An Optimization Framework for General Rate Splitting for General Multicast

Lingzhi Zhao, Ying Cui, *Member, IEEE*, Sheng Yang, *Member, IEEE*, and Shlomo Shamai (Shitz), *Life Fellow, IEEE*

Abstract—Immersive video, such as virtual reality (VR) and multi-view videos, is growing in popularity. Its wireless streaming is an instance of general multicast, extending conventional unicast and multicast, whose effective design is still open. This paper investigates general rate splitting for general multicast. Specifically, we consider a multi-carrier single-cell wireless network where a multi-antenna base station (BS) communicates to multiple single-antenna users via general multicast. We consider linear beamforming at the BS and joint decoding at each user in the slow fading and fast fading scenarios. In the slow fading scenario, we consider the maximization of the weighted sum average rate, which is a challenging nonconvex stochastic problem with numerous variables. To reduce computational complexity, we decouple the original nonconvex stochastic problem into multiple nonconvex deterministic problems, one for each system channel state. Then, we propose an iterative algorithm for each deterministic problem to obtain a Karush-Kuhn-Tucker (KKT) point using the concave-convex procedure (CCCP). In the fast fading scenario, we consider the maximization of the weighted sum ergodic rate. This problem is more challenging than the one for the slow fading scenario, as it is not separable. First, we propose a stochastic iterative algorithm to obtain a KKT point using stochastic successive convex approximation (SSCA) and the exact penalty method. Then, we propose two low-complexity iterative algorithms to obtain feasible points with promising performance for two cases of channel distributions using approximation and CCCP. The proposed optimization framework generalizes the existing ones for rate splitting for various types of services. Finally, we numerically show substantial gains of the proposed solutions over existing schemes in both scenarios and reveal the design insights of general rate splitting for general multicast.

Index Terms—General multicast, general rate splitting, linear beamforming, joint decoding, optimization, concave-convex procedure (CCCP), stochastic successive convex approximation (SSCA).

I. INTRODUCTION

CONVENTIONAL mobile Internet services include (traditional) video, audio, web browsing, social networking, software downloading, etc. These services can be supported by unicast, unicast with a common message, single-group multicast, and multi-group multicast. Immersive video, such as 360 video (projection of virtual reality (VR) spherical video onto a rectangle) and multi-view video (a key technique in free-viewpoint television, naked-eye 3D and VR), is growing in popularity. It is predicted that the VR market will reach 87.97 billion USD by 2025 [2]. When watching a 360 video, the tiles in a user's current FoV plus a safe margin are usually transmitted to the user in case of FoV change. On the other hand, when watching a multi-view video, a user's current view and adjacent views are usually transmitted to the user in case of a view switch. In wireless streaming of a popular immersive video to multiple users simultaneously, multiple messages (e.g., tiles for 360 video and views for multi-view video) are transmitted to each user, and one message may be intended for multiple users [3], [4], as illustrated in Fig. 1. This emerging service plays an important role in online gaming, self-driving, and cloud meeting, etc. but cannot perfectly adapt to the conventional transmission schemes mentioned above. This motivates us to consider general multicast (also referred to as general connection [8], [9] and general groupcast [10]) where one message can be intended for any user. Clearly, general multicast includes conventional unicast, unicast with a common message, single-group multicast, and multi-group multicast as special cases.

References [3], [4], [5], and [6] are pioneer works for supporting wireless streaming of a 360 video [3], [5] and wireless streaming of a multi-view video [4], [6], which are instances of general multicast. In [3], [4], [5], and [6], Orthogonal Multiple Access (OMA), such as Time Division Multiple Access (TDMA) [3], [4] and Orthogonal Frequency Division Multiple Access (OFDMA) [5], [6], is adopted to convert general multicast to per resource block single-group multicast. While the OMA-based mechanisms are easy to implement, spatial multiplexing gain is not exploited. On the other hand, non-orthogonal transmission mechanisms achieve higher transmission efficiency but are also more challenging

Lingzhi Zhao is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Ying Cui was with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China and also with the Henan Joint International Research Laboratory of Intelligent Networking and Data Analysis, School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China. She is now with the Internet of Things (IoT) Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511400, China and also with the Department of Electronics and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: yingcui@ust.hk).

Sheng Yang is with the Laboratory of Signals and Systems, CentraleSupélec-CNRS-Université Paris-Sud, 91192 Gif-sur-Yvette, France.

Shlomo Shamai (Shitz) is with the Technion–Israel Institute of Technology, Haifa 3200003, Israel.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3205508.

Digital Object Identifier 10.1109/TWC.2022.3205508

(a) 360 video   (b) Multi-view video
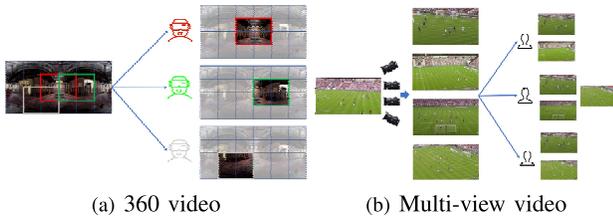
Fig. 1.   Applications of general multicast.

due to interference. Space Division Multiple Access (SDMA) and Non-Orthogonal Multiple Access (NOMA) are two solutions. The cost to suppress interference in SDMA can be high when the channels for some users are spatially aligned, while decoding interference in NOMA may not be possible when the interfering message rate is too high. Thus, SDMA and NOMA may have unsatisfactory performance. Rate splitting [7], [21] is introduced to partially suppress interference and partially decode interference to circumvent these limitations.

Rate splitting is originally proposed to effectively support unicast services [13]. Specifically, [13] investigates the simplest form of rate splitting for unicast, hereafter called 1-layer rate splitting, for the two-user interference channel. The idea of 1-layer rate splitting is as follows. First, each individual message is split into one private part and one common part, respectively. Then, the common parts of all the messages are re-assembled into one common message that is multicasted to all the users, and the private parts are unicasted to the corresponding receivers, respectively. In this way, part of the interference can be removed since it is decodable by design. Later, in [14], 1-layer rate splitting is applied to the multi-antenna broadcast channel (BC) and shown to provide a strict sum degree of freedom gain of a BC when only imperfect channel state information at the transmitter is available. In [15], [16], [17], [18], and [19], the authors investigate the precoder optimization of 1-layer rate splitting for unicast for Gaussian multiple-input multiple-output channels. In particular, the rates and beamforming vectors for common and private messages are optimized to maximize the sum rate [15], [19], worst-case rate [16], [17] or ergodic sum rate [18]. On the one hand, [20], [21] extend 1-layer rate splitting for unicast to general rate splitting for unicast and study the optimization of three-user unicast [20] and multi-user unicast [21], respectively, with linear precoding. On the other hand, [22], [23], [24], [25], [26] generalize 1-layer rate splitting for unicast to 1-layer rate splitting for unicast together with a multicast message intended for all users [22] and multi-group multicast [23], [24], [25], [26], respectively, and investigate the optimizations with linear beamforming. Note that most works [15], [16], [17], [19], [20], [21], [22], [24], [25], [26] focus on the slow fading scenario, whereas [18], [23] concentrates on the fast fading scenario.

Optimization-based random linear network coding designs for general multicast have been studied in [8] and [9] for wired networks. Besides, general rate splitting for general multicast has been studied in [10] for discrete memoryless broadcast channels. Here, we are interested in Gaussian fading channels and specifically the linear beamforming design from the optimization perspective. In general rate splitting for general multicast, each message intended for a user group is split into

sub-messages, one for each subset of users containing the user group. Then, the sub-messages intended for the same group of users are re-assembled and multicasted to the group. In this way, each user group decodes the desired message and part of the message of any other user group. Note that general rate splitting for general multicast produces more sub-messages and enables more flexible and effective interference reduction than brute-force application of general rate splitting for unicast [20], [21] to general multicast.[1] Besides, the optimizations of rate splitting for unicast and its slight generalization in [22], [23], [24], [25], and [26] cannot apply to general multicast. Therefore, for general multicast in both slow fading and fast fading scenarios, the optimization of general rate splitting with linear beamforming remains an open problem.

This paper intends to shed some light on the above issue. Specifically, we consider a multi-carrier single-cell wireless network, where a multi-antenna base station (BS) communicates to multiple single-antenna users via general multicast. Our main contributions are summarized below.

- We present general rate splitting for general multicast and illustrate its connection with conventional unicast, unicast with a common message, single-group multicast, and multi-group multicast. We adopt linear beamforming at the BS and joint decoding at each user and characterize the corresponding rate regions in the slow fading and fast fading scenarios.

- In the slow fading scenario, we optimize the transmission beamforming vectors and rates of sub-message units to maximize the weighted sum average rate. Note that the proposed problem formulation can reduce to those for general rate splitting for unicast [21], unicast with a common message [22], single-group multicast [27], and 1-layer rate splitting for multi-group multicast [24]. This problem is a challenging nonconvex stochastic problem with a large number of variables. To reduce computational complexity, we decouple the original nonconvex stochastic problem into multiple nonconvex deterministic problems, one for each system channel state. Then, for each nonconvex deterministic problem, we propose an iterative algorithm to obtain a Karush-Kuhn-Tucker (KKT) point using the concave-convex procedure (CCCP).

- In the fast fading scenario, we optimize the transmission beamforming vectors and rates of sub-message units to maximize the weighted sum ergodic rate and show that the problem formulation can reduce to the one for 1-layer rate splitting for unicast [18]. This problem is more challenging than the one for the slow fading scenario, as it is not separable. First, we propose a stochastic iterative algorithm to obtain a KKT point using stochastic successive convex approximation (SSCA) and the exact penalty method. Then, we propose two low-complexity iterative algorithms to obtain feasible points with promising performance for two cases of channel distributions, i.e., spatially correlated channel and independent and identically distributed (i.i.d.) channel, using approximation and

---

[1]One can brute-forcely apply general rate splitting [20], [21] for unicast to general multicast by treating a group of users who request the same message as a virtual user.

CCCP. It is noteworthy that the proposed optimization framework also provides general rate splitting designs for other service types in the fast fading scenario.
- We compare the complexities of the proposed solutions in the slow fading and fast fading scenarios. We also numerically demonstrate substantial gains of the proposed solutions over existing schemes in both scenarios and reveal the design insights of general rate splitting for general multicast.

Notation: We represent vectors by boldface lowercase letters (e.g., $\mathbf{x}$), matrices by boldface uppercase letters (e.g., $\mathbf{X}$), scalar constants by non-boldface letters (e.g., x), sets by calligraphic letters (e.g., $\mathcal{X}$), and sets of sets by boldface calligraphic letters (e.g., $\boldsymbol{\mathcal{X}}$). The notation x_i represents the i-th element of vector $\mathbf{x}$. The symbol $(\cdot)^H$ denotes complex conjugate transpose operator. $\|\cdot\|_2$ denotes the Euclidean norm of a vector. $\Re\{\cdot\}$ denotes the real part of a complex number. $\mathbb{E}[\cdot]$ denotes the statistical expectation. $\mathbf{I}_{M\times M}$ denotes the $M\times M$ identity matrix. $\mathbb{C}$ denotes the set of complex numbers.

II. SYSTEM MODEL

In this section, we first introduce general multicast in a single-cell wireless network and briefly illustrate its connection with unicast, unicast with a common message, single-group multicast, and multi-group multicast. Then, we present the physical layer model and general rate splitting with joint decoding for general multicast.

A. General Multicast

We consider a single-cell wireless network consisting of one BS and K users. Let $\mathcal{K} \triangleq \{1,\ldots,K\}$ denote the set of user indices. The BS has I independent messages. Let $\mathcal{I} \triangleq \{1,\ldots,I\}$ denote the set of the indices of I messages. We consider general multicast. Specifically, each user $k \in \mathcal{K}$ can request arbitrary I_k messages in $\mathcal{I}$, denoted by $\mathcal{I}_k \subseteq \mathcal{I}$, from the BS. We do not have any assumptions on $\mathcal{I}_k, \ k \in \mathcal{K}$ except that each message in $\mathcal{I}$ is requested by at least one user, i.e., $\cup_{k\in\mathcal{K}}\mathcal{I}_k = \mathcal{I}$.

To facilitate serving the K users, we partition the message set $\mathcal{I}$ according to the requests from the K users. For all $\mathcal{S} \subseteq \mathcal{K}, \mathcal{S} \neq \emptyset$, let

$$\mathcal{P}_\mathcal{S} \triangleq \left(\bigcap_{k\in\mathcal{S}}\mathcal{I}_k\right)\bigcap\left(\mathcal{I}-\bigcup_{k\in\mathcal{K}\backslash\mathcal{S}}\mathcal{I}_k\right) \quad (1)$$

denote the set of the indices of messages requested by each user in $\mathcal{S}$ and not requested by any user in $\mathcal{K}\backslash\mathcal{S}$ [3]. Define

$$\boldsymbol{\mathcal{P}} \triangleq \{\mathcal{P}_\mathcal{S}|\mathcal{P}_\mathcal{S} \neq \emptyset, \mathcal{S} \subseteq \mathcal{K}, \mathcal{S} \neq \emptyset\},$$
$$\boldsymbol{\mathcal{S}} \triangleq \{\mathcal{S}|\mathcal{P}_\mathcal{S} \neq \emptyset, \mathcal{S} \subseteq \mathcal{K}, \mathcal{S} \neq \emptyset\}.$$

Thus, $\boldsymbol{\mathcal{P}}$ forms a partition of $\mathcal{I}$ and $\boldsymbol{\mathcal{S}}$ specifies the user groups corresponding to the partition. We refer to each element in $\boldsymbol{\mathcal{P}}$ as a message unit.[2] We can see that different message units in $\boldsymbol{\mathcal{P}}$ are requested by different user groups in $\boldsymbol{\mathcal{S}}$.

Example 1 (Illustration of $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{S}}$ for Two-User Case): As illustrated in Fig. 2 (a), we consider $K = 2$, $I = 6$, $\mathcal{I}_1 = \{1,2,5,6\}$, $\mathcal{I}_2 = \{2,3,6,7\}$. Then, we have $\mathcal{P}_{\{1\}} = \{1,5\}$, $\mathcal{P}_{\{2\}} = \{3,7\}$, $\mathcal{P}_{\{1,2\}} = \{2,6\}$, $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_{\{1\}}, \mathcal{P}_{\{2\}}, \mathcal{P}_{\{1,2\}}\}$,
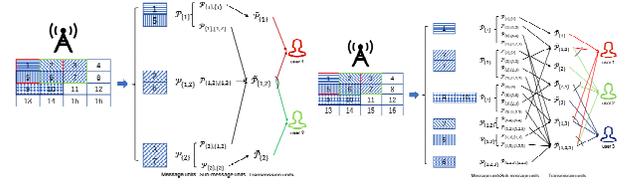
(a) $K = 2$, $I = 6$, $\mathcal{I}_1 = \{1,2,5,6\}$, $\mathcal{I}_2 = \{2,3,6,7\}$.

(b) $K = 3$, $I = 8$, $\mathcal{I}_1 = \{1,2,5,6\}$, $\mathcal{I}_2 = \{2,3,6,7\}$, $\mathcal{I}_3 = \{5,6,9,10\}$.

