that (iv) $h''(x)$ is monotonic decreasing for $x > 0$ and $h''(x) = 0$ has a unique solution. Moreover, (i) and (iv) imply that (v) $h'(x)$ is monotonic increasing for $x \in (0, x_\delta]$ and decreasing for $x \in (x_\delta, +\infty)$ for some $x_\delta > 0$. Combining (70), (i), and (v), we have $x \in (0, x_h]$ for $h(x) \geq 0$, where $x_h > 0$ is a constant threshold. Then $\bar{\gamma}_2 \leq \bar{\gamma}_3$ if and only if $\bar{\gamma}_2 \in (0, x_h]$. Note that $\bar{\gamma}_2$ is a function of $\vartheta$ whose analytical expression $\bar{\gamma}_2(\vartheta)$ is difficult to obtain. However, we have

$$\frac{d\bar{\gamma}_2}{d\vartheta} = -\frac{\frac{dR(\bar{\gamma}_2,\vartheta)}{d\vartheta}}{\frac{dR(\bar{\gamma}_2,\vartheta)}{d\bar{\gamma}_2}} > 0, \quad (74)$$

since

$$\frac{dR(\bar{\gamma}_2,\vartheta)}{d\vartheta} = -\sqrt{\frac{(\bar{\gamma}_2+2)\bar{\gamma}_2}{(1+\bar{\gamma}_2)^2}} < 0, \quad (75)$$

$$\frac{dR(\bar{\gamma}_2,\vartheta)}{d\bar{\gamma}_2} > 0. \quad (76)$$

One can verify that $\bar{\gamma}_2$ is monotonic increasing for $\vartheta > 0$ and $\bar{\gamma}_2 \leq x_h$ is equivalent to $\vartheta \leq \hat{\vartheta}$ for a threshold $\hat{\vartheta} > 0$ satisfying $\bar{\gamma}_2(\hat{\vartheta}) = x_h$. Note that $\hat{\vartheta}$ is a constant and we can estimate that $\hat{\vartheta} \approx 0.651\ 12$ using numerical methods, e.g., bisection method. This completes the proof.

## Appendix B
## Proof of Theorem 2

Defining $\iota_k(\mathbf{W}) = \mathbf{h}_k^H \mathbf{w}_{I(k)}$ and using $\alpha_k(\mathbf{W})$ and $\beta_k(\mathbf{W})$ are defined in (19) and (20), we have

$$\ln\left(1 + \frac{|\iota_k(\mathbf{W})|^2}{\alpha_k(\mathbf{W})}\right) = -\ln\left(1 - \frac{|\iota_k(\mathbf{W})|^2}{\beta_k(\mathbf{W})}\right), \quad (77)$$

which is jointly convex w.r.t. $\iota_k(\mathbf{W})$ and $\beta_k(\mathbf{W})$ [52]. Then (77) can be lower bounded by its first-order Taylor expansion as

$$\ln\left(1+\gamma_k(\mathbf{W})\right) \geq f_k^{(t)}(\mathbf{W}, \mathbf{W}^{(t)}) \triangleq \ln\left(1+\gamma_k(\mathbf{W}^{(t)})\right) - \gamma_k(\mathbf{W}^{(t)}) + 2\Re\left\{\frac{\iota_k^\star(\mathbf{W}^{(t)})\iota_k(\mathbf{W})}{\alpha_k(\mathbf{W}^{(t)})}\right\} - \frac{\beta_k(\mathbf{W})}{\alpha_k(\mathbf{W}^{(t)})} + \frac{\beta_k(\mathbf{W})}{\beta_k(\mathbf{W}^{(t)})}. \quad (78)$$

Note that $f_k^{(t)}(\mathbf{W}, \mathbf{W}^{(t)})$ is concave with $\mathbf{W}$.

