Fronthaul Compression and Precoding Optimization for NOMA-Based Joint Transmission of Broadcast and Unicast Services in C-RAN

Sungjun Ahn, Member, IEEE, Sung-Ik Park, Senior Member, IEEE, Jae-young Lee, Senior Member, IEEE, Namho Hur, Member, IEEE, Joonhyuk Kang, Member, IEEE

Abstract—This paper tackles a non-orthogonal multiplexing (NOMA)-based realization of broadcast-unicast overlay transmission in cloud-radio access networks (C-RANs). In particular, a fronthaul compression and precoding are jointly optimized to maximize the throughput in the considered network. The problem further optimizing the power allocation between broadcast and unicast signals is addressed as well. The algorithmic solutions are obtained by means of majorization minimization (MM)-type difference-of-convex (DC) programming. Moreover, the proposed solutions are further extended to cope with uncertainty of channel state information (CSI) to enable robust transmission. Computer simulations are performed to verify that the proposed scheme attains substantial achievable rate gain over the conventional decentralized precoding in entire parameter regions considered.

Index Terms—Non-orthogonal multiple access (NOMA), cloud-radio access network (C-RAN), fronthaul compression.

I. INTRODUCTION

MULTIMEDIA broadcast multicast service (MBMS) via broadband networks has been an attractive solution to lessen the cellular traffic for decades. Basically aimed at video streaming use cases, MBMS has been adopted in the 4th generation standard of 3rd Generation Partnership Project (3GPP), so-called Long-Term Evolution-Advanced (LTE-A)-Pro [1], [2]. Though MBMS over broadband is yet commercially successful, the broadcasting mode in LTE or further in the 5th Generation standard (5G) is foreseen promising [3]-[5], unlike to the previous premature foundations such as Digital Video Broadcasting-Handheld (DVB-H) and Media Forward Link Only (MediaFLO) [6]. Since the media consumption trend has been changed to streaming video contents in mobile/handheld devices, MBMS solution is expected to be further essential at managing data traffic. In fact, the increasing demand for point-to-multipoint (PTM) data distribution is not restricted to a video content delivery. Where many of device-to-device (D2D), machine-to-machine (M2M), and infrastructure-to-vehicle (X2V) services are based on the control/data message deliveries toward massive devices, broadcast/multicast transmission will be a key enabling technique for the emerging services in upcoming 5G era1 [7]-[10].

Meanwhile, one of the key features in the next generation communication networks is a centralized orchestration over multiple functional/architectural entities in the end-to-end system [11]-[13]. For 5G radio access network (RAN) in particular, several horizontal split options ranging from radio resource control (RRC)/packet data convergence protocol (PDCP) to common public radio interface (CPRI) splits have been suggested [14]. The physical layer split (option 7-a in [15]) in baseband level is widely known as cloud-RAN (C-RAN) in academia. In C-RANs, a cluster of densely deployed radio units, remote radio heads (RRHs) in other words, is controlled by a central unit (CU) through low-latency fronthaul links. This architecture allows a bundle of cells to associate aptly and thereby yields effective management on interference and traffic load [16]-[19]. In addition, a broadcast-broadband convergence, which has recently drawn significant attention [20], [21], can also be realized by such centralized orchestration [22]-[26]. Vigorously repromoted by the fully-Internet protocol (IP)-based development of Advanced Television Systems Committee (ATSC) 3.0 [27], the conjunction between terrestrial broadcasting and cloud-RAN has been addressed in various levels more over than a straightforward solution relying on multi-connectivity. One encouraging suggestion is a tower overlay, particularly at high power high towers (HPHTs) [28] where further expansions to the fully-integrated/hybrid networks can be considered as well.