Fig. 2. Wireless streaming of a tiled 360 video to multiple users. The 360 video is divided into 4×4 tiles. The users have different FoVs which overlap to certain extent.

and $\boldsymbol{\mathcal{S}} = \{\{1\},\{2\},\{1,2\}\}$. There are 3 message units that are requested by 3 groups of users, respectively. For example, message unit $\mathcal{P}_{\{1\}}$ is requested only by user 1 and message unit $\mathcal{P}_{\{1,2\}}$ is requested by user 1 and user 2.[3]

Example 2 (Illustration of $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{S}}$ for Three-User Case): As illustrated in Fig. 2 (b), we consider $K = 3$, $I = 8$, $\mathcal{I}_1 = \{1,2,5,6\}$, $\mathcal{I}_2 = \{2,3,6,7\}$, $\mathcal{I}_3 = \{5,6,9,10\}$. Then, we have $\mathcal{P}_{\{1\}} = \{1\}$, $\mathcal{P}_{\{2\}} = \{3,7\}$, $\mathcal{P}_{\{3\}} = \{9,10\}$, $\mathcal{P}_{\{1,2\}} = \{2\}$, $\mathcal{P}_{\{1,3\}} = \{5\}$, $\mathcal{P}_{\{1,2,3\}} = \{6\}$, $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_{\{1\}}, \mathcal{P}_{\{2\}}, \mathcal{P}_{\{3\}}, \mathcal{P}_{\{1,2\}}, \mathcal{P}_{\{1,3\}}, \mathcal{P}_{\{1,2,3\}}\}$, and $\boldsymbol{\mathcal{S}} = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{1,2,3\}\}$. There are 6 message units that are requested by 6 groups of users, respectively. For example, message unit $\mathcal{P}_{\{1\}}$ is requested only by user 1, message unit $\mathcal{P}_{\{1,2\}}$ is requested by user 1 and user 2, and message unit $\mathcal{P}_{\{1,2,3\}}$ is requested by user 1, user 2, and user 3.

Remark 1 (Connection with Unicast and Multicast): The considered general multicast includes conventional unicast, unicast with a common message, single-group multicast, and multi-group multicast as special cases. (i) When $I = K, I_k = 1, k \in \mathcal{K}$, and $\mathcal{I}_k \neq \mathcal{I}_{k'}, k, k' \in \mathcal{K}, k \neq k'$, general multicast reduces to unicast [15], [16], [17], [18], [19], [20], [21]. In this case, $\boldsymbol{\mathcal{P}} = \{\{1\},\{2\},\ldots,\{K\}\}$ and $\boldsymbol{\mathcal{S}} = \{\{1\},\{2\},\ldots,\{K\}\}$. (ii) When $I = K + 1, I_k = 2, k \in \mathcal{K}, \mathcal{I}_k \neq \mathcal{I}_{k'}, k, k' \in \mathcal{K}, k \neq k'$, and $|\cap_{k\in\mathcal{K}}\mathcal{I}_k| = 1$, general multicast reduces to unicast with a common message [22]. In this case, $\boldsymbol{\mathcal{P}} = \{\{1\},\{2\},\ldots,\{K\},\mathcal{K}\}$ and $\boldsymbol{\mathcal{S}} = \{\{1\},\{2\},\ldots,\{K\},\mathcal{K}\}$. (iii) When $I = 1$, implying $I_k = 1, k \in \mathcal{K}$, and $\mathcal{I}_k = \mathcal{I}_{k'}, k, k' \in \mathcal{K}, k \neq k'$, general multicast becomes single-group multicast [27]. In this case, $\boldsymbol{\mathcal{P}} = \{\{1\}\}$ and $\boldsymbol{\mathcal{S}} = \{\mathcal{K}\}$. (iv) When $1 < I < K$ and $I_k = 1, k \in \mathcal{K}$, general multicast reduces to multi-group (I-group) multicast [23], [24], [25], [26]. In this case, $\boldsymbol{\mathcal{P}} = \{\{1\},\{2\},\ldots,\{I\}\}$ and $\boldsymbol{\mathcal{S}} = \{\{k \in \mathcal{K}|\mathcal{I}_k=\{1\}\}, \{k \in \mathcal{K}|\mathcal{I}_k=\{2\}\},\ldots,\{k \in \mathcal{K}|\mathcal{I}_k=\{I\}\}\}$. The general multicast considered in this paper, general connection in [8] and [9], and general groupcast considered in [10] mean the same.

B. General Rate Splitting

We consider rate splitting in the most general form for general multicast to serve the K users [10]. It allows each user group to decode not only the desired message unit $\mathcal{P}_\mathcal{S}$ but also part of the message unit of any other user group, $\mathcal{P}_{\mathcal{S}'}$ for all

---

[2]$\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{S}}$ are assumed to be given in [10].

[3]This general multicast scenario coincides with unicast with a common message.

$\mathcal{S}' \neq \mathcal{S}, \mathcal{S}' \in \boldsymbol{\mathcal{S}}$, to flexibly reduce the interference level. For all $\mathcal{S} \in \boldsymbol{\mathcal{S}}$, let $\boldsymbol{\mathcal{G}}_\mathcal{S} \triangleq \{\mathcal{X} | \mathcal{S} \subseteq \mathcal{X} \subseteq \mathcal{K}\}$. Namely, $\boldsymbol{\mathcal{G}}_\mathcal{S}$ collects all $2^{K-|\mathcal{S}|}$ subsets of $\mathcal{K}$ that contain $\mathcal{S}$. Define $\boldsymbol{\mathcal{G}} \triangleq \bigcup_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} \boldsymbol{\mathcal{G}}_\mathcal{S}$. Obviously, $\boldsymbol{\mathcal{S}} \subseteq \boldsymbol{\mathcal{G}}$. First, we split each message unit $\mathcal{P}_\mathcal{S}$ into $2^{K-|\mathcal{S}|}$ sub-message units. Accordingly, the rate of the message unit $\mathcal{P}_\mathcal{S}$, denoted by $R_\mathcal{S}$, is split into the rates of the $2^{K-|\mathcal{S}|}$ sub-message units, denoted by $R_{\mathcal{S},\mathcal{G}}, \mathcal{G} \in \boldsymbol{\mathcal{G}}_\mathcal{S}$[4] i.e.,

$$R_\mathcal{S} = \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}_\mathcal{S}} R_{\mathcal{S},\mathcal{G}}, \ \mathcal{S} \in \boldsymbol{\mathcal{S}}. \tag{2}$$

Let $\boldsymbol{\mathcal{S}}_\mathcal{G} \triangleq \{\mathcal{S} \in \boldsymbol{\mathcal{S}} | \mathcal{S} \subseteq \mathcal{G}\}$. Then, for all $\mathcal{G} \in \boldsymbol{\mathcal{G}}$, we re-assemble the sub-message units to form a transmission unit $\widetilde{\mathcal{P}}_\mathcal{G}$ with rate:

$$\widetilde{R}_\mathcal{G} = \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}_\mathcal{G}} R_{\mathcal{S},\mathcal{G}}, \ \mathcal{G} \in \boldsymbol{\mathcal{G}}. \tag{3}$$

That is, we first split $|\boldsymbol{\mathcal{S}}|$ message units, $\mathcal{P}_\mathcal{S}, \mathcal{S} \in \boldsymbol{\mathcal{S}}$, into $\sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} 2^{K-|\mathcal{S}|}$ sub-message units and then we re-assemble these sub-message units to form $|\boldsymbol{\mathcal{G}}|$ transmission units, $\widetilde{\mathcal{P}}_\mathcal{G}, \mathcal{G} \in \boldsymbol{\mathcal{G}}$.

Example 3 (Illustration of $\boldsymbol{\mathcal{G}}$ and General Rate Splitting for Two-User Case): For Example 1, we have $\boldsymbol{\mathcal{G}}_{\{1\}} = \{\{1\}, \{1,2\}\}, \boldsymbol{\mathcal{G}}_{\{2\}} = \{\{2\}, \{1,2\}\}, \boldsymbol{\mathcal{G}} = \{\{1\}, \{2\}, \{1,2\}\}$. As shown in Fig. 2 (a), we first split 3 message units into 5 sub-message units and then re-assemble the 5 sub-message units to form 3 transmission units.

Example 4 (Illustration of $\boldsymbol{\mathcal{G}}$ and General Rate Splitting for Three-User Case): For Example 2, we have $\boldsymbol{\mathcal{G}}_{\{1\}} = \{\{1\}, \{1,2\}, \{1,3\}, \{1,2,3\}\}$, $\boldsymbol{\mathcal{G}}_{\{2\}} = \{\{2\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$, $\boldsymbol{\mathcal{G}}_{\{3\}} = \{\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$, $\boldsymbol{\mathcal{G}}_{\{1,2\}} = \{\{1,2\}, \{1,2,3\}\}$, $\boldsymbol{\mathcal{G}}_{\{1,3\}} = \{\{1,3\}, \{1,2,3\}\}$, $\boldsymbol{\mathcal{G}} = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. As shown in Fig. 2 (b), we first split 6 message units into 17 sub-message units and then re-assemble the 17 sub-message units to form 7 transmission units.

Remark 2 (Connection with Rate Splitting for Unicast and Multicast): (i) When general multicast degrades to unicast, the proposed general rate splitting reduces to the general rate splitting for unicast proposed in our previous work [21], which extends the one-layer rate splitting for unicast [18]. (ii) When general multicast degrades to unicast with a common message, the proposed general rate splitting reduces to the one-layer rate splitting for unicast with a common message [22]. (iii) When general multicast degrades to single-group multicast, the proposed general rate splitting reduces to the conventional single-group multicast transmission [27]. (iv) When general multicast degrades to multi-group multicast, the proposed general rate splitting reduces to the one-layer rate splitting for multi-group multicast [24].

C. Physical Layer Model

The BS is equipped with M antennas, and each user has one antenna. We consider a multi-carrier system. Let N and $\mathcal{N} \triangleq \{1, 2, \ldots, N\}$ denote the number of subcarriers and the set of subcarrier indices, respectively. The bandwidth of each subcarrier is B (in Hz). We consider a discrete-time

system, i.e., time is divided into fixed-length slots. We adopt the block fading model, i.e., for each user and subcarrier, the channel remains constant within each slot and is independent and identically distributed (i.i.d.) over slots. Let $\vartheta_{k,n} \in \mathcal{H}$ denote the M-dimensional random channel vector for user k and subcarrier n, where $\mathcal{H} \subseteq \mathbb{C}^M$ denotes the M-dimensional channel state space. Let $\vartheta \triangleq (\vartheta_{k,n})_{k \in \mathcal{K}, n \in \mathcal{N}} \in \mathcal{H}^{KN}$ denote the random system channel state, where $\mathcal{H}^{KN}$ represents the system channel state space. Besides, let $\mathbf{h} \triangleq (\mathbf{h}_{k,n})_{k \in \mathcal{K}, n \in \mathcal{N}}$ denote a realization of ϑ in one slot, where $\mathbf{h}_{k,n}$ is a realization of $\vartheta_{k,n}$. Assume that user $k \in \mathcal{K}$ knows his channel state $\mathbf{h}_k \triangleq (\mathbf{h}_{k,n})_{n \in \mathcal{N}}$ in each slot. We consider two channel models which are presented below.

1) Slow Fading: We assume that the system channel state is known to the BS in each slot and adopt transmission rate adaptation over slots. We consider channel coding within each slot and across the N subcarriers. For the sake of exposition, we consider an arbitrary slot with system channel state $\mathbf{h}$ and any user group $\mathcal{S} \in \boldsymbol{\mathcal{S}}$. Let $\mathcal{P}_\mathcal{S}(\mathbf{h})$ denote the component of $\mathcal{P}_\mathcal{S}$ to be transmitted to user group $\mathcal{S}$ in the slot, also referred to as message unit, and $R_\mathcal{S}(\mathbf{h})$ denote the corresponding (transmission) rate of $\mathcal{P}_\mathcal{S}(\mathbf{h})$. Then, the average rate of message unit $\mathcal{P}_\mathcal{S}$ is given by $\mathbb{E}[R_\mathcal{S}(\vartheta)]$. We apply the general rate splitting scheme presented in Section II-B to transmit $\mathcal{P}_\mathcal{S}(\mathbf{h}), \mathcal{S} \in \boldsymbol{\mathcal{S}}$ to all user groups in $\boldsymbol{\mathcal{S}}$ in the slot. Specifically, for all $\mathcal{S} \in \boldsymbol{\mathcal{S}}$, message unit $\mathcal{P}_\mathcal{S}(\mathbf{h})$ is split into $2^{K-|\mathcal{S}|}$ sub-message units. The rate $R_\mathcal{S}(\mathbf{h})$ of $\mathcal{P}_\mathcal{S}(\mathbf{h})$ and the rates of the sub-message unit, denoted by $R_{\mathcal{S},\mathcal{G}}(\mathbf{h}), \mathcal{G} \in \boldsymbol{\mathcal{G}}_\mathcal{S}$, satisfy:

$$R_\mathcal{S}(\mathbf{h}) = \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}_\mathcal{S}} R_{\mathcal{S},\mathcal{G}}(\mathbf{h}), \ \mathcal{S} \in \boldsymbol{\mathcal{S}}, \ \mathbf{h} \in \mathcal{H}^{KN}. \tag{4}$$

Then, for all $\mathcal{G} \in \boldsymbol{\mathcal{G}}$, we re-assemble sub-message units to form transmission unit $\widetilde{\mathcal{P}}_\mathcal{G}(\mathbf{h})$ with rate:

$$\widetilde{R}_\mathcal{G}(\mathbf{h}) = \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}_\mathcal{G}} R_{\mathcal{S},\mathcal{G}}(\mathbf{h}), \ \mathcal{G} \in \boldsymbol{\mathcal{G}}, \ \mathbf{h} \in \mathcal{H}^{KN}. \tag{5}$$

For all $\mathcal{G} \in \boldsymbol{\mathcal{G}}$, transmission unit $\widetilde{\mathcal{P}}_\mathcal{G}(\mathbf{h})$ is encoded (channel coding) into codewords that span over the N subcarriers. Let $s_{\mathcal{G},n}(\mathbf{h}) \in \mathbb{C}$ denote a symbol for $\widetilde{\mathcal{P}}_\mathcal{G}(\mathbf{h})$ that is transmitted on the n-th subcarrier. For all $n \in \mathcal{N}$, let $\mathbf{s}_n(\mathbf{h}) \triangleq (s_{\mathcal{G},n}(\mathbf{h}))_{\mathcal{G} \in \boldsymbol{\mathcal{G}}}$, and assume that $\mathbb{E}[\mathbf{s}_n(\mathbf{h})\mathbf{s}_n^H(\mathbf{h})] = \mathbf{I}$. We consider linear beamforming. For all $n \in \mathcal{N}$, let $\mathbf{w}_{\mathcal{G},n}(\mathbf{h}) \in \mathbb{C}^{M \times 1}$ denote the beamforming vector for transmitting $\widetilde{\mathcal{P}}_\mathcal{G}(\mathbf{h})$ on subcarrier n in the slot. Using superposition coding, the transmitted signal on subcarrier n, denoted by $\mathbf{x}_n(\mathbf{h}) \in \mathbb{C}^{M \times 1}$, is given by:

$$\mathbf{x}_n(\mathbf{h}) = \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \mathbf{w}_{\mathcal{G},n}(\mathbf{h}) s_{\mathcal{G},n}(\mathbf{h}), \ n \in \mathcal{N}, \ \mathbf{h} \in \mathcal{H}^{KN}. \tag{6}$$