Meanwhile, since $\sqrt{V\left(\gamma_k(\mathbf{W})\right)}$ is a concave function of $V\left(\gamma_k(\mathbf{W})\right)$, an upper bound of $\sqrt{V\left(\gamma_k(\mathbf{W})\right)}$ can be obtained as

$$\sqrt{V\left(\gamma_k(\mathbf{W})\right)} \leq \frac{\sqrt{V\left(\gamma_k(\mathbf{W}^{(t)})\right)}}{2}\left(1+\frac{1}{V\left(\gamma_k(\mathbf{W}^{(t)})\right)}\right) - \frac{1}{2\sqrt{V\left(\gamma_k(\mathbf{W}^{(t)})\right)}}\left(\frac{\alpha_k(\mathbf{W})}{\beta_k(\mathbf{W})}\right)^2. \quad (79)$$

Moreover, since $x^2$ is a convex function, by setting $x = \alpha_k(\mathbf{W})/\beta_k(\mathbf{W})$, we obtain

$$\left(\frac{\alpha_k(\mathbf{W})}{\beta_k(\mathbf{W})}\right)^2 \geq \frac{2\alpha_k(\mathbf{W}^{(t)})\alpha_k(\mathbf{W})}{\beta_k(\mathbf{W}^{(t)})\beta_k(\mathbf{W})} - \left(\frac{\alpha_k(\mathbf{W}^{(t)})}{\beta_k(\mathbf{W}^{(t)})}\right)^2. \quad (80)$$

Similarly, since the function $|x|^2/y$ is jointly convex with respect to $x \in \mathbb{C}$ and $y > 0$, we have

$$\frac{\alpha_k(\mathbf{W})}{\beta_k(\mathbf{W})} \geq \frac{2\left(\sum\limits_{m\neq I(k)} \Re\left\{(\mathbf{w}_m^{(t)})^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{w}_m\right\} + \sigma_k^2\right)}{\beta_k(\mathbf{W}^{(t)})} - \frac{\alpha_k(\mathbf{W}^{(t)})\beta_k(\mathbf{W})}{\beta_k^2(\mathbf{W}^{(t)})}. \quad (81)$$

Theorem 1 can be proved by combining (78), (79), (80), and (81).

## Appendix C
## Proof of Theorem 3

The Lagrangian of (35) is given by

$$\mathcal{L}_\mu = \sum\nolimits_{m=1}^{M} \left|\Gamma_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Lambda_{k,m}^{(n)}\right|^2 + \mu\Theta_k(\mathbf{\Gamma}_k), \quad (82)$$

where $\Theta_k(\mathbf{\Gamma}_k)$ is defined in (28), and $\mu \geq 0$ is the dual variable for (35b). The optimal solution of (35) is characterized by the KKT conditions:

$$\frac{1}{2}\frac{\partial \mathcal{L}_\mu}{\partial \Gamma_{k,m}} = \quad (83)$$

$$\begin{cases} \Gamma_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Lambda_{k,m}^{(n)} - \mu \mathbf{h}_k^H \mathbf{w}_m^{(t)} = 0, & \text{if } m = I(k), \\ \Gamma_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Lambda_{k,m}^{(n)} + \mu\hat{\gamma}_{I(k)}\Gamma_{k,m} = 0, & \text{otherwise}, \end{cases}$$

$$\Theta_k(\mathbf{\Gamma}_k) \leq 0, \quad (84)$$

$$\mu \geq 0, \quad (85)$$

$$\mu\Theta_k(\mathbf{\Gamma}_k) = 0, \quad (86)$$

The optimal solution (36) can be immediately obtained from (83). Substituting (36) into $\Theta_k(\mathbf{\Gamma}_k)$, we have

$$\psi(\mu) = \Theta_k(\mathbf{\Gamma}_k^*) = \frac{\hat{\gamma}_{I(k)} \sum\limits_{m\neq I(k)} |\mathbf{h}_k^H \mathbf{w}_m^{(n)} - \Lambda_{k,m}^{(n)}|^2}{(1+\mu\hat{\gamma}_{I(k)})^2} - 2\Re\left\{(\mathbf{w}_{I(k)}^{(t)})^H \mathbf{h}_k (\mathbf{h}_k^H \mathbf{w}_{I(k)}^{(n)} - \Lambda_{k,I(k)}^{(n)})\right\} + (1-2\mu)|\mathbf{h}_k^H \mathbf{w}_{I(k)}^{(t)}|^2 + \hat{\gamma}_{I(k)}\sigma_k^2, \quad (87)$$

where $\psi(\mu)$ is monotonically decreasing for $\mu \geq 0$. According to (86), $\mu^* = 0$ if $\psi(0) < 0$; otherwise, there exists a unique root $\mu^* > 0$ such that $\psi(\mu^*) = 0$. Note that solving the equation $\psi(\mu^*) = 0$ involves finding the roots of a cubic function, where closed-form expressions for the roots can be obtained similar to [11].