Once broadcast and unicast services are decided to be co-located in the same network infrastructure, they can be multiplexed more efficiently on the top of non-orthogonal multiple access (NOMA) technology [29]-[31]. NOMA is one example of superposition coding embodiments, which combines multiple physical layer pipes (PLPs) within the same radio resource block while the signal powers are differently assigned for each PLP [32], [33]. In addition to the theoretic validations on superposition coding concept over decades [34], [35], the latest realization of non-orthogonal multiplexing (NOM), so-called a layered-division multiplexing (LDM), has been adopted in ATSC 3.0 standard as a mandatory technology [36]-[38]. A significant throughput gain of LDM over the tradi-

---

1e.g., datacast, automotive industry, intelligent transport system (ITS), connected car, and augmented reality (AR).
tion of orthogonal multiplexing schemes\textsuperscript{2} has been well-verified in practice via extensive demonstrations, simulations, and field trials\textsuperscript{3} [39]. NOMA approach is known especially suitable to broadcast-unicast combination scenarios since NOM particularly benefits unequal protection (UEP) configurations [40], [41]. The stochastic performance of NOMA-based broadcast-unicast combination has been analyzed in [29] within large-scale network viewpoint, but a specific design problem in system-scope under deterministic environment has yet been addressed.

This paper hence proposes a joint design of fronthaul compression and precoding for NOMA-based broadcast-unicast co-transmission, which particularly runs in C-RAN with finite capacity fronthaul links. More specifically, the compression noise levels among fronthaul links and the precoding matrices for each RRH are simultaneously optimized to maximize the achievable sum unicast-rate, while guaranteeing a predefined Co-transmission, which particularly runs in C-RAN with finite compression and precoding for NOMA-based broadcast-unicast combination has been addressed.

Two possible NOMA configurations are separately tackled: 1) broadcast at core layer (CL) and unicast at enhanced layer (EL)\textsuperscript{4} and 2) vice versa, while the former is mainly addressed due to the latter’s structural vulnerability on user privacy. The problems of interest belong to a class of difference-of-convex (DC) formulation, an objective form widely investigated in precoding optimization literature [42]. However, our approach differs from the typical precoding optimizations since the EL rate is piecewise differentiable, where the EL rate drops when the user fails at successive interference cancellation (SIC). We accordingly emphasize the impact of fronthaul capacity and target broadcast service rate on unicast rates from such discontinuity, which is realized by cross-layer interference (CLI). The proposed design is obtained by majorization-minimization (MM) programming, which is still feasible to find a stationary point for the piecewise differentiable objectives [43]. Furthermore, the proposed designs are additionally refined to attain robustness against the channel state information (CSI) uncertainty at the transmitter side.

The remainder of this paper is organized as follows. In Section II, the considered system model is illustrated. Section III and IV formulate the problems we aim at and provide the algorithmic solutions for them, under perfect and imperfect CSIT condition, respectively. The numerical verification of the proposed algorithms is drawn in Section V, and Section VI concludes the paper with some remarks.

\textbf{Notation:} Tr[A] and |A| denote a trace and determinant of a matrix A, respectively; log(·) stands for a logarithm operation with base 2; |A| for a set A indicates the cardinality of A; [·]\textsuperscript{T} and [·]\textsuperscript{H} denote a transpose and Hermitian operations, respectively, and 1(·) denotes an indicator function.

\section{System Model}

This paper considers a broadcast-unicast co-transmission based on NOMA in multiple-input multiple output (MIMO) C-RAN, which is depicted in Fig. 1. The C-RAN of interest consists of |N\textsubscript{T}| RRHs controlled by a CU and |N\textsubscript{R}| receivers (Rxs), where N\textsubscript{T} and N\textsubscript{R} denote the sets of RRHs and Rxs, respectively. It is assumed that the i\textsubscript{th} RRH is equipped with N\textsubscript{T},i transmit antennas where totally N\textsubscript{T} = \sum\textsubscript{i\in N\textsubscript{T}} N\textsubscript{T},i antennas are deployed at the transmitter side. On the other hand, N\textsubscript{R},j\in N\textsubscript{R} receive antennas are assumed to be mounted in the jth Rx, yielding total N\textsubscript{R} = \sum\textsubscript{j\in N\textsubscript{R}} N\textsubscript{R},j receive antennas to exist in the MIMO C-RAN. Each RRH modulates and transmits the broadcast and unicast data symbols delivered from the CU, having them multiplexed in NOMA-based manner. The RRHs can transmit broadcast datas within CL above the EL unics, and the opposite is also possible.