The transmission power on subcarrier $n \in \mathcal{N}$ is given by $\sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \|\mathbf{w}_{\mathcal{G},n}(\mathbf{h})\|_2^2$, and the total transmission power is given by $\sum_{n \in \mathcal{N}} \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \|\mathbf{w}_{\mathcal{G},n}(\mathbf{h})\|_2^2$. In the slow fading scenario, the total transmission power constraint is given by:

$$\sum_{n \in \mathcal{N}} \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \|\mathbf{w}_{\mathcal{G},n}(\mathbf{h})\|_2^2 \leq P, \ \mathbf{h} \in \mathcal{H}^{KN}. \tag{7}$$

Here, P denotes the transmission power budget. Define $\boldsymbol{\mathcal{G}}^{(k)} \triangleq \{\mathcal{G} \in \boldsymbol{\mathcal{G}} | k \in \mathcal{G}\}, k \in \mathcal{K}$. Then, the received signal at user $k \in \mathcal{K}$ on subcarrier $n \in \mathcal{N}$, denoted by $y_{k,n}(\mathbf{h}) \in \mathbb{C}$, is given by:

$$y_{k,n}(\mathbf{h}) = \mathbf{h}_{k,n}^H \mathbf{x}_n(\mathbf{h}) + z_{k,n}$$

---

[4]When $\mathcal{S} = \mathcal{K}$, $\boldsymbol{\mathcal{G}}_\mathcal{S} = \{\mathcal{S}\}$ and the message unit $\mathcal{P}_\mathcal{S}$ will not be split. For ease of exposition, we let $R_\mathcal{S} = R_{\mathcal{S},\mathcal{S}}$ for $\mathcal{S} = \mathcal{K}$.

$$= \mathbf{h}_{k,n}^H \sum\nolimits_{\mathcal{G} \in \mathcal{G}^{(k)}} \mathbf{w}_{\mathcal{G},n}(\mathbf{h}) s_{\mathcal{G},n}(\mathbf{h})$$
$$+ \mathbf{h}_{k,n}^H \sum\nolimits_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} \mathbf{w}_{\mathcal{G}',n}(\mathbf{h}) s_{\mathcal{G}',n}(\mathbf{h}) + z_{k,n},$$
$$k \in \mathcal{K}, \ n \in \mathcal{N}, \mathbf{h} \in \mathcal{H}^{KN}, \tag{8}$$

where the last equality is due to (6), and $z_{k,n} \sim \mathcal{CN}(0, \sigma^2)$ is the additive white gaussian noise (AWGN). In (8), the first term represents the desired signal, and the second represents the interference. It is noteworthy that the main idea of rate splitting is to make the undesired messages partially decodable in order to reduce interference [21]. To exploit the full potential of the general rate splitting for general multicast, we consider joint decoding at each user.[5] That is, each user $k \in \mathcal{K}$ jointly decodes the desired transmission units $\widetilde{\mathcal{P}}_{\mathcal{G}}(\mathbf{h}), \mathcal{G} \in \mathcal{G}^{(k)}$. Thus, in the slow fading scenario, the achievable rate region of the transmission units in the slot with system channel state $\mathbf{h}$ is described by the following constraints:

$$\sum_{\mathcal{G} \in \mathcal{X}} \widetilde{R}_{\mathcal{G}}(\mathbf{h})$$
$$\leq B \sum_{n \in \mathcal{N}} \log_2 \left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h})|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2} \right),$$
$$\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, \mathbf{h} \in \mathcal{H}^{KN}, \tag{9}$$

where $\widetilde{R}_{\mathcal{G}}(\mathbf{h})$ is given by (5).

Example 5 (Illustration of Physical Layer Design for Two-user Case in Slow Fading): For Example 1, the rates of message units $\mathcal{P}_{\{1\}}(\mathbf{h})$, $\mathcal{P}_{\{2\}}(\mathbf{h})$, and $\mathcal{P}_{\{1,2\}}(\mathbf{h})$ are $R_{\{1\}}(\mathbf{h})$, $R_{\{2\}}(\mathbf{h})$, and $R_{\{1,2\}}(\mathbf{h})$, respectively. The rates of transmission units $\tilde{\mathcal{P}}_{\{1\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2\}}(\mathbf{h})$, and $\tilde{\mathcal{P}}_{\{1,2\}}(\mathbf{h})$ are $\tilde{R}_{\{1\}}(\mathbf{h}) = R_{\{1\},\{1\}}(\mathbf{h})$, $\tilde{R}_{\{2\}}(\mathbf{h}) = R_{\{2\},\{2\}}(\mathbf{h})$, and $\tilde{R}_{\{1,2\}}(\mathbf{h}) = R_{\{1,2\}}(\mathbf{h}) + R_{\{1\},\{1,2\}}(\mathbf{h}) + R_{\{2\},\{1,2\}}(\mathbf{h})$, respectively. The beamforming vectors for transmitting $\tilde{\mathcal{P}}_{\{1\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2\}}(\mathbf{h})$, and $\tilde{\mathcal{P}}_{\{1,2\}}(\mathbf{h})$ on subcarrier $n \in \mathcal{N}$ are $\mathbf{w}_{\{1\},n}(\mathbf{h})$, $\mathbf{w}_{\{2\},n}(\mathbf{h})$, and $\mathbf{w}_{\{1,2\},n}(\mathbf{h})$, respectively.

Example 6 (Illustration of Physical Layer Design for Three-user Case in Slow Fading): For Example 2, the rates of message units $\mathcal{P}_{\{1\}}(\mathbf{h})$, $\mathcal{P}_{\{2\}}(\mathbf{h})$, $\mathcal{P}_{\{3\}}(\mathbf{h})$, $\mathcal{P}_{\{1,2\}}(\mathbf{h})$, $\mathcal{P}_{\{1,3\}}(\mathbf{h})$, and $\mathcal{P}_{\{1,2,3\}}(\mathbf{h})$ are $R_{\{1\}}(\mathbf{h})$, $R_{\{2\}}(\mathbf{h})$, $R_{\{3\}}(\mathbf{h})$, $R_{\{1,2\}}(\mathbf{h})$, $R_{\{1,3\}}(\mathbf{h})$, and $R_{\{1,2,3\}}(\mathbf{h})$, respectively. The rates of transmission units $\tilde{\mathcal{P}}_{\{1\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{3\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{1,2\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{1,3\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2,3\}}(\mathbf{h})$, and $\tilde{\mathcal{P}}_{\{1,2,3\}}(\mathbf{h})$ are $\tilde{R}_{\{1\}}(\mathbf{h}) = R_{\{1\},\{1\}}(\mathbf{h})$, $\tilde{R}_{\{2\}}(\mathbf{h}) = R_{\{2\},\{2\}}(\mathbf{h})$, $\tilde{R}_{\{3\}}(\mathbf{h}) = R_{\{3\},\{3\}}(\mathbf{h})$, $\tilde{R}_{\{1,2\}}(\mathbf{h}) = R_{\{1,2\},\{1,2\}}(\mathbf{h}) + R_{\{1\},\{1,2\}}(\mathbf{h}) + R_{\{2\},\{1,2\}}(\mathbf{h})$, $\tilde{R}_{\{1,3\}}(\mathbf{h}) = R_{\{1,3\},\{1,3\}}(\mathbf{h}) + R_{\{1\},\{1,3\}}(\mathbf{h}) + R_{\{3\},\{1,3\}}(\mathbf{h})$, $\tilde{R}_{\{2,3\}}(\mathbf{h}) = R_{\{2\},\{2,3\}}(\mathbf{h}) + R_{\{3\},\{2,3\}}(\mathbf{h})$, and $\tilde{R}_{\{1,2,3\}}(\mathbf{h}) = R_{\{1,2,3\}}(\mathbf{h}) + R_{\{1,3\},\{1,2,3\}}(\mathbf{h}) + R_{\{1,2\},\{1,2,3\}}(\mathbf{h}) + R_{\{1\},\{1,2,3\}}(\mathbf{h}) + R_{\{2\},\{1,2,3\}}(\mathbf{h}) + R_{\{3\},\{1,2,3\}}(\mathbf{h})$, respectively. The beamforming vectors for transmitting $\tilde{\mathcal{P}}_{\{1\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{3\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{1,2\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{1,3\}}(\mathbf{h})$, $\tilde{\mathcal{P}}_{\{2,3\}}(\mathbf{h})$, and $\tilde{\mathcal{P}}_{\{1,2,3\}}(\mathbf{h})$ on subcarrier $n \in \mathcal{N}$ are $\mathbf{w}_{\{1\},n}(\mathbf{h})$, $\mathbf{w}_{\{2\},n}(\mathbf{h})$, $\mathbf{w}_{\{3\},n}(\mathbf{h})$, $\mathbf{w}_{\{1,2\},n}(\mathbf{h})$, $\mathbf{w}_{\{1,3\},n}(\mathbf{h})$, $\mathbf{w}_{\{2,3\},n}(\mathbf{h})$, and $\mathbf{w}_{\{1,2,3\},n}(\mathbf{h})$, respectively.

[5]We can easily extend it to successive decoding as in [21].

2) Fast Fading: We assume that the system channel state is unknown to the BS in each slot, but its distribution is known to the BS. We consider channel coding over many slots and across the N subcarriers. We apply the general rate splitting scheme presented in Section II-B to transmit $\mathcal{P}_{\mathcal{S}}, \mathcal{S} \in \mathcal{S}$ to all user groups in $\mathcal{S}$ over many slots. For all $\mathcal{G} \in \mathcal{G}$, transmission unit $\widetilde{\mathcal{P}}_{\mathcal{G}}$ is encoded into codewords that span over the N subcarriers and many slots. Let $s_{\mathcal{G},n} \in \mathbb{C}$ denote a symbol for $\widetilde{\mathcal{P}}_{\mathcal{G}}$ that is transmitted on the n-th subcarrier. Let $\mathbf{w}_{\mathcal{G},n} \in \mathbb{C}^{M \times 1}$ denote the constant beamforming vector for transmitting $\widetilde{\mathcal{P}}_{\mathcal{G}}$ on subcarrier n over these slots. Similarly, using superposition coding, the transmitted signal on subcarrier n is given by:

$$\mathbf{x}_n = \sum\nolimits_{\mathcal{G} \in \mathcal{G}} \mathbf{w}_{\mathcal{G},n} s_{\mathcal{G},n}, \ n \in \mathcal{N}. \tag{10}$$

The total transmission power is given by $\sum_{n \in \mathcal{N}} \sum_{\mathcal{G} \in \mathcal{G}} \|\mathbf{w}_{\mathcal{G},n}\|_2^2$. In the fast fading scenario, the total transmission power constraint is given by:

$$\sum_{n \in \mathcal{N}} \sum_{\mathcal{G} \in \mathcal{G}} \|\mathbf{w}_{\mathcal{G},n}\|_2^2 \leq P. \tag{11}$$

Then, in a slot with system channel state $\mathbf{h}$, the received signal at user k on subcarrier n, denoted by $y_{k,n} \in \mathbb{C}$, is given by:

$$y_{k,n}$$
$$= \mathbf{h}_{k,n}^H \mathbf{x}_n + z_{k,n}$$
$$= \mathbf{h}_{k,n}^H \sum\nolimits_{\mathcal{G} \in \mathcal{G}^{(k)}} \mathbf{w}_{\mathcal{G},n} s_{\mathcal{G},n}$$
$$+ \mathbf{h}_{k,n}^H \sum\nolimits_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} \mathbf{w}_{\mathcal{G}',n} s_{\mathcal{G}',n} + z_{k,n}, \ k \in \mathcal{K}, \ n \in \mathcal{N},$$

where the last equality is due to (10). Similarly, we consider joint decoding at each user. The achievable ergodic rate region of the transmission units over these slots is given by:

$$\sum_{\mathcal{G} \in \mathcal{X}} \widetilde{R}_{\mathcal{G}}$$
$$\leq B \sum_{n \in \mathcal{N}} \mathbb{E} \left[\log_2 \left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}|^2} \right) \right],$$
$$\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, \tag{12}$$

where $\widetilde{R}_{\mathcal{G}}$ is given by (3).

Example 7 (Illustration of Physical Layer Design for Two-user Case in Fast Fading): For Example 1, the rates of message units $\mathcal{P}_{\{1\}}$, $\mathcal{P}_{\{2\}}$, and $\mathcal{P}_{\{1,2\}}$ are $R_{\{1\}}$, $R_{\{2\}}$, and $R_{\{1,2\}}$, respectively. The rates of transmission units $\tilde{\mathcal{P}}_{\{1\}}$, $\tilde{\mathcal{P}}_{\{2\}}$, and $\tilde{\mathcal{P}}_{\{1,2\}}$ are $\tilde{R}_{\{1\}} = R_{\{1\},\{1\}}$, $\tilde{R}_{\{2\}} = R_{\{2\},\{2\}}$, and $\tilde{R}_{\{1,2\}} = R_{\{1,2\}} + R_{\{1\},\{1,2\}} + R_{\{2\},\{1,2\}}$, respectively. The beamforming vectors for transmitting $\tilde{\mathcal{P}}_{\{1\}}$, $\tilde{\mathcal{P}}_{\{2\}}$, and $\tilde{\mathcal{P}}_{\{1,2\}}$ on subcarrier $n \in \mathcal{N}$ are $\mathbf{w}_{\{1\},n}$, $\mathbf{w}_{\{2\},n}$, and $\mathbf{w}_{\{1,2\},n}$, respectively.

Example 8 (Illustration of Physical Layer Design for Three-user Case in Fast Fading): For Example 2, the rates of message units $\mathcal{P}_{\{1\}}$, $\mathcal{P}_{\{2\}}$, $\mathcal{P}_{\{3\}}$, $\mathcal{P}_{\{1,2\}}$, $\mathcal{P}_{\{1,3\}}$, and $\mathcal{P}_{\{1,2,3\}}$ are $R_{\{1\}}$, $R_{\{2\}}$, $R_{\{3\}}$, $R_{\{1,2\}}$, $R_{\{1,3\}}$, and $R_{\{1,2,3\}}$, respectively. The rates of transmission units $\tilde{\mathcal{P}}_{\{1\}}$, $\tilde{\mathcal{P}}_{\{2\}}$, $\tilde{\mathcal{P}}_{\{3\}}$, $\tilde{\mathcal{P}}_{\{1,2\}}$, $\tilde{\mathcal{P}}_{\{1,3\}}$, $\tilde{\mathcal{P}}_{\{2,3\}}$, and $\tilde{\mathcal{P}}_{\{1,2,3\}}$ are $\tilde{R}_{\{1\}} = R_{\{1\},\{1\}}$, $\tilde{R}_{\{2\}} = R_{\{2\},\{2\}}$, $\tilde{R}_{\{3\}} = R_{\{3\},\{3\}}$, $\tilde{R}_{\{1,2\}} = R_{\{1,2\},\{1,2\}} + R_{\{1\},\{1,2\}} + R_{\{2\},\{1,2\}}$, $\tilde{R}_{\{1,3\}} = R_{\{1,3\},\{1,3\}} + R_{\{1\},\{1,3\}} + R_{\{3\},\{1,3\}}$, $\tilde{R}_{\{2,3\}} = R_{\{2\},\{2,3\}} + R_{\{3\},\{2,3\}}$, and $\tilde{R}_{\{1,2,3\}} =$
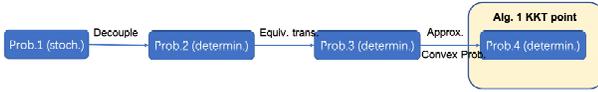
Fig. 3.   Illustration of the proposed solution framework for the slow fading scenario. "determin." and "stoch." are short for "deterministic" and "stochastic", respectively. "trans.", "Equiv.", "Approx.", and "Prob." are short for "transformation", "Equivalent", "Approximation", and "Problem", respectively.