## Appendix D
## Proof of Theorem 4

The Lagrangian of (37) is given by

$$\mathcal{L}_\varrho = \varrho\Upsilon_k(d_k, \mathbf{q}_k) + \sum\nolimits_{m=1}^{M} \left|q_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Psi_{k,m}^{(n)}\right|^2 + \left|d_k - r^{(n)} + u_k^{(n)}\right|^2, \quad (88)$$

where $\Upsilon_k(d_k, \mathbf{q}_k)$ is defined in (29), and $\varrho \geq 0$ is the dual variable for (37b). The optimal solution of (37) is characterized by the KKT conditions:

$$\frac{\partial \mathcal{L}_\varrho}{\partial q_{k,m}} = 2(q_{k,m} - \mathbf{h}_k^H \mathbf{w}_m^{(n)} + \Psi_{k,m}^{(n)}) + 2\varrho c_k^{(t)} q_{k,m} - \varrho(\delta_{k,m}^{(t)})^\star = 0, \quad (89)$$

$$\frac{\partial \mathcal{L}_\varrho}{\partial d_k} = 2(d_k - r^{(n)} + u_k^{(n)}) + \varrho = 0, \quad (90)$$

$$\varrho \geq 0, \quad (91)$$

$$\Upsilon_k(d_k, \mathbf{q}_k) \leq 0, \quad (92)$$

$$\varrho\Upsilon_k(d_k, \mathbf{q}_k) = 0. \quad (93)$$

The optimal solution (38) can be obtained from (89) and (90). Substituting (38) into $\Upsilon_k(d_k, \mathbf{q}_k)$, we have

$$\phi(\varrho) = \Upsilon_k(d_k^*, \mathbf{q}_k^*) = \frac{2r^{(n)} - 2u_k^{(n)} - \varrho}{2} - a_k^{(t)} - \sum_{m=1}^{M} \Re\left\{\delta_{k,m}^{(t)} \frac{2\mathbf{h}_k^H \mathbf{w}_m^{(n)} - 2\Psi_{k,m}^{(n)} + \varrho(\delta_{k,m}^{(t)})^\star}{2 + 2\varrho c_k^{(t)}}\right\} + c_k^{(t)} \sum_{m=1}^{M} \left|\frac{2\mathbf{h}_k^H \mathbf{w}_m^{(n)} - 2\Psi_{k,m}^{(n)} + \varrho(\delta_{k,m}^{(t)})^\star}{2 + 2\varrho c_k^{(t)}}\right|^2. \quad (94)$$

We can show that $\phi(\varrho)$ is a monotonically decreasing function for $\varrho \geq 0$, since the first-order derivative of $\phi(\varrho)$ satisfies

$$\phi'(\varrho) = -\sum_{m=1}^{M} \frac{\left|2(\delta_{k,m}^{(t)})^\star - 4c_k^{(t)}(\mathbf{h}_k^H \mathbf{w}_m^{(n)} - \Psi_{k,m}^{(n)})\right|^2}{\left(2 + 2\varrho c_k^{(t)}\right)^3} - \frac{1}{2} < 0. \quad (95)$$

According to (93), $\varrho^* = 0$ if $\phi(0) < 0$; otherwise, there exists a unique root $\varrho^* > 0$ such that $\phi(\varrho^*) = 0$. Here $\varrho^*$ can be obtained using bisection or Newton methods.

## Appendix E
## Proof of Lemma 4

Here we prove Lemma 4 by contradiction, similar to [14]. Suppose that Lemma 4 is not true. Then the optimal multicast beamforming vectors can be written as

$$\mathbf{w}_m = \sum\nolimits_{k\in\mathcal{K}_m} \xi_k \mathbf{h}_k + \sum\nolimits_{t=1}^{N_t - K_m} \zeta_{t,m}\mathbf{e}_{t,m}, \ \forall m \in \mathcal{M}, \quad (96)$$

where $\{\mathbf{e}_{t,m}\}_{t=1}^{N_t-K_m}$ is an orthogonal basis of the orthogonal complement of the subspace spanned by $\{\mathbf{h}_k\}_{k\in\mathcal{K}_m}$, and at

least one of $\zeta_{t,m},\ t = 1,\ldots,N_t - K_m$ is non-zero . The received signal of user $k$ is expressed as