The rest of this paper will address those two strategies separately, and further compare them through simulations.

Within the dedicated fronthaul link, the encoded unicast data symbol s\textsubscript{i} = [s\textsubscript{T},i,1, \ldots, s\textsubscript{T},i,|N\textsubscript{T}|]\textsuperscript{T} precoded by matrix W\textsubscript{T},i\in N\textsubscript{T} is delivered to the ith RRH, while the broadcast data symbol s\textsubscript{b},i is delivered without precoding. s\textsubscript{b},i\in N\textsubscript{R} is the unicast data symbol for the jth Rx, and s\textsubscript{b},j and s\textsubscript{b},i\in N\textsubscript{R} are assumed to be independent and identically distributed (i.i.d.) zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables with unit variance. Note that the desired broadcast symbol is s\textsubscript{b} = [s\textsubscript{T},1, \ldots, s\textsubscript{T},|N\textsubscript{T}|]\textsuperscript{T}, which is constructed by gathering the s\textsubscript{b},i\in N\textsubscript{T} s\textsubscript{b} separately transmitted at each RRH. In short, the considered C-RAN operates as a distributed antenna system (DAS) transmitting s\textsubscript{b} for broadcast service. Besides, the fronthaul links are limited to have finite capacities, i.e., the capacity of the fronthaul link to the i\textsubscript{th} RRH is given \textbackslash C\textsubscript{i}. Accordingly, the broadcast stream delivery is assumed to be limited up to the predefined rate R\textsubscript{b} under the network operator’s agreement. Hence the data symbols are compressed before being forwarded to RRHs, in order to meet the capacity constraint of the corresponding fronthaul links. That is, the compressed version of unicast and broadcast signals

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig1.png}
\caption{Illustration of broadcast-unicast co-transmission based on MIMO C-RAN with |N\textsubscript{T}| RRHs and |N\textsubscript{R}| Rxs.}
\end{figure}
A. Broadcast in Core Layer and Unicast in Enhanced Layer (BCUE)

The case that broadcast and unicast streams are conveyed in NOMA CL and EL, respectively, is first addressed. Reasonably understood, this case does nothing with privacy issue where the unicast Rxs need to demodulate only the common broadcast signals, not the unicast signals oriented to the other users. The signal transmitted at the ith RRH is modeled as
\[ \mathbf{x}_{i} = \mathbf{W}_{i}^T \mathbf{s} + \mathbf{q}_{i}, \]
and \( \mathbf{x}_{b,i} = \mathbf{W}_{b,i}^T \mathbf{s} + \mathbf{q}_{b,i} \), respectively, are conveyed to the \( i \in N_{\text{cl}} \)th RRH, where the quantization noise \( \mathbf{q}_{b,i} \) and \( \mathbf{q}_{i} \) are i.i.d. ZMCSG random variables with covariance matrices \( \Omega_{\mathbf{q}_{b}}, i \in N_{\text{cl}} \) and \( I/N_{\text{cl}}, i \in N_{\text{el}} \), respectively. The portion of quantization noise in broadcast stream is determined by the ratio between \( \alpha_{b} \in (0, 1] \) and \( 1 - \alpha_{b} \), and \( \mathbf{w}_{b,i} \in N_{\text{cl}} \) indicates the broadcasting signal power at the ith RRH.

A channel from the \( i \in N_{\text{cl}} \)th RRH to the \( j \in N_{\text{el}} \)th Rx is modeled in matrix form as \( \beta_{i,j} \mathbf{H}_{j,i} \in \mathbb{C}^{N_{\text{cl}},N_{\text{el}}}, j \) where \( \beta_{i,j} \) is a propagation-loss coefficient and \( \mathbf{H}_{j,i} \) denotes a Rayleigh fading channel matrix which is assumed to have i.i.d. ZMCSG random entries with unit variance. For the sake of convenience, we comprehensively present the overall channel at the jth Rx as \( \mathbf{H}_{j} = [\beta_{j,1} \mathbf{H}_{1,j}, \cdots, \beta_{j,N_{\text{cl}}} \mathbf{H}_{N_{\text{cl}},j}]^T \).