$R_{\{1,2,3\}} + R_{\{1,3\},\{1,2,3\}} + R_{\{1,2\},\{1,2,3\}} + R_{\{1\},\{1,2,3\}} + R_{\{2\},\{1,2,3\}} + R_{\{3\},\{1,2,3\}}$, respectively. The beamforming vectors for transmitting $\tilde{\mathcal{P}}_{\{1\}}, \tilde{\mathcal{P}}_{\{2\}}, \tilde{\mathcal{P}}_{\{3\}}, \tilde{\mathcal{P}}_{\{1,2\}}, \tilde{\mathcal{P}}_{\{1,3\}}, \tilde{\mathcal{P}}_{\{2,3\}},$ and $\tilde{\mathcal{P}}_{\{1,2,3\}}$ on subcarrier $n \in \mathcal{N}$ are $\mathbf{w}_{\{1\},n}, \mathbf{w}_{\{2\},n}, \mathbf{w}_{\{3\},n}, \mathbf{w}_{\{1,2\},n}, \mathbf{w}_{\{1,3\},n}, \mathbf{w}_{\{2,3\},n},$ and $\mathbf{w}_{\{1,2,3\},n}$, respectively.

III. Average Rate Maximization in Slow Fading Scenario

This section considers the slow fading scenario and maximizes the weighted sum average rate under the achievable rate constraints and total transmission power constraints. The proposed solution framework for the slow fading scenario is illustrated in Fig. 3.

A. Optimization Problem Formulation

We would like to optimize the transmission beamforming vectors $\mathbf{w}(\mathbf{h}) \triangleq (\mathbf{w}_{\mathcal{G},n}(\mathbf{h}))_{\mathcal{G} \in \mathcal{G}, n \in \mathcal{N}}, \mathbf{h} \in \mathcal{H}^{KN}$ and rates of the sub-message units $\mathbf{R}(\mathbf{h}) \triangleq (R_{\mathcal{S},\mathcal{G}}(\mathbf{h}))_{\mathcal{S} \in \mathcal{S}, \mathcal{G} \in \mathcal{G}}, \mathbf{h} \in \mathcal{H}^{KN}$ to maximize the weighted sum average rate,[6] $\sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} \mathbb{E}[R_{\mathcal{S}}(\boldsymbol{\vartheta})]$, where the coefficient $\alpha_{\mathcal{S}} \geq 0$ denotes the weight for message unit $\mathcal{P}_{\mathcal{S}}$, subject to the total transmission power constraints in (7) and the achievable rate constraints in (9). Therefore, we formulate the following optimization problem.

Problem 1 (Weighted Sum Average Rate Maximization):

$$U^{(\text{slow})\star}$$
$$\triangleq \max_{(\mathbf{w}(\mathbf{h}))_{\mathbf{h} \in \mathcal{H}^{KN}}, (\mathbf{R}(\mathbf{h}))_{\mathbf{h} \in \mathcal{H}^{KN}} \succeq 0} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} \mathbb{E}[R_{\mathcal{S}}(\boldsymbol{\vartheta})]$$
$$\text{s.t.} \quad (7), (9),$$

where $R_{\mathcal{S}}(\cdot)$ is given by (4).

As the objective function and constraints of Problem 1 are seperable with $\mathbf{h} \in \mathcal{H}^{KN}$, Problem 1 is seperable and can be readily solved by solving the following problems for all $\mathbf{h} \in \mathcal{H}^{KN}$ in parallel.

Problem 2 (Weighted Sum Rate Maximization for $\mathbf{h}$):

$$\max_{\mathbf{w}(\mathbf{h}), \mathbf{R}(\mathbf{h}) \succeq 0} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} R_{\mathcal{S}}(\mathbf{h})$$
$$\text{s.t.} \quad \sum_{n \in \mathcal{N}} \sum_{\mathcal{G} \in \mathcal{G}} \|\mathbf{w}_{\mathcal{G},n}(\mathbf{h})\|_2^2 \leq P, \tag{13}$$

$$\sum_{\mathcal{G} \in \mathcal{X}} \widetilde{R}_{\mathcal{G}}(\mathbf{h})$$
$$\leq B \sum_{n \in \mathcal{N}} \log_2 \left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h})|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2} \right),$$
$$\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}. \tag{14}$$

[6] The proposed problem formulation and solution method can be readily extended to maximize the sum average rate and worst-case average rate as in [21].

Remark 3 (Connection with Rate Splitting for Unicast and Multicast): (i) When general multicast degrades to unicast, Problem 2 reduces to the weighted sum rate maximization problem for general rate splitting for unicast in [21]. (ii) When general multicast degrades to unicast with a common message, Problem 2 can be viewed as a generalization of the weighted sum rate maximization for unicast with a common message in [22]. (iii) When general multicast degrades to single-group multicast, Problem 2 reduces to the rate maximization problem for single-group multicast in [27]. (iv) When general multicast degrades to multi-group multicast, Problem 2 can be viewed as a generalization of the weighted sum rate maximization for multi-group multicast in [24].

Therefore, we can focus on solving Problem 2 for all $\mathbf{h}$. Note that the objective function is linear, the constraint in (13) is convex, and the constraints in (14) are nonconvex. Thus, Problem 2 is nonconvex.[7]

B. KKT Point

In this subsection, we propose an iterative algorithm to obtain a KKT point of Problem 2 using CCCP. First, we transform Problem 2 into the following equivalent problem by introducing auxiliary variables $\mathbf{e}(\mathbf{h}) \triangleq (e_{k,n,\mathcal{X}}(\mathbf{h}))_{\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}}$ and $\mathbf{u}(\mathbf{h}) \triangleq (u_{k,n,\mathcal{X}}(\mathbf{h}))_{\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}}$ and extra constraints (15)–(17), shown at the bottom of the next page.

Problem 3 (Equivalent DC Problem of Problem 2):

$$\max_{\mathbf{w}(\mathbf{h}), \mathbf{R}(\mathbf{h}) \succeq 0, \mathbf{e}(\mathbf{h}), \mathbf{u}(\mathbf{h})} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} \sum_{\mathcal{G} \in \mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}(\mathbf{h})$$
$$\text{s.t.} \quad (13), (15), (16), (17).$$

Let $(\mathbf{w}^\star(\mathbf{h}), \mathbf{R}^\star(\mathbf{h}), \mathbf{e}^\star(\mathbf{h}), \mathbf{u}^\star(\mathbf{h}))$ denote an optimal solution of Problem 3.

Lemma 1 (Equivalence Between Problem 2 and Problem 3): $(\mathbf{w}^\star(\mathbf{h}), \mathbf{R}^\star(\mathbf{h}), \mathbf{e}^\star(\mathbf{h}), \mathbf{u}^\star(\mathbf{h}))$ satisfies $2^{e_{k,n,\mathcal{X}}^\star(\mathbf{h})/B} = u_{k,n,\mathcal{X}}^\star(\mathbf{h}), \mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}$. Furthermore, Problem 2 and Problem 3 are equivalent.

Proof 1: First, by introducing auxiliary variables $\mathbf{e}(\mathbf{h})$ and $\mathbf{u}(\mathbf{h})$ and extra constraints:

$$2^{e_{k,n,\mathcal{X}}(\mathbf{h})/B} = u_{k,n,\mathcal{X}}(\mathbf{h}), \mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{19}$$

we can equivalently transform Problem 2 into the following problem:

$$\max_{\mathbf{w}(\mathbf{h}), \mathbf{R}(\mathbf{h}) \succeq 0, \mathbf{e}(\mathbf{h}), \mathbf{u}(\mathbf{h})} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} \sum_{\mathcal{G} \in \mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}(\mathbf{h})$$
$$\text{s.t.} \quad (13), (15), (17), (19)$$

Let $(\mathbf{w}^\ddagger(\mathbf{h}), \mathbf{R}^\ddagger(\mathbf{h}), \mathbf{e}^\ddagger(\mathbf{h}), \mathbf{u}^\ddagger(\mathbf{h}))$ denote an optimal solution. It is obvious that $(\mathbf{w}^\ddagger(\mathbf{h}), \mathbf{R}^\ddagger(\mathbf{h}))$ is an optimal solution of Problem 2. Next, we transform the above problem to Problem 3 by relaxing the constrains in (19) to the constraints in (16). By contradiction and the monotonicity of the objective function with respect to (w.r.t.) $\mathbf{R}$ in Problem 3,

[7] There are generally no effective methods for solving a nonconvex problem optimally. The goal of solving a nonconvex problem is usually to design an iterative algorithm to obtain a stationary point or a KKT point (which satisfies necessary conditions for optimality if strong duality holds).

we can show that the constraints in (16) are active at the optimal solution. Thus, $(\mathbf{w}^{\ddagger}(\mathbf{h}), \mathbf{R}^{\ddagger}(\mathbf{h}), \mathbf{e}^{\ddagger}(\mathbf{h}), \mathbf{u}^{\ddagger}(\mathbf{h}))$ is an optimal solution of Problem 3. Therefore, we can show Lemma 1. ∎

Based on Lemma 1, solving Problem 2 is equivalent to solving Problem 3. Problem 3 is a difference of convex functions (DC) programming (one type of nonconvex problems), and a KKT point can be obtained by CCCP [21].[8] The main idea of CCCP is to solve a sequence of successively refined approximate convex problems, each of which is obtained by linearizing the concave part and preserving the remaining convex part in the DC problem. Specifically, at the i-th iteration, the approximate convex problem of Problem 3 is given as follows. Let $(\mathbf{w}^{(i)}(\mathbf{h}), \mathbf{R}^{(i)}(\mathbf{h}), \mathbf{e}^{(i)}(\mathbf{h}), \mathbf{u}^{(i)}(\mathbf{h}))$ denote an optimal solution of the following problem.

Problem 4 (Approximation of Problem 3 at Iteration i):

$$\max_{\mathbf{w}(\mathbf{h}), \mathbf{R}(\mathbf{h}) \succeq 0, \mathbf{e}(\mathbf{h}), \mathbf{u}(\mathbf{h})} \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G} \in \mathcal{G}_{\mathcal{S}}} R_{\mathcal{S}, \mathcal{G}}(\mathbf{h})$$

$$\text{s.t.} \quad (13), (15), (16),$$

$$L_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}); \mathbf{w}_n^{(i-1)}(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h})) \leq 0,$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{20}$$

where $\mathbf{w}_n^{(i-1)}(\mathbf{h}) \triangleq (\mathbf{w}_{\mathcal{G},n}^{(i-1)}(\mathbf{h}))_{\mathcal{G} \in \boldsymbol{\mathcal{X}} \cup (\boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)})}$, $\mathbf{w}_n(\mathbf{h}) \triangleq (\mathbf{w}_{\mathcal{G},n}(\mathbf{h}))_{\mathcal{G} \in \boldsymbol{\mathcal{X}} \cup (\boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)})}$, and $L_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}); \mathbf{w}_n^{(i-1)}(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}))$ is given by (18), shown at the bottom of the page.

Problem 4 is convex and can be solved efficiently using standard convex optimization methods [29]. Problem 4 has $MN|\boldsymbol{\mathcal{G}}| + \sum_{S \in \boldsymbol{\mathcal{S}}} 2^{K-|\mathcal{S}|} + 2N \sum_{k \in \mathcal{K}} (2^{|\boldsymbol{\mathcal{G}}^{(k)}|} - 1)$ variables and $1 + (2N+1) \sum_{k \in \mathcal{K}} (2^{|\boldsymbol{\mathcal{G}}^{(k)}|} - 1)$ constraints. Thus, when an

interior point method is applied, the computational complexity for solving Problem 4 is $\mathcal{O}(K^{3.5} 2^{1.75 \times 2^K})$ as $K \to \infty$ [29]. The details of CCCP for obtaining a KKT point of Problem 3 are summarized in Algorithm 1.[9] As the number of iterations of Algorithm 1 does not scale with the problem size [30], the computational complexity for Algorithm 1 is the same as that for solving Problem 4, i.e., $\mathcal{O}(K^{3.5} 2^{1.75 \times 2^K})$ as $K \to \infty$.

Theorem 1 (Convergence of Algorithm 1): As $i \to \infty$, $(\mathbf{w}^{(i)}(\mathbf{h}), \mathbf{R}^{(i)}(\mathbf{h}), \mathbf{e}^{(i)}(\mathbf{h}), \mathbf{u}^{(i)}(\mathbf{h}))$ obtained by Algorithm 1 converges to a KKT point of Problem 3 [28].

Proof 2: The constraints in (13), (15), (16) are convex, and the constraint function in (17) can be regarded as a difference between two convex functions, i.e., $\sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2$ and $(\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h})|^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2)/u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h})$. Therefore, Problem 3 is a DC programming. Linearizing $(\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h})|^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2)/u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h})$ at $(\mathbf{w}_n^{(i-1)}(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}))$ and preserving $\sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2$ give $L_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}); \mathbf{w}_n^{(i-1)}(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}))$ in (18). Thus, Algorithm 1 implements CCCP. It has been validated in [28] that solving DC programming through CCCP always returns a KKT point. Therefore, we can show Theorem 1. ∎

This paper focuses mainly on exploiting the full potential of general rate splitting for general multicast. The computational complexity of Algorithm 1 can be formidable with a large K. We can use successive decoding to reduce the computational

---

[8]CCCP can exploit the partial convexity and usually converges faster to a KKT point than conventional gradient methods.

[9]In practice, we can run each of Algorithms 1-4 multiple times with randomly selected feasible initial points to obtain multiple KKT points and choose the KKT point with the best objective value.

$$\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}_{\mathcal{G}}} R_{\mathcal{S}, \mathcal{G}}(\mathbf{h}) = \sum_{n \in \mathcal{N}} e_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}), \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, \tag{15}$$

$$2^{e_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h})/B} \leq u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}), \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{16}$$

$$\sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2 - \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h})|^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2}{u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h})} \leq 0,$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}. \tag{17}$$

$$L_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}); \mathbf{w}_n^{(i-1)}(\mathbf{h}), u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}))$$

$$\triangleq \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h})|^2 + \sigma^2$$

$$+ \left(\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}^{(i-1)}(\mathbf{h})|^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}^{(i-1)}(\mathbf{h})|^2 + \sigma^2 \right) u_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{h}) \Big/ \left(u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}) \right)^2$$

$$- \left(2\Re \left\{ \sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^{(i-1)H}(\mathbf{h}) \mathbf{h}_{k,n} \mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G},n}(\mathbf{h}) + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \setminus \boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^{(i-1)H}(\mathbf{h}) \mathbf{h}_{k,n} \mathbf{h}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}(\mathbf{h}) \right\} + 2\sigma^2 \right) \Big/ u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}(\mathbf{h}),$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}. \tag{18}$$

Algorithm 1 Obtaining a KKT Point of Problem 3

1: Initialization: Choose any feasible point of Problem 3 $(\mathbf{w}^{(0)}(\mathbf{h}), \mathbf{R}^{(0)}(\mathbf{h}), \mathbf{e}^{(0)}(\mathbf{h}), \mathbf{u}^{(0)}(\mathbf{h}))$ and set $i = 0$.

2: **repeat**

3: Obtain an optimal solution $(\mathbf{w}^{(i)}(\mathbf{h}), \mathbf{R}^{(i)}(\mathbf{h}), \mathbf{e}^{(i)}(\mathbf{h}), \mathbf{u}^{(i)}(\mathbf{h}))$ of Problem 4 with an interior point method.