$$\begin{aligned} y_k =&\mathbf{h}_k^H \mathbf{w}_{I(k)} s_{I(k)} + \sum\nolimits_{i\neq I(k), i\in\mathcal{M}} \mathbf{h}_k^H \mathbf{w}_i s_i + n_k \\ \overset{(a)}{=}&\xi_k \mathbf{h}_k^H \mathbf{h}_k s_{I(k)} + \sum\nolimits_{i\neq I(k), i\in\mathcal{M}} \mathbf{h}_k^H \left(\sum_{t=1}^{N_t-K_i} \zeta_{t,i}\mathbf{e}_{t,i}\right) s_i \\ &+ n_k, \end{aligned} \tag{97}$$

where $(a)$ is due to $\mathbf{h}_k^H \mathbf{e}_{t,I(k)} = 0,\ t = 1,\ldots N_t - K_{I(k)}$, and $\lim_{N_t\to\infty} \mathbf{h}_k^H \mathbf{h}_i / N_t = 0,\ i \neq k$. According to (97), the second term of (96) does not contribute to the desired signal, and only causes extra interference to the users in other groups.

Now let us construct another beamformer by setting $\zeta_{t,m} \equiv 0$ in (97) but keeping the transmit power unchanged, i.e.,

$$\mathbf{w}'_m = \sum\nolimits_{k\in\mathcal{K}_m} \eta_m \xi_k \mathbf{h}_k,\ \forall m \in \mathcal{M}, \tag{98}$$

where

$$\eta_m = \frac{\|\mathbf{w}_m\|_2}{\sqrt{\sum_{k\in\mathcal{K}_m} \|\xi_k \mathbf{h}_k\|_2^2}} \geq 1. \tag{99}$$

Then the received signal of user $k$ with (98) is given by

$$\begin{aligned} y_k &= \mathbf{h}_k^H \mathbf{w}'_{I(k)} s_{I(k)} + \sum\nolimits_{i\neq I(k), i\in\mathcal{M}} \mathbf{h}_k^H \mathbf{w}'_i s_i + n_k \\ &= \eta_{I(k)} \xi_k \mathbf{h}_k^H \mathbf{h}_k s_{I(k)} + n_k. \end{aligned} \tag{100}$$

Therefore, we must have $\gamma_k(\mathbf{W}) \leq \gamma_k(\mathbf{W}'), \forall k \in \mathcal{K}$, where $\mathbf{W}' = [\mathbf{w}'_1, \ldots, \mathbf{w}'_M]$. Finally, since $R(\gamma, \vartheta)$ is a monotonically increasing function in the effective SINR regime, the system can always achieve higher performance with (98) than that with (96) for problems $\mathbb{P}_1$ and $\mathbb{P}_2$, contradicting the optimality of (96). This completes the proof.

## Appendix F
## Proof of Theorem 5

For given $\vartheta > 0$, $R(\gamma, \vartheta)$ is monotonic increasing with $\gamma$ for $\gamma \geq \bar{\gamma}_2$. Then $\mathbb{P}_{12}$ can be equivalently rewritten as

$$\begin{aligned} \max_{\{\kappa_k | k\in\mathcal{K}_m\}}\ \min_k\ & g_k \kappa_k^2 \\ \text{s.t. }& (54b), (54c). \end{aligned} \tag{101}$$

If problem $\mathbb{P}_{12}$ is feasible, one can verify that

$$g_k \kappa_k^2 = g_{k'} \kappa_{k'}^2, \forall k \neq k',\ k, k' \in \mathcal{K}_m, \tag{102}$$

holds at the optimality. Then combining (54b), the optimal solution of problem $\mathbb{P}_{12}$ is given by

$$(\kappa_k^*)^2 = \frac{1}{\sum_{i\in\mathcal{K}_m} (\Gamma_k \sigma_i^2)/(\Gamma_i \sigma_k^2)},\ \forall k \in \mathcal{K}_m. \tag{103}$$

which leads to (55), as combining coefficients $\{\kappa_k\}_{k=1}^K$ are real positive. Substituting (103) into (54c), we have that problem $\mathbb{P}_{12}$ is feasible if and only if $\widehat{\Gamma}_m p_m \geq \hat{\gamma}_m$, where $\widehat{\Gamma}_m = E/\left(\sum_{i\in\mathcal{K}_m} \sigma_i^2/\Gamma_i\right)$. The optimal objective value of $\mathbb{P}_{12}$ is thus $R(\widehat{\Gamma}_m p_m, \vartheta_m)$, if the problem is feasible.