**Fig. 2:** Block diagram of broadcast-unicast co-transmission system based on C-RAN with NOMA.

\[ \mathbf{x}_{i} = \mathbf{W}_{i}^T \mathbf{s} + \mathbf{q}_{i}, \]

\[ \mathbf{x}_{b,i} = \mathbf{W}_{b,i}^T \mathbf{s} + \mathbf{q}_{b,i}, \]

\[ \mathbf{x}_{i} = \mathbf{W}_{i}^T \mathbf{s} + \mathbf{q}_{i}, \]

\[ \mathbf{x}_{b,i} = \mathbf{W}_{b,i}^T \mathbf{s} + \mathbf{q}_{b,i}, \]

\[ \mathbf{x}_{i} = \mathbf{W}_{i}^T \mathbf{s} + \mathbf{q}_{i}, \]

\[ \mathbf{x}_{b,i} = \mathbf{W}_{b,i}^T \mathbf{s} + \mathbf{q}_{b,i}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

\[ \mathbf{y}_{j} = \sum_{k \in N_{\text{cl}} \setminus j} \mathbf{H}_{j} \mathbf{W}_{k}^T \mathbf{x}_{k} + \mathbf{n}_{j}, \]

where \( \mathbf{x}_{i} = [\mathbf{x}_{1}, \cdots, \mathbf{x}_{N_{\text{cl}}}]^T \), \( \mathbf{W}_{j} = [\mathbf{W}_{j1}, \cdots, \mathbf{W}_{jN_{\text{cl}}}]^T \), \( \mathbf{q}_{b,j} = [\mathbf{q}_{b,1,j}, \cdots, \mathbf{q}_{b,N_{\text{cl}},j}]^T \), and \( \mathbf{q}_{j} = [\mathbf{q}_{1,j}, \cdots, \mathbf{q}_{N_{\text{cl}},j}]^T \) are defined for notational simplicity. Moreover, \( \mathbf{W}_{j} \) denotes a precoding matrix applied to \( \mathbf{s}_{j} \), which is a submatrix of \( \mathbf{W} \): the \( (k-1)N_{\text{cl}} + 1 \)th column to \( kN_{\text{cl}} \)th column matrix of \( \mathbf{W} \) in precise. The difference of \( \mathbf{w}_{b,j} \) among RRs is considered in a diagonal matrix

\[ \mathbf{D} = \text{diag} \left( \mathbf{w}_{b,1} \mathbf{I}_{N_{\text{cl}}}, \cdots, \mathbf{w}_{b,N_{\text{cl}}} \mathbf{I}_{N_{\text{cl}}} \right), \]

where \( \text{diag}(\mathbf{a}) \) stands for a diagonal matrix whose diagonal entries are the elements of the vector \( \mathbf{a} \). In addition, \( \mathbf{n}_{j} \) denotes an additive white Gaussian noise (AWGN) at the jth Rx, which is a ZMCSG random variable whose covariance matrix is given by \( \mathbf{D} \).

B. Unicast in Core Layer and Broadcast in Enhanced Layer (UCBE)

Despite it has a demerit requiring demodulation of other users’ signals to extract broadcast messages, UCBE multiplexing can reduce a computation required at unicast-only users. Within UCBE networks, unicast users only have to decode CL signal, and hence do not need to perform SIC procedure.

The signal model for UCBE does not differ from (1) and (2) in Section II-A. However, in this case, the unicast signals are transmitted \( 1/\delta \) times stronger than broadcast signals, i.e.,

\[ \mathbf{w}_{b,j} = \delta \mathbf{w}_{b,j} \mathbf{I}_{N_{\text{cl}}} \mathbf{W}_{j} + \mathbf{n}_{j}, \]

Consequently the unicast users suffer from inter-layer interference (ILI) due to EL broadcast signal.