4: Set $i = i + 1$.

5: **until** the convergence criterion $\|(\mathbf{w}^{(i)}(\mathbf{h}), \mathbf{R}^{(i)}(\mathbf{h}), \mathbf{e}^{(i)}(\mathbf{h}), \mathbf{u}^{(i)}(\mathbf{h})) - (\mathbf{w}^{(i-1)}(\mathbf{h}), \mathbf{R}^{(i-1)}(\mathbf{h}), \mathbf{e}^{(i-1)}(\mathbf{h}), \mathbf{u}^{(i-1)}(\mathbf{h}))\|_2 \leq \epsilon$ is met.
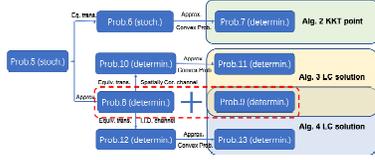
Fig. 4. Illustration of the proposed solution framework for the fast fading scenario. "Cor." and "LC" are short for "Correlated" and "Low Complexity", respectively. The interpretations of the remaining short forms can be found in Fig 3.

complexity to $\mathcal{O}(K^{1.5}|\mathcal{S}|^{1.5}2^{2K})$ as in [21].[10] We can also apply rate splitting with a reduced number of layers together with successive decoding to further reduce the computational complexity to $\mathcal{O}(K^{1.5}|\mathcal{S}|^{1.5}(|\mathcal{G}_{\text{lb}}|^2 + |\mathcal{G}_{\text{lb}}||\mathcal{S}|K + |\mathcal{S}|^2K^2))$ as in [21], for some $\mathcal{G}_{\text{lb}}$ satisfying $\mathcal{S} \subseteq \mathcal{G}_{\text{lb}} \subseteq \mathcal{G}$. Note that $|\mathcal{G}_{\text{lb}}|$ represents the reduced number of layers and satisfies $|\mathcal{S}| \leq |\mathcal{G}_{\text{lb}}| \leq 2^K - 1$.[11] Low-complexity optimization methods are beyond the scope of this paper.

IV. ERGODIC RATE MAXIMIZATION IN FAST FADING SCENARIO

This section considers the fast fading scenario and maximizes the weighted sum ergodic rate under the total transmission power constraint and achievable ergodic rate constraints. The proposed solution framework for the fast fading scenario is illustrated in Fig. 4. It is noteworthy that in the fast fading scenario when general multicast degrades to unicast, the proposed solutions based on general rate splitting outperform the one based on 1-layer rate splitting in [18] due to the fact that our solution can exploit the full potential of rate splitting and obtain a KKT point.

A. Optimization Problem Formulation

We would like to optimize the transmission beamforming vectors $\mathbf{w} \triangleq (\mathbf{w}_{\mathcal{G},n})_{\mathcal{G} \in \mathcal{G}, n \in \mathcal{N}}$ and rates of sub-message units $\mathbf{R} \triangleq (R_{\mathcal{S},\mathcal{G}})_{\mathcal{S} \in \mathcal{S}, \mathcal{G} \in \mathcal{G}}$ to maximize the weighted sum ergodic rate, $\sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} R_{\mathcal{S}}$, subject to the total transmission power constraint in (11) and the achievable ergodic rate constraints in (12). Therefore, we formulate the following problem.

Problem 5 (Weighted Sum Ergodic Rate Maximization):

$$U^{(\text{fast})\star} \triangleq \max_{\mathbf{w}, \mathbf{R} \succeq 0} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} R_{\mathcal{S}}$$
$$\text{s.t. } (11), (12),$$

where $R_{\mathcal{S}}$ is given by (2).

---

[10]Note that $|\mathcal{S}|$ may not scale with K, and how $|\mathcal{S}|$ scales with K relies on the user requests. In the case of $|\mathcal{S}| = \mathcal{O}(1)$, the reduced computational complexity is $\mathcal{O}(K^{1.5}2^{2K})$, as $K \to \infty$.

[11]In the case of $|\mathcal{S}| = \mathcal{O}(1)$ and $|\mathcal{G}_{\text{lb}}| = \mathcal{O}(1)$, the reduced computational complexity is $\mathcal{O}(N^{3.5}K^{3.5})$ as $K \to \infty$.

Remark 4 (Connection with Rate Splitting for Unicast and Multicast): When general multicast degrades to unicast and general rate splitting reduces to 1-layer rate splitting, Problem 5 can reduce to the weighted sum ergodic rate maximization problem for unicast in [18]. Besides, Problem 5 can be regards as a generalization the weighted sum ergodic rate maximization problem for other service types.

Note that the objective function is linear, the constraint in (11) is convex, and the constraints in (12) involve expectations and are nonconvex. Thus, Problem 5 is a more challenging nonconvex stochastic problem than the one in the slow fading scenario as it is not separable and has stochastic constraints.

B. KKT Point

In this subsection, we propose a stochastic iterative algorithm to obtain a KKT point of Problem 5 using SSCA together with the exact penalty method [32]. First, we equivalently transform Problem 5 to a stochastic optimization problem whose objective function is the weighted sum of the original objective function and the penalty for violating the nonconvex constraints with expectations in (12).

Problem 6 (Equivalent Problem of Problem 5):

$$\max_{\mathbf{w}, \mathbf{R} \succeq 0, \mathbf{s} \succeq 0} \sum_{\mathcal{S} \in \mathcal{S}} \alpha_{\mathcal{S}} \sum_{\mathcal{G} \in \mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}} - \rho \sum_{k \in \mathcal{K}} \sum_{\mathcal{X} \subseteq \mathcal{G}^{(k)}} s_{\mathcal{X},k}$$

$$\text{s.t. } (11), \quad \sum_{\mathcal{G} \in \mathcal{X}} \widetilde{R}_{\mathcal{G}}$$

$$-B \sum_{n \in \mathcal{N}} \mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}|^2}\right)\right]$$
$$\leq s_{\mathcal{X},k}, \quad \mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}, \qquad (22)$$

where $\mathbf{s} \triangleq (s_{\mathcal{X},k})_{\mathcal{X} \subseteq \mathcal{G}^{(k)}, k \in \mathcal{K}}$ are slack variables, and $\rho > 0$ is a penalty parameter that trades off the original objective function and the slack penalty term.

Next, at iteration i, we choose:

$$\overline{F}_{\mathcal{X},k,n}^{(i)}(\mathbf{w}) = (1 - \omega^{(i)})\overline{F}_{\mathcal{X},k,n}^{(i-1)}(\mathbf{w})$$
$$+ \omega^{(i)}\hat{f}_{\mathcal{X},k,n}(\mathbf{w}; \mathbf{w}^{(i-1)}; \mathbf{h}^{(i)}), \quad \mathcal{X} \subseteq \mathcal{G}^{(k)},$$
$$k \in \mathcal{K}, n \in \mathcal{N},$$

with $\overline{F}_{\mathcal{X},k,n}^{(0)}(\mathbf{w}) = 0$ as a convex surrogate function of $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}|^2}\right)\right]$ in (22), where $\mathbf{h}^{(i)}$ is the realization of the random system channel state at iteration i, $\omega^{(i)}$ is a stepsize satisfying $\omega^{(i)} > 0$, $\lim_{t \to \infty} \omega^{(i)} = 0, \sum_{i=1}^{\infty} \omega^{(i)} = \infty$, and $\hat{f}_{\mathcal{X},k,n}(\mathbf{w}; \mathbf{w}^{(i-1)}, ; \mathbf{h}^{(i)})$ is a convex approximation of $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \mathcal{X}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}|^2}\right)\right]$ around $\mathbf{w}^{(i-1)}$ which is given by (21), shown at the bottom of the next page. In (21), $\tau_{\mathcal{X},k,n} > 0$ can be any constant, and the term $\tau_{\mathcal{X},k,n}\|\mathbf{w} - \mathbf{w}^{(i-1)}\|^2$ is used to ensure strong convexity. Then, the approximation convex problem at iteration i is given as follows.

Problem 7 (Approximation of Problem 5 at i-th Iteration):

$$\max_{\mathbf{w}, \mathbf{R} \succeq 0, \mathbf{s} \succeq 0} \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G} \in \boldsymbol{\mathcal{G}}_{\mathcal{S}}} R_{\mathcal{S}, \mathcal{G}} + \rho \sum_{k \in \mathcal{K}} \sum_{\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}} s_{\boldsymbol{\mathcal{X}}, k}$$

$$\text{s.t. } (11),$$

$$\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} - B \sum_{n \in \mathcal{N}} \overline{F}^{(i)}_{\boldsymbol{\mathcal{X}}, k, n}(\mathbf{w}) \leq s_{\boldsymbol{\mathcal{X}}, k},$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}. \tag{23}$$

Let $(\overline{\mathbf{w}}^{(i)}, \mathbf{R}^{(i)}, \mathbf{s}^{(i)})$ denote an optimal solution of Problem 7.

It is noted that Problem 7 is a convex optimization problem which is always feasible (as $s_{\boldsymbol{\mathcal{X}}, k}$ can be arbitrarily large). We can obtain an optimal solution of Problem 7, $(\overline{\mathbf{w}}^{(i)}, \mathbf{R}^{(i)}, \mathbf{s}^{(i)})$, with the convex optimization techniques [29]. Given $\overline{\mathbf{w}}^{(i)}$, we then update $\mathbf{w}^{(i)}$ according to:

$$\mathbf{w}^{(i)} = (1 - \gamma^{(i)})\mathbf{w}^{(i-1)} + \gamma^{(i)}\overline{\mathbf{w}}^{(i)}, \tag{24}$$

where $\gamma^{(i)}$ is a step size satisfying:

$$\gamma^{(i)} > 0, \ \lim_{t \to \infty} \gamma^{(i)} = 0, \ \sum_{i=1}^{\infty} \gamma^{(i)} = \infty,$$

$$\sum_{t=1}^{\infty} \left(\gamma^{(i)} \right)^2 < \infty, \ \lim_{t \to \infty} \frac{\gamma^{(i)}}{\omega^{(i)}} = 0.$$

The detailed procedure is summarized in Algorithm 2. Note that Problem 7 has $MN|\boldsymbol{\mathcal{G}}| + \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} 2^{K-|\mathcal{S}|} + \sum_{k \in \mathcal{K}} (2^{|\boldsymbol{\mathcal{G}}^{(k)}|} - 1)$ variables and $1 + \sum_{k \in \mathcal{K}} (2^{|\boldsymbol{\mathcal{G}}^{(k)}|} - 1)$ constraints. Thus, when an interior point method is applied, the computational complexity for solving Problem 7 is $\mathcal{O}(K^{3.5} 2^{1.75 \times 2^K})$ as $K \to \infty$ [29]. Note that the number of iterations of Algorithm 2, T, is fixed, the computational complexity for Algorithm 2 is the same as that for solving Problem 7, i.e., $\mathcal{O}(K^{3.5} 2^{1.75 \times 2^K})$ as $K \to \infty$. As discussed in Section III-B, we can apply rate splitting with a reduced number of layers together with successive decoding to reduce the computational complexity to $\mathcal{O}(K^{1.5}|\boldsymbol{\mathcal{S}}|^{1.5}(|\boldsymbol{\mathcal{G}}_{\text{lb}}|^2 + |\boldsymbol{\mathcal{G}}_{\text{lb}}||\boldsymbol{\mathcal{S}}|K + |\boldsymbol{\mathcal{S}}|^2 K^2))$ as in [21].

Theorem 2 (Convergence of Algorithm 2): $\{(\mathbf{w}^{(i)}, \mathbf{R}^{(i)}, \mathbf{s}^{(i)})\}$ generated by Algorithm 2 has a limit point, denote by $(\mathbf{w}^{\dagger}, \mathbf{R}^{\dagger}, \mathbf{s}^{\dagger})$, and the following statements hold. (i) $(\mathbf{w}^{\dagger}, \mathbf{R}^{\dagger}, \mathbf{s}^{\dagger})$ is a KKT point of Problem 6; (ii) If $\mathbf{s}^{\dagger} = \mathbf{0}$, then $(\mathbf{w}^{\dagger}, \mathbf{R}^{\dagger})$ is a KKT point of Problem 5.

Algorithm 2 SSCA for Obtaining a KKT Point of Problem 6

1: Initialization: Choose any feasible point $\mathbf{w}^{(0)}$ of Problem 7 as the initial point and a sufficiently large $\rho > 0$.
2: **for** $i = 0, 1, \dots, T$ **do**
3:   Obtain an optimal solution $(\overline{\mathbf{w}}^{(i)}, \mathbf{R}^{(i)}, \mathbf{s}^{(i)})$ of Problem 7 with an interior point method, and update $\mathbf{w}^{(i)}$ according to (24).
4: **end for**
5: **Output**: $(\mathbf{w}^{(T)}, \mathbf{R}^{(T)}, \mathbf{s}^{(T)})$

Proof 3: Problem 5 and $\overline{F}^{(i)}_{\boldsymbol{\mathcal{X}}, k, n}(\mathbf{w})$ satisfy Assumption 1 and Assumption 2 in [32], respectively. Additionally, Algorithm 2 implements the stochastic iterative algorithm proposed in [32]. Therefore, by Theorem 2 of [32], we can show that statement (i) and statement (ii) hold. ∎

We can run Algorithm 2 multiple times, each with a random initial point $\mathbf{w}^{(0)}$ and a sufficiently large $\rho > 0$, until a KKT point $(\mathbf{w}^{\dagger}, \mathbf{R}^{\dagger}, \mathbf{s}^{\dagger})$ of Problem 6 with $\mathbf{s}^{\dagger} = \mathbf{0}$, i.e., a KKT point $(\mathbf{w}^{\dagger}, \mathbf{R}^{\dagger})$ of Problem 5, is obtained.