## Appendix G
## Proof of Theorem 7

The Lagrangian of problem $\mathbb{P}_{15}$ is given by

$$\begin{aligned} \mathcal{L}_\tau =& \sum\nolimits_{m=1}^M \omega_m R(\widehat{\Gamma}_m p_m, \vartheta_m) + \tau\left(1 - \sum\nolimits_{m=1}^M p_m\right) \\ &+ \sum\nolimits_{m=1}^M \varsigma_m (p_m - \hat{\gamma}_m/\widehat{\Gamma}_m), \end{aligned} \tag{104}$$

where $\tau$ and $\varsigma_m \geq 0, m \in \mathcal{M}$, are the dual variables for constraints (59b) and (56c), respectively. According to Lemma 2, problem $\mathbb{P}_{15}$ is convex when $\hat{\gamma}_m \geq \bar{\gamma}_{3,m}, \forall m \in \mathcal{M}$. Thus, its optimal solution can be obtained by solving the KKT conditions:

$$\frac{\partial \mathcal{L}_\tau}{\partial p_m} = \widehat{\Gamma}_m \omega_m \Delta(\widehat{\Gamma}_m p_m, \vartheta_m) - \tau + \varsigma_m = 0,\ \forall m, \tag{105}$$

$$\sum\nolimits_{m=1}^M p_m = 1, \tag{106}$$

$$\varsigma_m (p_m - \hat{\gamma}_m/\widehat{\Gamma}_m) = 0,\ m \in \mathcal{M}, \tag{107}$$

$$\varsigma_m \geq 0,\ m \in \mathcal{M}, \tag{108}$$

$$p_m \geq \hat{\gamma}_m/\widehat{\Gamma}_m,\ m \in \mathcal{M}, \tag{109}$$

where $\Delta(x, y)$ is defined in (64). Substituting (105) into (107) and (108), the above KKT conditions simplify into

$$\widehat{\Gamma}_m \omega_m \Delta(\widehat{\Gamma}_m p_m, \vartheta_m) - \tau \leq 0,\ m \in \mathcal{M}, \tag{110}$$

$$1 - \sum\nolimits_{m=1}^M p_m = 0, \tag{111}$$

$$p_m \geq \hat{\gamma}_m/\widehat{\Gamma}_m,\ m \in \mathcal{M}, \tag{112}$$

$$\left(\widehat{\Gamma}_m \omega_m \Delta(\widehat{\Gamma}_m p_m, \vartheta_m) - \tau\right)(p_m - \hat{\gamma}_m/\widehat{\Gamma}_m) = 0, \forall m. \tag{113}$$

Note that $\Delta(\gamma, \vartheta) \geq 0$ monotonically decreases with $\gamma \geq \bar{\gamma}_3$for given $\vartheta$ according to Lemma 2. Then, according to (113), if the optimal dual variable $\tau^* \leq \omega_m \widehat{\Gamma}_m \Delta(\hat{\gamma}_m, \vartheta_m)$, the optimal solution $p_m^* = \breve{p}_m$, where $\breve{p}_m \geq \hat{\gamma}_m/\widehat{\Gamma}_m$ is the unique solution satisfying $\widehat{\Gamma}_m \omega_m \Delta(\widehat{\Gamma}_m \breve{p}_m, \vartheta_m) = \tau^*$; otherwise, $p_m^* = \hat{\gamma}_m/\widehat{\Gamma}_m$. Moreover, since $\sum_{m=1}^M p_m^*$ is a monotonically decreasing function of $\tau^*$, the optimal dual variable $\tau^*$ is unique and can be obtained with bisection method. The proof is completed.

## References


[1] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos *et al.*, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1733–1757, 3rd Quart., 2018.

[2] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Mar. 2017.

[3] Y. Yang, P. Wang, C. Wang, and F. Liu, "An eMBMS based congestion control scheme in cellular-VANET heterogeneous networks," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst. (ITSC)*, Qingdao, China, Oct. 2014, pp. 1–5.

[4] S. Roger, D. Martin-Sacristan, D. Garcia-Roger, J. F. Monserrat *et al.*, "Low-latency layer-2-based multicast scheme for localized V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2962–2975, Aug. 2019.

[5] D. Striccoli, G. Piro, and G. Boggia, "Multicast and broadcast services over mobile networks: A survey on standardized approaches and scientific outcomes," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1020–1063, 2nd Quart., 2019.