III. Fronthaul Compression and Precoding Optimization with Perfect CSIT

This section jointly optimizes \( \Omega_{\mathbf{u}} \) and \( \mathbf{W} \) in order to maximize the achievable sum-unicast rate. Perfect CSI at the transmitter (CSIT) is assumed while the imperfect CSIT case will be addressed in the next section. In addition, the solutions for BCUE and UCBE configurations are both provided, while those configurations are addressed separately.

A. Broadcast in Core Layer and Unicast in Enhanced Layer (BCUE)

As described in Fig. 2, receiving NOMA signals proceed in a hierarchical manner. Since SIC should be performed in advance to decode EL signal, the effective model of EL signal received at the jth Rx can be written by

\[ \mathbf{y}_{j}^{\text{EL}} = \mathbf{H}_{j}^T \mathbf{W}_{j}^T \mathbf{s}_{j} + \mathbf{n}_{j} \]

where the comprehensive notations \( \mathbf{y}_{j}^{\text{EL}} = [\mathbf{y}_{j1}, \cdots, \mathbf{y}_{jN_{\text{cl}}}]^T \), \( \mathbf{W}_{j} = [\mathbf{W}_{j1}, \cdots, \mathbf{W}_{jN_{\text{cl}}}]^T \), \( \mathbf{q}_{b,j} = [\mathbf{q}_{b,1,j}, \cdots, \mathbf{q}_{b,N_{\text{cl}},j}]^T \), and \( \mathbf{q}_{j} = [\mathbf{q}_{1,j}, \cdots, \mathbf{q}_{N_{\text{cl}},j}]^T \) are defined for notational simplicity. Moreover, \( \mathbf{W}_{j} \) denotes a precoding matrix applied to \( \mathbf{s}_{j} \), which is a submatrix of \( \mathbf{W} \): the \( (k-1)N_{\text{cl}} + 1 \)th column to \( kN_{\text{cl}} \)th column matrix of \( \mathbf{W} \) in precise. The difference of \( \mathbf{w}_{b,j} \) among RRs is considered in a diagonal matrix

\[ \mathbf{D} = \text{diag} \left( \mathbf{w}_{b,1} \mathbf{I}_{N_{\text{cl}}}, \cdots, \mathbf{w}_{b,N_{\text{cl}}} \mathbf{I}_{N_{\text{cl}}} \right), \]

where \( \text{diag}(\mathbf{a}) \) stands for a diagonal matrix whose diagonal entries are the elements of the vector \( \mathbf{a} \). In addition, \( \mathbf{n}_{j} \) denotes an additive white Gaussian noise (AWGN) at the jth Rx, which is a ZMCSG random variable whose covariance matrix is given by \( 1/N_{\text{cl}} \mathbf{I} \).
where $z_{n,j}$ denotes an additional noise due to a mis-detection. In case that CL signal is mis-detected, SIC can lay an additional noise as well as allowing the ILI still remain. We here model this additional noise independent to $s_n$, to be ZMCG random variable with covariance matrix $1/N_T I$. Note that the ILI components arise only when the CL detection is failed (the CL detection failure at the $j$th Rx is indicated by $E_{CL}^j$). According to rate-distortion theorem [44], the compression for broadcast stream is determined to satisfy $R_{b,j} = N_T \log (\alpha_b + 1) - N_T \log (1 - \alpha_b) - N_T \log (1 - \alpha_b)$ ($\iff \alpha_b = 1 - 2/N_T \epsilon_{NL^j}$). In addition, the CL rate achievable at the $j$th Rx is given by

$$R_{CL}^j(\delta) = \log |I + H_j^H (D + WW^H + \Omega_j) H_j| - \log |I + H_j^H (1 - \alpha_b)D + WW^H + \Omega_j) H_j|. \quad (6)$$