C. Low-Complexity Solutions

It is noteworthy that a stochastic iterative algorithm usually converges slowly and has low computation efficiency compared with its deterministic counterpart. Thus, in this subsection, we propose two low-complexity iterative algorithms to obtain feasible points of Problem 5 with promising performance for two cases of channel distributions using approximation and CCCP. By approximating $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\boldsymbol{\vartheta}^H_{k,n} \mathbf{w}_{\mathcal{G}, n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} |\boldsymbol{\vartheta}^H_{k,n} \mathbf{w}_{\mathcal{G}', n}|^2}\right)\right]$ in (12) with [31]:

$$\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \mathbb{E}\left[|\boldsymbol{\vartheta}^H_{k,n} \mathbf{w}_{\mathcal{G}, n}|^2\right]}{\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} \mathbb{E}\left[|\boldsymbol{\vartheta}^H_{k,n} \mathbf{w}_{\mathcal{G}', n}|^2\right]}\right)$$
$$= \log_2\left(1 + \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \mathbf{w}^H_{\mathcal{G}, n} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}, n}}{\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}^H_{\mathcal{G}', n} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}', n}}\right),$$

where $\mathbf{Q}_{k,n} \triangleq \mathbb{E}\left[\boldsymbol{\vartheta}_{k,n} \boldsymbol{\vartheta}^H_{k,n}\right] \succeq \mathbf{0}$ represents the channel covariance matrix for user k and subcarrier n, we approximate the stochastic nonconvex problem in Problem 5 with the following deterministic nonconvex problem.

$$\hat{f}_{\boldsymbol{\mathcal{X}}, k, n}(\mathbf{w}; \mathbf{w}^{(i-1)}, ; \mathbf{h}^{(i)})$$

$$\triangleq \log_2\left(1 + \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}^{(i)H}_{k,n} \mathbf{w}^{(i-1)}_{\mathcal{G}, n}|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}^{(i)H}_{k,n} \mathbf{w}^{(i-1)}_{\mathcal{G}', n}|^2}\right)$$

$$+ \frac{2\Re\left\{\sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}^{H(i-1)}_{\mathcal{G}', n} \mathbf{h}^{(i)}_{k,n} \mathbf{h}^{(i)H}_{k,n} (\mathbf{w}_{\mathcal{G}', n} - \mathbf{w}^{(i-1)}_{\mathcal{G}', n}) + \sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \mathbf{w}^{H(i-1)}_{\mathcal{G}, n} \mathbf{h}^{(i)}_{k,n} \mathbf{h}^{(i)H}_{k,n} (\mathbf{w}_{\mathcal{G}, n} - \mathbf{w}^{(i-1)}_{\mathcal{G}, n})\right\}}{\ln 2 (\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}^{(i)H}_{k,n} \mathbf{w}^{(i-1)}_{\mathcal{G}', n}|^2 + \sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\mathbf{h}^{(i)H}_{k,n} \mathbf{w}^{(i-1)}_{\mathcal{G}, n}|^2)}$$

$$- \frac{2\Re\left\{\sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}^{H(i-1)}_{\mathcal{G}', n} \mathbf{h}^{(i)}_{k,n} \mathbf{h}^{(i)H}_{k,n} (\mathbf{w}_{\mathcal{G}', n} - \mathbf{w}^{(i-1)}_{\mathcal{G}', n})\right\}}{\ln 2 (\sigma^2 + \sum_{\mathcal{G}' \in \boldsymbol{\mathcal{G}} \backslash \boldsymbol{\mathcal{G}}^{(k)}} |\mathbf{h}^{(i)H}_{k,n} \mathbf{w}^{(i-1)}_{\mathcal{G}', n}|^2)} + \tau_{\boldsymbol{\mathcal{X}}, k, n} \|\mathbf{w} - \mathbf{w}^{(i-1)}\|^2, \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, \ k \in \mathcal{K}, \ n \in \mathcal{N}. \tag{21}$$

Problem 8 (Approximate Problem of Problem 5 for Obtaining $\mathbf{w}^\dagger$*):*

$$\max_{\mathbf{w},\mathbf{R}\succeq 0} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \ (11), \sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}}$$

$$\leq B \sum_{n\in\mathcal{N}} \log_2\left(1 + \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^H \mathbf{Q}_{k,n}\mathbf{w}_{\mathcal{G},n}}{\sigma^2 + \sum_{\mathcal{G}'\in\mathcal{G}\setminus\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n}\mathbf{w}_{\mathcal{G}',n}}\right),$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}. \quad (25)$$

Let $\mathbf{w}^\dagger$ denote the beamforming vectors corresponding to a KKT point of Problem 8 (which may not be feasible for Problem 5 because of the adopted approximation). Then, construct the following deterministic linear programming parametrized by $\mathbf{w}^\dagger$.

Problem 9 (Approximate Problem of Problem 5 for Obtaining $\mathbf{R}^\dagger$*):*

$$\max_{\mathbf{R}\succeq 0} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}}$$

$$\leq B \sum_{n\in\mathcal{N}} \mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}^\dagger|^2}{\sigma^2 + \sum_{\mathcal{G}'\in\mathcal{G}\setminus\boldsymbol{\mathcal{G}}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}^\dagger|^2}\right)\right],$$

$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}. \quad (26)$$

Let $\mathbf{R}^\dagger$ denote an optimal solution of Problem 9.

Lemma 2 (Feasibility of $(\mathbf{w}^\dagger,\mathbf{R}^\dagger)$ *for Problem 5):* $(\mathbf{w}^\dagger,\mathbf{R}^\dagger)$ is a feasible point of Problem 5.

Proof 4: $\mathbf{w}^\dagger$ satisfies the constraint in (11). Besides, $(\mathbf{w}^\dagger,\mathbf{R}^\dagger)$ satisfies the constraints in (12). Thus, $(\mathbf{w}^\dagger,\mathbf{R}^\dagger)$ is a feasible point of Problem 5. ∎

As the approximation error is usually small, $(\mathbf{w}^\dagger,\mathbf{R}^\dagger)$ can be viewed as an approximate solution of Problem 5. Based on the numerical calculation of $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}^\dagger|^2}{\sigma^2 + \sum_{\mathcal{G}'\in\mathcal{G}\setminus\boldsymbol{\mathcal{G}}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}^\dagger|^2}\right)\right]$, $\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}$, we can easily solve Problem 9 using standard optimization methods. Problem 9 has $\sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} 2^{K-|\mathcal{S}|}$ variables and $\sum_{k\in\mathcal{K}} 2^{|\boldsymbol{\mathcal{G}}^{(k)}|} - K$ constraints. Thus, the computational complexity for solving Problem 9 using an interior point method is $\mathcal{O}(K^{1.5}2^{0.75\times 2^K})$. Moreover, it is expected that the deterministic nonconvex problem in Problem 8 can be more effectively solved than the stochastic nonconvex problem in Problem 5. In what follows, we concentrate on solving Problem 8 in two cases of channel distributions, namely spatially correlated channel ($\mathbf{Q}_{k,n}$ has non-zero non-diagonal elements, for all $k\in\mathcal{K}, n\in\mathcal{N}$) and independent and identically distributed (i.i.d.) channel ($\mathbf{Q}_{k,n}$ is a diagonal matrix with identical diagonal elements, for all $k\in\mathcal{K}, n\in\mathcal{N}$).

1) Spatially Correlated Channel: In this part, we consider the case of spatially correlated channel, i.e., $\mathbf{Q}_{k,n}$ has non-zero non-diagonal elements, for all $k\in\mathcal{K}, n\in\mathcal{N}$. First, using the same method as in Section III, Problem 8 can be transformed into the following DC programming by introducing auxiliary variables and extra constraints.

Problem 10 (Equivalent DC Problem of Problem 8):

$$\max_{\mathbf{w},\mathbf{R}\succeq 0,\mathbf{e},\mathbf{u}} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \ (11), (29),$$

$$\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}},\mathcal{S}\subseteq\mathcal{G}} R_{\mathcal{S},\mathcal{G}} \leq \sum_{n\in\mathcal{N}} e_{k,n,\boldsymbol{\mathcal{X}}}, \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}, \quad (27)$$

$$2^{\frac{e_{k,n,\boldsymbol{\mathcal{X}}}}{B}} \leq u_{k,n,\boldsymbol{\mathcal{X}}}, \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}, n\in\mathcal{N}, \quad (28)$$

where $\mathbf{e} \triangleq (e_{k,n,\boldsymbol{\mathcal{X}}})_{\boldsymbol{\mathcal{X}}\subseteq\boldsymbol{\mathcal{G}}^{(k)},k\in\mathcal{K},n\in\mathcal{N}}$, $\mathbf{u} \triangleq (u_{k,n,\boldsymbol{\mathcal{X}}})_{\boldsymbol{\mathcal{X}}\subseteq\boldsymbol{\mathcal{G}}^{(k)},k\in\mathcal{K},n\in\mathcal{N}}$ (with a slight abuse of notation), and (29) is shown at the bottom of the next page. Let $(\mathbf{w}^\star,\mathbf{R}^\star,\mathbf{e}^\star,\mathbf{u}^\star)$ denote an optimal solution of Problem 10.

Lemma 3 (Equivalence Between Problem 8 and Problem 10): $(\mathbf{w}^\star,\mathbf{R}^\star,\mathbf{e}^\star,\mathbf{u}^\star)$ satisfies $2^{\frac{e_{k,n,\boldsymbol{\mathcal{X}}}^\star}{B}} = u_{k,n,\boldsymbol{\mathcal{X}}}^\star$, $\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k\in\mathcal{K}, n\in\mathcal{N}$. Furthermore, Problem 8 and Problem 10 are equivalent.

Proof 5: The proof is similar to the one for Lemma 1 and is omitted here. ∎

Note that $\mathbf{e}$ and $\mathbf{u}$ are auxiliary variables, and (27), (28), and (29), shown at the bottom of the next page, are extra constraints. Analogously, the approximation convex problem of Problem 10 at iteration i is given below. Let $(\mathbf{w}^{(i)},\mathbf{R}^{(i)},\mathbf{e}^{(i)},\mathbf{u}^{(i)})$ denote an optimal solution of the following problem.

Problem 11 (Approximation of Problem 10 at Iteration i):

$$\max_{\mathbf{w},\mathbf{R}\succeq 0,\mathbf{e},\mathbf{u}} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\mathcal{G}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \ (11), (27), (28),$$

$$Q_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n, u_{k,n,\boldsymbol{\mathcal{X}}}; \mathbf{w}_n^{(i-1)}, u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}) \leq 0, \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)},$$

$$k\in\mathcal{K}, n\in\mathcal{N}, \quad (31)$$

where $\mathbf{w}_n \triangleq (\mathbf{w}_{\mathcal{G},n})_{\mathcal{G}\in\boldsymbol{\mathcal{X}}\cup(\mathcal{G}\setminus\boldsymbol{\mathcal{G}}^{(k)})}$, $\mathbf{w}_n^{(i-1)} \triangleq (\mathbf{w}_{\mathcal{G},n}^{(i-1)})_{\mathcal{G}\in\boldsymbol{\mathcal{X}}\cup(\mathcal{G}\setminus\boldsymbol{\mathcal{G}}^{(k)})}$, and $Q_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n, u_{k,n,\boldsymbol{\mathcal{X}}}; \mathbf{w}_n^{(i-1)}, u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)})$ is given by (30), shown at the bottom of the next page.

Similarly, Problem 11 is a convex problem and can be solved using standard convex optimization methods. Problem 11 has $MN|\mathcal{G}| + \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} 2^{K-|\mathcal{S}|} + 2N\sum_{k\in\mathcal{K}}(2^{|\boldsymbol{\mathcal{G}}^{(k)}|}-1)$ variables and $1+(2N+1)\sum_{k\in\mathcal{K}}(2^{|\boldsymbol{\mathcal{G}}^{(k)}|}-1)$ constraints. Thus, when an interior point method is applied, the computational complexity for solving Problem 11 is $\mathcal{O}(K^{3.5}2^{1.75\times 2^K})$ as $K\to\infty$ [29]. The details for obtaining a feasible point of Problem 5 are summarized in Algorithm 3. Similarly, as the number of iterations of Algorithm 1 does not scale with the problem size [30], we can conclude that the computational complexity of Steps 2-5 in Algorithm 3 is $\mathcal{O}(K^{3.5}2^{1.75\times 2^K})$ as $K\to\infty$. Recall that the computational complexity for solving Problem 9 is $\mathcal{O}(K^{1.5}2^{0.75\times 2^K})$. Thus, the computational complexity of Algorithm 3 is $\mathcal{O}(K^{3.5}2^{1.75\times 2^K})$ as $K\to\infty$. Similarly, we can reduce the computational complexity to $\mathcal{O}(K^{1.5}|\boldsymbol{\mathcal{S}}|^{1.5}(|\boldsymbol{\mathcal{G}}_{\text{lb}}|^2 + |\boldsymbol{\mathcal{G}}_{\text{lb}}||\boldsymbol{\mathcal{S}}|K + |\boldsymbol{\mathcal{S}}|^2 K^2))$ as in [21]. Analogously, we have the following convergence result.

Theorem 3 (Convergence of Algorithm 3): As $i\to\infty$, $(\mathbf{w}^{(i)},\mathbf{R}^{(i)},\mathbf{e}^{(i)},\mathbf{u}^{(i)})$ obtained by Steps 1-5 of Algorithm 3 converges to a KKT point of Problem 10 [28].

Algorithm 3 Obtaining a Feasible Point of Problem 5

1: Initialization: Choose any feasible point $(\mathbf{w}^{(0)}, \mathbf{R}^{(0)}, \mathbf{e}^{(0)}, \mathbf{u}^{(0)})$, and set $i = 0$.
2: **repeat**
3: Obtain an optimal solution $(\mathbf{w}^{(i)}, \mathbf{R}^{(i)}, \mathbf{e}^{(i)}, \mathbf{u}^{(i)})$ of Problem 11 with an interior point method.
4: Set $i = i + 1$.
5: **until** the convergence criterion $\|(\mathbf{w}^{(i)}, \mathbf{R}^{(i)}, \mathbf{e}^{(i)}, \mathbf{u}^{(i)}) - (\mathbf{w}^{(i-1)}, \mathbf{R}^{(i-1)}, \mathbf{e}^{(i-1)}, \mathbf{u}^{(i-1)})\|_2 \le \epsilon$ is met.
6: Set $\mathbf{w}^\dagger = \mathbf{w}^{(i)}$.
7: Calculate $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G},n}^\dagger|^2}{\sigma^2 + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} |\boldsymbol{\vartheta}_{k,n}^H \mathbf{w}_{\mathcal{G}',n}^\dagger|^2}\right)\right], \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}$ numerically and obtain an optimal solution $\mathbf{R}^\dagger$ of Problem 9 using standard optimization methods.

Proof 6: The constraints in (11), (27), (28) are convex, and the constraint function in (29) can be regarded as a difference between two convex functions, i.e., $\sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2$ and $(\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n} + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2)/u_{k,n,\boldsymbol{\mathcal{X}}}$. Therefore, Problem 10 is a DC programming. Similarly, linearizing $(\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n} + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2)/u_{k,n,\boldsymbol{\mathcal{X}}}$ at $(\mathbf{w}_n^{(i-1)}, u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)})$ and preserving $\sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2$ give $Q_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n, u_{k,n,\boldsymbol{\mathcal{X}}}; \mathbf{w}_n^{(i-1)}, u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)})$ in (30). Thus, Steps 1-5 of Algorithm 3 implements CCCP. It has been validated in [28] that solving DC programming through CCCP always returns a KKT point. Therefore, we can show Theorem 3. ∎

2) I.I.D. Channel: In this part, we consider i.i.d. channel, i.e., $\mathbf{Q}_{k,n} = \lambda \mathbf{I}_{M \times M}$ is a diagonal matrix with identical diagonal elements, for all $k \in \mathcal{K}, n \in \mathcal{N}$. By changing variables, i.e., letting $t_{\mathcal{G},n} \triangleq \|\mathbf{w}_{\mathcal{G},n}\|_2^2 \ge 0$, Problem 8 can be equivalently converted into the following problem.

Problem 12 (Equivalent Problem of Problem 8 for i.i.d. Channel):

$$\max_{\mathbf{t}\succeq 0, \mathbf{R}\succeq 0} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\boldsymbol{\mathcal{G}}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}}\sum_{\mathcal{G}\in\boldsymbol{\mathcal{G}}} t_{\mathcal{G},n} \le P, \tag{32}$$

$$\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} \le B \sum_{n\in\mathcal{N}} \log_2\left(1 + \frac{\lambda \sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} t_{\mathcal{G},n}}{\sigma^2 + \lambda \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} t_{\mathcal{G}',n}}\right),$$
$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, \tag{33}$$

where $\mathbf{t} \triangleq (t_{\mathcal{G},n})_{\mathcal{G}\in\boldsymbol{\mathcal{G}},n\in\mathcal{N}}$. Let $(\mathbf{t}^\star, \mathbf{R}^\star)$ denote an optimal solution of Problem 12.