[6] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni *et al.*, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] M. Dong and Q. Wang, "Multi-group multicast beamforming: Optimal structure and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 3738–3753, Jun. 2020.

[13] O. T. Demir and T. E. Tuncer, "Antenna selection and hybrid beamforming for simultaneous wireless information and power transfer in multi-group multicasting systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6948–6962, Oct. 2016.

[14] 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.

[15] W. Wu, C. Xiao, and X. Gao, "Multicell massive MIMO multicast transmission with finite-alphabet inputs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6747–6760, Jul. 2019.

[16] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen *et al.*, "Max-min fair transmit precoding for multi-group multicasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.

[17] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Multicast multigroup precoding and user scheduling for frame-based satellite communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4695–4707, Sep. 2015.

[18] M. Kaliszan, E. Pollakis, and S. Stanczak, "Multigroup multicast with application-layer coding: Beamforming for maximum weighted sum rate," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Paris, France, Apr. 2012, pp. 2270–2275.

[19] A. Z. Yalcin and M. Yuksel, "Precoder design for multi-group multicasting with a common message," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7302–7315, Oct. 2019.

[20] K.-X. Li, L. You, J. Wang, and X. Gao, "Physical layer multicasting in massive MIMO systems with statistical CSIT," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1651–1665, Feb. 2020.

[21] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[22] K. S. Kim, D. K. Kim, C.-B. Chae, S. Choi *et al.*, "Ultrareliable and low-latency communication techniques for tactile internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.

[23] H. Lee and Y.-C. Ko, "Physical layer enhancements for ultra-reliable low-latency communications in 5G new radio systems," *IEEE Commun. Standards Mag.*, vol. 5, no. 4, pp. 112–122, Dec. 2021.

[24] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, Jul. 2020.

[25] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regim," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[26] S. He, Z. An, J. Zhu, J. Zhang *et al.*, "Beamforming design for multiuser URLLC with finite blocklength transmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, Dec. 2021.

[27] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7184–7200, Nov. 2020.

[28] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah *et al.*, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[29] A. A. Nasir, H. D. Tuan, H. Q. Ngo, T. Q. Duong, and H. V. Poor, "Cell-free massive MIMO in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5861–5871, Sep. 2021.

[30] Y. Wang, V. W. S. Wong, and J. Wang, "Flexible rate-splitting multiple access with finite blocklength," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1398–1412, May 2023.

[31] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[32] C. Sun, C. She, C. Yang, T. Q. S. Quek *et al.*, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.

[33] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

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

[35] A. Martinez and A. G. I. Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Appl. Workshop (ITA)*, La Jolla, CA, USA, Feb. 2011, pp. 1–6.

[36] J. Ostman, A. Lancho, G. Durisi, and L. Sanguinetti, "URLLC with massive MIMO: Analysis and design at finite blocklength," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6387–6401, Oct. 2021.

[37] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on laplace integrals and their asymptotic approximations," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854–6883, Dec. 2016.

[38] B. Yin, J. Tang, and M. Wen, "Connectivity maximization in non-orthogonal network slicing enabled industrial internet-of-things with multiple services," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5642–5656, Aug. 2023.

[39] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Massive mu-MIMO downlink TDD systems with linear precoding and downlink pilots," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, Oct. 2013, pp. 293–298.

[40] J. Wang and D. P. Palomar, "Worst-case robust MIMO transmission with imperfect channel knowledge," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3086–3100, Aug. 2009.

[41] J. Wang, G. Scutari, and D. P. Palomar, "Robust MIMO cognitive radio via game theory," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1183–1201, Mar. 2011.

[42] J. Wang, M. Bengtsson, B. Ottersten, and D. P. Palomar, "Robust MIMO precoding for several classes of channel uncertainty," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3056–3070, Jun. 2013.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

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

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

[46] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899–3911, Jul. 2015.

[47] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[48] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[49] H. Cramér, *Random Variables and Probability Distributions*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

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

[51] APS Mosek, "The mosek optimization toolbox for matlab," 2019. [Online]. Available: http://www.mosek.com

[52] H. H. M. Tam, H. D. Tuan, and D. T. Ngo, "Successive convex quadratic programming for quality-of-service management in full-duplex mu-MIMO multicell networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2340–2353, Jun. 2016.