When the worst-case is considered, the ambiguity of $WW^H + \Omega_j$ in (6) can be removed by regarding maximal ratio transmission (MRT)-based precoding and compression for CL rate calculation, i.e., $WW^H + \Omega_j = \delta \sum_{i \in N_k} \tilde{P}_i/(1 + \delta) / \|H_j H_j^H\|$. Let $\tilde{P}_i$ to denote the transmission power constraint at the $i$th RRH. This alternative formulation leads to the modified CL rate of

$$R_{CL}^j(\delta) = \log |I + H_j^H D'H_j + \delta \sum_{i \in N_k} \tilde{P}_i H_j H_j^H H_j^H / (1 + \delta) \|H_j H_j^H\| |H_j H_j^H| - \log |I + (1 - \alpha_b) H_j^H D'H_j + \delta \sum_{i \in N_k} \tilde{P}_i H_j H_j^H H_j^H / (1 + \delta) \|H_j H_j^H\| |H_j H_j^H|. \quad (7)$$

where $D'$ is defined as $D' \triangleq 1/(1 + \delta) \text{diag}([\tilde{P}_1 I_{N_T,1}, \tilde{P}_2 I_{N_T,2}, \cdots, \tilde{P}_{N_k} I_{N_T,N_k}]).$

**Scenario 1. Optimizing fronthaul compression and precoding under given $\delta$**

In general, if broadcast programs are intended to be multiplexed in non-orthogonal manner, i.e., LDMA, the network operators determine a proper $\delta$ before the service launch. The $\delta$ is chosen to meet the target coverages, which are also related to the target quality of service (QoS). Let us denote the CL reception rate required for broadcast service decoding by $R_{CL}^b$. $R_{CL}^b$ is an alternative expression of a threshold of visibility (ToV), and is an upper bound of the target bitrate of the broadcast service. Thus, the network operator may determine $\delta$ to allow the expected CL reception rate at the arbitrary undetermined users to be $> R_{CL}^b$. Note that $R_{CL}^b$ also behaves as a threshold CL reception rate required for CL cancellation at the unicast users. The Rx is regarded to be in $R_{CL}^b$ when $R_{CL}^j(\delta) < R_{CL}^b$, in other words. Therefore, the condition indicator $1(R_{CL}^j(\delta) > R_{CL}^b)$ is henceforth replaced to $1(R_{CL}^j(\delta) < R_{CL}^b).

**Algorithm 1 Difference-of-concave (DC) algorithm for joint optimization of $W$ and $\Omega_j$ in problem (8)**

**Input:** Channel $H$, injection level $\delta$, fronthaul link capacity set $\{C_i\}_{i \in N_T}$, broadcast program stream rate $R_b$, the maximum transmit power set $\{\tilde{P}_i\}_{i \in N_k}$, and the convergence threshold $\epsilon > 0$.

**Initialization:** $n = 0$, $\{V_{j}^{(0)}\}_{j \in N_k}$, $\{\Omega_{a,i}^{(0)}\}_{i \in N_T}$.

**Repeat**

1. $n \leftarrow n + 1$
2. (Majorization-Minimization) Update $\{V_{j}^{(n)}\}_{j \in N_k}$ and $\{\Omega_{a,i}^{(n)}\}_{i \in N_T}$ by the solution of the problem (9):

$$\arg \max_{(V, \Omega)} \sum_{j \in N_k} R_j(V, \Omega, \delta | V^{(n-1)}, \Omega^{(n-1)} - \log |V_j| - \log |\Omega_{a,i}| \leq C_i - R_b, \text{Tr}[V_j + \Omega_{a,i}] \leq T_{Cl}[V_j + T_{Cl}[V_j + \Omega_{a,i}]) \leq \frac{\delta}{1 + \delta} \tilde{P}_i, \text{rank}(V_j) \geq N_r, V_j \geq 0.$$

**Until** a convergence criterion $\sum_{j \in N_k} ||V_j^{(n)} - V_j^{(n-1)}||_2 \leq \epsilon$ is satisfied.

**Output:** Compression noise matrix $\{\Omega_{a,i}^{(n)}\}_{i \in N_T}$ and precoding matrices $\{W_j^{(n)}\}_{j \in N_k}$ extracted from the solution $\{V_j\}_{j \in N_k} = \{V_j^{(n)}\}_{j \in N_k}$ via rank reduction.