Note that Problem 12 for i.i.d. channel is irrelevant to M, as we intend to optimize the time-invariant beamforming vectors according to channel statistics in the fast fading scenario. Problem 10 has $MN|\boldsymbol{\mathcal{G}}|$ complex optimization variables for beamforming (i.e., $\mathbf{w}$) whereas Problem 12 has $N|\boldsymbol{\mathcal{G}}|$ real optimization variables for beamforming (i.e., $\mathbf{t}$). Thus, it is expected that Problem 12 has lower computational complexity than Problem 10. Besides, we have the following result.

Lemma 4 (Equivalence Between Problem 8 and Problem 12): Any $(\mathbf{w}^\star, \mathbf{R}^\star)$ with $\|\mathbf{w}_{\mathcal{G},n}^\star\|_2^2 = t_{\mathcal{G},n}^\star$ is an optimal solution of Problem 8.

Proof 7: When $\mathbf{Q}_{k,n} = \lambda \mathbf{I}$, we can transform $\log_2\left(1 + \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n}}{\sigma^2 + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n}}\right)$ to $\log_2\left(1 + \frac{\lambda \sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} t_{\mathcal{G},n}}{\sigma^2 + \lambda \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} t_{\mathcal{G}',n}}\right)$ by letting $t_{\mathcal{G},n} \triangleq \|\mathbf{w}_{\mathcal{G},n}\|_2^2 \ge 0$. Thus, the constraints in (11) and (25) can be equivalently converted to the constraints in (32) and (33), respectively. Thus, Problem 12 is equal to Problem 8. ∎

Problem 12 is a DC programming, and we can obtain a KKT point using CCCP. Specifically, the approximation convex problem at iteration i is given below.

Problem 13 (Approximation of Problem 12 at Iteration i):

$$\max_{\mathbf{t}\succeq 0, \mathbf{R}\succeq 0} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}} \alpha_{\mathcal{S}} \sum_{\mathcal{G}\in\boldsymbol{\mathcal{G}}_{\mathcal{S}}} R_{\mathcal{S},\mathcal{G}}$$

$$\text{s.t.} \quad (32),$$

$$\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} - B \sum_{n\in\mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}\cup\boldsymbol{\mathcal{X}}} t_{\mathcal{G}',n}\right)$$
$$+ B \sum_{n\in\mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} t_{\mathcal{G}',n}^{(i-1)}\right)$$
$$+ \frac{B}{\ln 2} \sum_{n\in\mathcal{N}} \frac{\sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} (t_{\mathcal{G}',n} - t_{\mathcal{G}',n}^{(i-1)})}{(\sigma^2 + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} t_{\mathcal{G}',n}^{(i-1)})} \le 0,$$
$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}. \tag{34}$$

$$\sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2 - \frac{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n} + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2}{u_{k,n,\boldsymbol{\mathcal{X}}}}$$
$$\le 0, \ \boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{29}$$

$$Q_{k,n,\boldsymbol{\mathcal{X}}}(\mathbf{w}_n, u_{k,n,\boldsymbol{\mathcal{X}}}; \mathbf{w}_n^{(i-1)}, u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)})$$
$$\triangleq \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^H \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2$$
$$+ \left(\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^{(i-1)H} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n}^{(i-1)} + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^{(i-1)H} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n}^{(i-1)} + \sigma^2\right) u_{k,n,\boldsymbol{\mathcal{X}}} \Big/ \left(u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)}\right)^2$$
$$- \left(2\Re\left\{\sum_{\mathcal{G}\in\boldsymbol{\mathcal{X}}} \mathbf{w}_{\mathcal{G},n}^{(i-1)H} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G},n} + \sum_{\mathcal{G}'\in\boldsymbol{\mathcal{G}}\backslash\boldsymbol{\mathcal{G}}^{(k)}} \mathbf{w}_{\mathcal{G}',n}^{(i-1)H} \mathbf{Q}_{k,n} \mathbf{w}_{\mathcal{G}',n} + \sigma^2\right\} + 2\sigma^2\right) \Big/ u_{k,n,\boldsymbol{\mathcal{X}}}^{(i-1)},$$
$$\boldsymbol{\mathcal{X}} \subseteq \boldsymbol{\mathcal{G}}^{(k)}, k \in \mathcal{K}, n \in \mathcal{N}. \tag{30}$$

Algorithm 4 Obtaining a Feasible Point of Problem 5

1: Initialization: Choose any feasible point $(\mathbf{t}^{(0)}, \mathbf{R}^{(0)})$, and set $i = 0$.
2: **repeat**
3: Obtain an optimal solution $(\mathbf{t}^{(i)}, \mathbf{R}^{(i)})$ of Problem 13 with an interior point method.
4: Set $i = i + 1$.
5: **until** the convergence criterion $\|(\mathbf{t}^{(i)}, \mathbf{R}^{(i)}) - (\mathbf{t}^{(i-1)}, \mathbf{R}^{(i-1)})\|_2 \le \epsilon$ is met.
6: Set $\mathbf{w}^\dagger = \mathbf{w}^{(i)}$, where $\|\mathbf{w}_{\mathcal{G},n}^{(i)}\|_2^2 = t_{\mathcal{G},n}^{(i)}$.
7: Calculate $\mathbb{E}\left[\log_2\left(1 + \frac{\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} |\vartheta_{k,n}^H \mathbf{w}_{\mathcal{G},n}^\dagger|^2}{\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} |\vartheta_{k,n}^H \mathbf{w}_{\mathcal{G}',n}^\dagger|^2}\right)\right], \boldsymbol{\mathcal{X}} \subseteq$
$\mathcal{G}^{(k)}, k \in \mathcal{K}$ numerically and obtain an optimal solution $\mathbf{R}^\dagger$ of Problem 9 using standard optimization methods.

Let $(\mathbf{t}^{(i)}, \mathbf{R}^{(i)})$ denote an optimal solution of Problem 13, where $\mathbf{t}^{(i)} \triangleq (t_{\mathcal{G},n}^{(i)})_{\mathcal{G} \in \boldsymbol{\mathcal{X}} \cup (\mathcal{G} \setminus \mathcal{G}^{(k)}), n \in \mathcal{N}}$.

Problem 13 is a convex problem and can be solved using standard convex optimization methods. Note that Problem 13 has $N|\mathcal{G}| + \sum_{S \in \boldsymbol{\mathcal{S}}} 2^{K - |\mathcal{S}|}$ variables and $1 + \sum_{k \in \mathcal{K}} (2^{|\mathcal{G}^{(k)}|} - 1)$ constraints. Thus, when an interior point method is applied, the computational complexity for solving Problem 13 is $\mathcal{O}(K^{1.5} 2^{0.75 \times 2^K + 4K})$ as $K \to \infty$ [29]. The details for obtaining a feasible point of Problem 5 are summarized in Algorithm 4. Similarly, the computational complexity of Steps 2-5 in Algorithm 4 is $\mathcal{O}(K^{1.5} 2^{0.75 \times 2^K + 4K})$ as $K \to \infty$. Noting that the computational complexity for solving Problem 9 is $\mathcal{O}(K^{1.5} 2^{0.75 \times 2^K})$, the computational complexity of Algorithm 4 is $\mathcal{O}(K^{1.5} 2^{0.75 \times 2^K + 4K})$ as $K \to \infty$. Similarly, we can reduce the computational complexity to $\mathcal{O}(K^{1.5} |\boldsymbol{\mathcal{S}}|^{1.5} (|\mathcal{G}_{\text{lb}}|^2 + |\mathcal{G}_{\text{lb}}||\boldsymbol{\mathcal{S}}| + |\boldsymbol{\mathcal{S}}|^2))$ as in [21]. Analogously, we have the following convergence result.

Theorem 4 (Convergence of Algorithm 4): As $i \to \infty$, $(\mathbf{t}^{(i)}, \mathbf{R}^{(i)})$ obtained by Steps 1-5 of Algorithm 4 converges to a KKT point of Problem 13 [28].

Proof 8: The constraint in (32) is linear and each constraint function of (33), $\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} - B \sum_{n \in \mathcal{N}} \log_2\left(1 + \frac{\lambda \sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} t_{\mathcal{G},n}}{\sigma^2 + \lambda \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} t_{\mathcal{G}',n}}\right)$ can be regarded as a difference between two convex functions, $\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} - B \sum_{n \in \mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)} \cup \boldsymbol{\mathcal{X}}} t_{\mathcal{G}',n}\right)$ and $-B \sum_{n \in \mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} t_{\mathcal{G}',n}\right)$. Thus, Problem 12 is a DC programming. Linearizing $-B \sum_{n \in \mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)}} t_{\mathcal{G}',n}\right)$ at $\mathbf{t}^{(i)}$ and preserving $\sum_{\mathcal{G} \in \boldsymbol{\mathcal{X}}} \widetilde{R}_{\mathcal{G}} - B \sum_{n \in \mathcal{N}} \log_2\left(\sigma^2 + \sum_{\mathcal{G}' \in \mathcal{G} \setminus \mathcal{G}^{(k)} \cup \boldsymbol{\mathcal{X}}} t_{\mathcal{G}',n}\right)$ give the constraints in (34). Thus, Step 1-5 of Algorithm 4 implements CCCP. It has been validated in [28] that solving DC programming through CCCP always returns a KKT point. Therefore, we can show Theorem 4. ∎

V. COMPARISONS BETWEEN SLOW FADING SCENARIO AND FAST FADING SCENARIO

First, we compare the computational complexities for the weighted sum average rate maximization in the slow fading scenario and the weighted sum ergodic rate maximization in the fast fading scenario. Algorithm 1 (for obtaining a KKT point in the slow fading scenario based on CCCP) has higher computational complexity than Algorithm 2 (for obtaining a KKT point in the fast fading scenario based on SSCA). However, noting that SSCA converges more slowly than CCCP in general, Algorithm 2 has longer computational time to achieve satisfactory performance than Algorithm 1. Besides, Algorithm 3 (for obtaining a low complexity solution in the fast fading scenario with spatially correlated channel based on CCCP) has the same computational complexity as Algorithm 1. Last, Algorithm 4 (for obtaining a low complexity solution in the fast fading scenario with i.i.d. channel based on CCCP) has lower computational complexity than Algorithm 1.

Next, we compare the practical implementation complexities for the solutions in slow fading scenario and the fast fading scenario. In the slow fading scenario, the BS must obtain the instantaneous channel condition and adjust the beamforming vectors at each slot according to the instantaneous system channel state. By contrast, in the fast fading scenario, the BS requires only the knowledge of channel statistics and adopts constant beamforming vectors, relying on channel statistics over slots. Therefore, the practical implementation complexities in the slow fading scenario are obviously higher than the those in the fast fading scenario.

Finally, we compare the optimal values of the weighted sum average rate maximization in the slow fading scenario and weighted sum ergodic rate maximization in the fast fading scenario. As the beamforming vectors in the slow fading scenario are adjusted at each slot according to the instantaneous system channel state, it is expected that the optimal value in the slow fading scenario is larger than that in the fast fading scenario. Although the gain in the optimal value cannot be shown analytical, it will be numerically verified in Section VI.

VI. NUMERICAL RESULTS

In this section, we numerically evaluate the proposed solutions obtained by Algorithm 1 for the slow fading scenario, namely Slow-Prop-RS and Algorithm 2, Algorithm 3, and Algorithm 4 for the fast fading scenario, namely Fast-Prop-RS, Fast-Prop-COR-RS, and Fast-Prop-IID-RS, respectively. In the slow fading scenario, we consider three baseline schemes, namely Slow-1L-RS, Slow-NOMA, and Slow-OFDMA. Slow-1L-RS and Slow-NOMA (both with joint decoding for fair comparison) extend 1-layer rate splitting (where each message is split into one private part and only one common part) [16] and NOMA [34], both for unicast in one slot, to general multicast in the considered slow fading scenario. More specifically, Slow-1L-RS and Slow-NOMA implement Algorithm 1 to obtain KKT points of Problem 2 with $\mathcal{G}_{\mathcal{S}} = \{\mathcal{S}, \mathcal{K}\}, \mathcal{S} \in \boldsymbol{\mathcal{S}}$ and with $\mathcal{G}_{\mathcal{S}} = \{\mathcal{S}\}, \mathcal{S} \in \boldsymbol{\mathcal{S}}$, respectively.[12] Slow-OFDMA adopts OFDMA and considers the maximum ratio transmission (MRT) on each subcarrier and optimizes the subcarrier and power allocation [5]. In the fast fading scenario, we consider one baseline scheme, namely Fast-1L-RS.[13] Similarly, Fast-1L-RS (with joint decoding)

---

[12]Slow-NOMA adopts NOMA with joint decoding, which outperforms the commonly adopted NOMA with successive decoding [34]. Note that it is challenging to optimize the subsequent decoding order for NOMA in general multicast.

[13]As existing works on NOMA and OFDMA mainly concentrate on one single slot, we do not consider their extensions to the fast fading scenario, whose performance can be far from satisfactory.

(a) Number of antennas M.   (b) Power budget P.   (c) Number of user groups G.
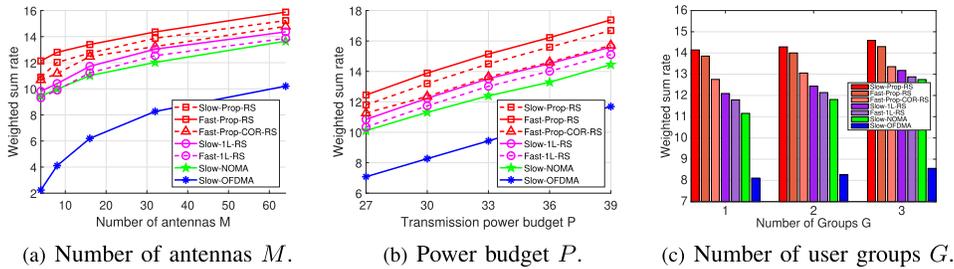
Fig. 5.   Weighted sum rate versus the system parameters in the case of spatially correlated channel.

(a) Slow-Prop-RS.   (b) Fast-Prop-RS.   (c) Fast-Prop-COR-RS.
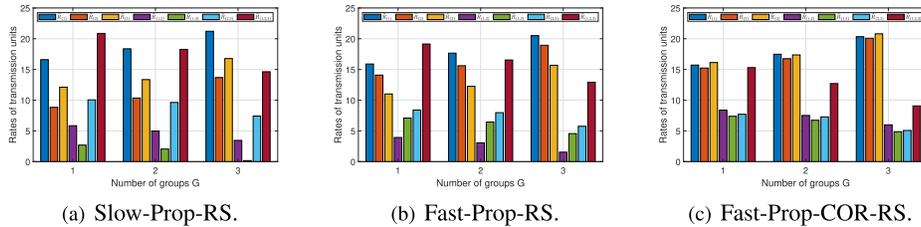
Fig. 6.   Rates of transmission units of the proposed solutions versus G in the case of spatially correlated channel.

extends 1-layer rate splitting for unicast in the fast fading scenario [18] to general multicast in the fast fading scenario. In particular, Fast-1L-RS implements Algorithm 2 to obtain a KKT point of Problem 5 with $\mathcal{G}_{\mathcal{S}} = \{\mathcal{S}, \mathcal{K}\}, \mathcal{S} \in \boldsymbol{\mathcal{S}}$. Note that for the proposed solutions, $\mathcal{G}_{\mathcal{S}} = \{\mathcal{X} | \mathcal{S} \subseteq \mathcal{X} \subseteq \mathcal{K}\}$. All algorithms adopt the same stopping criterion that the change in the objective function between two consecutive iterations is smaller than 0.1. In the slow fading scenario, we generate 100 random realizations of ϑ, solve the weighted sum rate maximization problem for each realization, and evaluate the weighted sum average rate of each scheme over the 100 random realizations. In the sequel, the weighted sum average rate in the slow fading scenario and the weighted sum ergodic rate in the fast fading scenario are both termed the weighted sum rate for ease of presentation.

In the simulation, we set $K = 3$, $I = 7$, $I_1 = \{1, 4, 5, 7\}$, $I_2 = \{2, 4, 6, 7\}$, and $I_3 = \{3, 5, 6, 7\}$. As a result, we have $\mathcal{P}_{\{1\}} = \{1\}$, $\mathcal{P}_{\{2\}} = \{2\}$, $\mathcal{P}_{\{3\}} = \{3\}$, $\mathcal{P}_{\{1,2\}} = \{4\}$, $\mathcal{P}_{\{1,3\}} = \{5\}$, $\mathcal{P}_{\{2,3\}} = \{6\}$, and $\mathcal{P}_{\{1,2,3\}} = \{7\}$. Additionally, we set $\alpha_{\mathcal{S}} = 1/7, \mathcal{S} \in \boldsymbol{\mathcal{S}}$, $B = 30$ kHz, $N = 72$, and $\sigma^2 = 10^{-9}$ W. We consider two cases of channel distributions, i.e., spatially correlated channel with the correlation following the one-ring scattering model as in [21] and i.i.d. channel with $\vartheta_{k,n} \sim \mathcal{CN}(0, \mathbf{I}_{M \times M}), k \in \mathcal{K}, n \in \mathcal{N}$. Note that the channel covariance matrix in the case of a spatially correlated channel is normalized for a fair comparison. When applying the one-ring scattering model, let G denote the number of user groups. We set the same angular spreads for the G groups and the same azimuth angle for the users in each group as in [21]. Note that G is related to the channel correlation among users. Specifically, the correlation increases as G decreases. When $G = 1$, users belong to one group and have the same channel covariance matrix. When $G = 3$, users are in different groups and have different channel covariance matrices.

First, we show the weighted sum rate of each scheme in the case of the spatially correlated channel. Fig. 5 illustrates the weighted sum rate versus the number of transmit antennas M, the total transmission power budget P, and the number of user

groups G in the one-ring scattering model, respectively, in the case of the spatially correlated channel. From Fig. 5, we have the following observations. Firstly, the weighted sum rate of each scheme increases with M, P, and G. Secondly, in the slow fading and fast fading scenarios, the proposed solutions outperform the corresponding baseline schemes. The gain of Slow-Prop-RS (Fast-Prop-RS or Fast-Prop-COR-RS) over Slow-1L-RS (Fast-1L-RS) is because each proposed solution unleashes the full potential of the flexibility of rate splitting. The gain of Slow-Prop-RS over Slow-NOMA arises because the cost for Slow-NOMA to suppress interference is high. In contrast, rate splitting together with joint decoding partially decodes interference and partially treats interference as noise. The gain of Slow-Prop-RS over Slow-OFDMA comes from an effective nonorthogonal transmission design. Thirdly, Slow-Prop-RS (Slow-1L-RS) outperforms Fast-Prop-RS (Fast-1L-RS). The gain arises from the fact that Slow-Prop-RS (Slow-1L-RS) optimizes the beamforming vectors at each slot according to the instantaneous system channel state, whereas Fast-Prop-RS (Fast-1L-RS) adopts the time-invariant beamforming vectors which are optimized according to the channel statistics. Additionally, Fig. 5 (c) shows that the gain of Slow-Prop-RS (Fast-Prop-RS or Fast-Prop-COR-RS) over Slow-1L-RS (Fast-1L-RS) and the gain of Slow-Prop-RS over Slow-NOMA increase as G decreases, demonstrating the advantage of flexibly dealing with interference in the presence of channel correlation among users. Fig. 6 shows the rates of the transmission units in each proposed solution versus the number of user groups G in the case of the spatially correlated channel. We can see that for each proposed solution, $\widetilde{R}_{\{1\}}, \widetilde{R}_{\{2\}}$, and $\widetilde{R}_{\{3\}}$ increase with G, whereas $\widetilde{R}_{\{1,2\}}, \widetilde{R}_{\{1,3\}}, \widetilde{R}_{\{2,3\}}$, and $\widetilde{R}_{\{1,2,3\}}$ decrease with G. This is because as channel correlation among the users decreases, it is efficient to decode less interference and treat more interference as noise.

Next, we show the weighted sum rate of each scheme in the case of the i.i.d. channel. Fig. 7 (a) and (b) plot the weighted sum rate versus the number of transmit antennas M and the

(a) Number of antennas M.   (b) Power budget P.
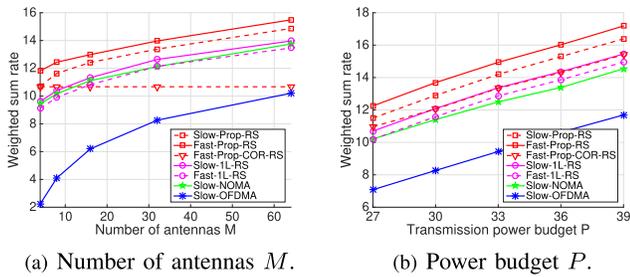
Fig. 7. Weighted sum rate versus the system parameters in the case of i.i.d. channel.
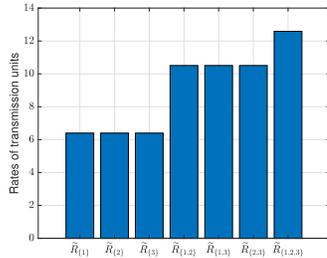
Fig. 8. Rates of transmission units of Fast-Prop-IID-RS.

total transmission power budget P in the case of the i.i.d. channel. We observe from these two figures that the weighted sum rate of Fast-Prop-IID-RS does not change with M. This is because Problem 12 is irrelevant to M. Besides, we have the same observations as from Fig. 5 (a) and (b). Fig. 8 shows the rates of the transmission units of Fast-Prop-IID-RS in the case of the i.i.d. channel. We can see from Fig. 8 that $\widetilde{R}_{\{1\}}, \widetilde{R}_{\{2\}}$, and $\widetilde{R}_{\{3\}}$ are identical, and $\widetilde{R}_{\{1,2\}}, \widetilde{R}_{\{1,3\}}$, and $\widetilde{R}_{\{2,3\}}$ are identical. The reasons are as follows. The general multicast setup is symmetric w.r.t. all users and their requesting messages, i.i.d. channel is considered, and time-invariant beamforming is adopted in the fast fading scenario.

VII. CONCLUSION

While applications such as content delivery are responsible for a large and increasing fraction of Internet traffic, general multicast communication will play a central role for future 6G and beyond networks. This paper investigated the optimization of general rate splitting for general multicast. We optimized the transmission beamforming vectors and rates of sub-message units to maximize the weighted sum average rate and the weighted sum ergodic rate in the slow fading and fast fading scenarios, respectively. We proposed iterative algorithms to obtain KKT points and low-complexity solutions in both scenarios using various optimization techniques. The proposed optimization framework generalizes the existing ones for rate splitting for unicast, unicast with a common message, single-group multicast, and multi-group multicast. Numerical results demonstrate notable gains of the proposed solutions over existing schemes and reveal the impact of channel correlation among users on the performance of general rate splitting for general multicast.

There are still some key aspects that we leave for future investigations. One direction is to go beyond linear approaches and investigate nonlinear precoders such as binning [10], [35].

Another interesting perspective is general multicast with partial channel state information at the transmitter side [21], [36].

REFERENCES

[1] L. Zhao, Y. Cui, S. Yang, S. Shamai, Y. Han, and Y. Zhang, "Rate splitting for general multicast," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 3936–3941.

[2] Mordor Intelligence. (Jan. 2020). *Virtual Reality (VR) Market—Growth, Trends, and Forecast (2020–2025)*. [Online]. Available: https://www.mordorintelligence.com/industry-reports/virtualrealitymarket/

[3] K. Long, Y. Cui, C. Ye, and Z. Liu, "Optimal wireless streaming of multi-quality 360 VR video by exploiting natural, relative smoothness-enabled, and transcoding-enabled multicast opportunities," *IEEE Trans. Multimedia*, vol. 23, pp. 3670–3683, 2021.

[4] W. Xu, Y. Cui, and Z. Liu, "Optimal multi-view video transmission in multiuser wireless networks by exploiting natural and view synthesis-enabled multicast opportunities," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1494–1507, Mar. 2020.

[5] C. Guo, L. Zhao, Y. Cui, Z. Liu, and D. Ng, "Power-efficient transmission of multi-quality tiled 360 VR video in MIMO-OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5408–5422, Aug. 2021.

[6] W. Xu, Y. Cui, Z. Liu, and H. Li, "Optimal multi-view video transmission in OFDMA systems," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 667–671, Mar. 2020.

[7] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," 2022, *arXiv:2201.03192*.

[8] Y. Cui, M. Médard, E. Yeh, D. Leith, F. Lai, and K. R. Duffy, "A linear network code construction for general integer connections based on the constraint satisfaction problem," *IEEE/ACM Trans. Netw.*, vol. 25, no. 6, pp. 3441–3454, Dec. 2017.

[9] Y. Cui, M. Medard, E. Yeh, D. Leith, and K. R. Duffy, "Optimization-based linear network coding for general connections of continuous flows," *IEEE/ACM Trans. Netw.*, vol. 26, no. 5, pp. 2033–2047, Oct. 2018.

[10] H. Romero and M. K. Varanasi, "Rate splitting, superposition coding and binning for groupcasting over the broadcast channel: A general framework," 2020, *arXiv:2011.04745*.

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

[12] Z. Li, S. Yang, and S. Shamai, "On linearly precoded rate splitting for Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4693–4709, Jul. 2021.

[13] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 1, pp. 49–60, Jan. 1981.

[14] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan. 2013.

[15] J. Park, J. Choi, N. Lee, W. Shin, and H. V. Poor, "Rate-splitting multiple access for downlink MIMO: A generalized power iteration approach," 2021, *arXiv:2108.06844*.

[16] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[17] Y. Mao, B. Clerckx, J. Zhang, V. O. K. Li, and M. A. Arafah, "Max-min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6362–6376, Oct. 2020.

[18] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[19] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5988–6002, Sep. 2021.

[20] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 133, May 2018.

[21] Z. Li, C. Ye, Y. Cui, S. Yang, and S. Shamai, "Rate splitting for multi-antenna downlink: Precoder design and practical implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1910–1924, Jun. 2020.

[22] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[23] L. Yin, B. Clerckx, and Y. Mao, "Rate-splitting multiple access for multi-antenna broadcast channels with statistical CSIT," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Mar. 2021, pp. 1–6.

[24] H. Joudeh and B. Clerckx, "Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7276–7289, Nov. 2017.

[25] A. Z. Yalcin, M. Yuksel, and B. Clerckx, "Rate splitting for multi-group multicasting with a common message," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12281–12285, Oct. 2020.

[26] H. Chen, D. Mi, B. Clerckx, Z. Chu, J. Shi, and P. Xiao, "Joint power and subcarrier allocation optimization for multigroup multicast systems with rate splitting," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2306–2310, Feb. 2020.

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

[28] 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.

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

[30] F. Facchinei, V. Kungurtsev, L. Lampariello, and G. Scutari, "Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity," *Math. Oper. Res.*, vol. 46, no. 2, pp. 595–627, Feb. 2021.

[31] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.

[32] C. Ye and Y. Cui, "Stochastic successive convex approximation for general stochastic optimization problems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 755–759, Jun. 2020.

[33] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[34] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[35] R. Tang, S. Xie, and Y. Wu, "On the achievable rate region of the K-receiver broadcast channels via exhaustive message splitting," *Entropy*, vol. 23, no. 11, p. 1408, Oct. 2021.

[36] Y. Mao and B. Clerckx, "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6775–6791, Nov. 2020.

[37] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, "Rate splitting multiple access in C-RAN: A scalable and robust design," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5727–5743, Sep. 2021.

Ying Cui (Member, IEEE) received the B.Eng. degree in electronic and information engineering from Xi'an Jiao Tong University, China, in 2007, and the Ph.D. degree from The Hong Kong University of Science and Technology, Hong Kong, SAR, China, in 2012. She held visiting positions at Yale University, USA, in 2011 and Macquarie University, Australia, in 2012. From June 2012 to December 2014, she was a Post-Doctoral Research Associate at Northeastern University, USA, and then at the Massachusetts Institute of Technology, USA. From January 2015 to July 2022, she was an Associate Professor with Shanghai Jiao Tong University, China. Since August 2022, she has been an Associate Professor with the Internet of Things (IoT) Thrust, The Hong Kong University of Science and Technology (Guangzhou), China, and an Affiliate Associate Professor with the Department of Electronics and Computer Engineering, The Hong Kong University of Science and Technology. Her current research interests include optimization, learning, the IoT communications, mobile edge caching and computing, and multimedia transmission. She was selected to the Thousand Talents Plan for Young Professionals of China in 2013. She received the Best Paper Award from IEEE ICC 2015 and IEEE GLOBECOM 2021. She serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Sheng Yang (Member, IEEE) received the B.E. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001, the Engineering and M.Sc. degrees in electrical engineering from Telecom ParisTech, Paris, France, in 2004, and the Ph.D. degree from Université de Pierre et Marie Curie (Paris VI) in 2007. From October 2007 to November 2008, he was with the Motorola Research Center, Gif-sur-Yvette, France, as a Senior Staff Research Engineer. Since December 2008, he has been with CentraleSupélec, Paris-Saclay University, where he is currently a Full Professor. Since April 2015, he also holds an honorary associate professorship with the Department of Electrical and Electronic Engineering, The University of Hong Kong (HKU). He received the 2015 IEEE ComSoc Young Researcher Award for the Europe, Middle East, and Africa Region (EMEA). He was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2015 to 2020. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Lingzhi Zhao received the B.E. degree in communication engineering from Shanghai University, China, in 2019, and the M.S. degree in communication engineering from Shanghai Jiao Tong University, China, in 2022. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Illinois Urbana–Champaign, USA. His research interests include multimedia transmission and wireless communication.

Shlomo Shamai (Shitz) (Life Fellow, IEEE) is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, where he is currently a Technion Distinguished Professor and the William Fondiller Chair of Telecommunications. His research interests encompasses a wide spectrum of topics in information theory and statistical communications. He is an URSI Fellow, a member of the Israeli Academy of Sciences and Humanities, and a Foreign Member of the U.S. National Academy of Engineering. He was a recipient of the 2011 Claude E. Shannon Award, the 2014 Rothschild Prize in Mathematics/Computer Sciences and Engineering, and the 2017 IEEE Richard W. Hamming Medal. He was a co-recipient of the 2018 Third Bell Labs Prize for Shaping the Future of Information and Communications Technology.

He was a recipient of numerous technical and paper awards and recognitions and is listed as a highly cited researcher (computer science) from 2013 to 2018. He served as an Associate Editor for the Special Issue on Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY, and has also served twice on the Board of Governors of the Information Theory Society. He has also served on the Executive Editorial Board for the IEEE TRANSACTIONS ON INFORMATION THEORY, the IEEE Information Theory Society Nominations and Appointments Committee, the IEEE Information Theory Society, and the Shannon Award Committee.