In case $\exists \delta$ is given in advance, the achievable sum-unicast rate maximization problem can be formulated as

$$\max_{W, \Omega_{a,i} \in N_k} \sum_{j \in N_k} R_{CL}^l(W, \Omega_{a,i}) - \log |W_i H_i^H + \Omega_{a,i}| - \log |\Omega_{a,i}| \leq C_i - R_b, \text{Tr}[W_i H_i^H + \Omega_{a,i}] \leq \frac{\delta}{1 + \delta} \tilde{P}_i, \text{rank}(V_j) = N_r, V_j \geq 0.$$

for $\forall j \in N_r$ where the achievable unicast rate at the $j$th Rx $R_{CL}^j(W, \Omega_j, \delta)$ is given by (5). The problem designing the compression and precoding matrices illustrated above can be carried out by dealing a rank-relaxed problem, which copes with $V \triangleq WW^H$ as an explanatory variable instead of $W$. Thereby the problem (8) is reformulated as

$$\max_{V, \Omega_{a,i} \in N_k} \sum_{j \in N_k} R_{CL}^l(V, \Omega_{a,i}) - \log |V_i H_i^H + \Omega_{a,i}| - \log |\Omega_{a,i}| \leq C_i - R_b, \text{Tr}[V_i H_i^H + \Omega_{a,i}] \leq \frac{\delta}{1 + \delta} \tilde{P}_i, \text{rank}(V_j) = N_r, V_j \geq 0.$$

for $\forall i \in N_T$ and $\forall j \in N_R$, where $V_i \triangleq W_i H_i^H$ and $V_j \triangleq W_j H_j^H$ are defined for rank-relaxation. This problem
can be solved by MM type DC algorithm. The detailed procedure for the proposed compression/precoding optimization is elaborated in Algorithm 1. In each iterates of Algorithm 1, the program uses a concave upper bound of \( R_j^{EL}(V, \Omega_u, \delta) \), which is obtained by linearizing the negative terms of (5) around the current iterate:

\[
\begin{align*}
\hat{R}_j(V, \Omega_u, \delta | V^{(n-1)}, \Omega_u^{(n-1)}) &= \left\{ \log |H_j^H(V + \Omega_u + (1 - \alpha_b)D)H_j + I| - f(\{V_{N_k}^{(j)}, \Omega_u^{(j)}\}) \right\} I(R_j^{CL}(\delta) \geq R_b^{ref}) \\
&+ \left\{ \log |H_j^H(V + \Omega_u + (1 - \alpha_b)D)H_j + I| - f(\{V_{N_k}^{(j)}, \Omega_u^{(j)}\}, D) \right\} I(R_j^{CL}(\delta) < R_b^{ref})
\end{align*}
\]

(10)

as objective function, where the function \( f(A, B) \) is the first-order Taylor approximation of the log-det function \( \log |A| + 1/\ln 2 \text{ Tr}(A^{-1}(B - A)) \), having the slack variables \( \Xi_1(V_{N_k}^{(j)}, \Omega_u) \equiv H_j^H(\sum_{k \in N_k} V_k^{(n-1)} + \Omega_u^{(n-1)})H_j + I, \Xi_2(V_{N_k}^{(j)}, \Omega_u) \equiv H_j^H(\sum_{k \in N_k} V_k + \Omega_u)H_j + I, \Xi_3(V_{N_k}^{(j)}, \Omega_u, D) \equiv H_j^H(\sum_{k \in N_k} V_k + \Omega_u + (1 - \alpha_b)D)H_j + I, \Xi_4(V_{N_k}^{(j)}, \Omega_u, D) \equiv H_j^H(\sum_{k \in N_k} V_k + \Omega_u + (1 - \alpha_b)D)H_j + I \) for its input.

The iterative procedure consequently obtains a stationary point of \((V, \Omega_u)\), followed by D. Turning to the original problem (8), the net solution \( W \) is extracted from the solution \( V \) of the relaxed problem by rank-reduction introduced in [45]. That is, the submatrices of solution \( W \) are given by \( W_j^{ NEW} = \nu_{max}^j(V_j) \), where \( \nu_{max}^j(A) \) denotes a unitary matrix whose columns are the eigenvectors of the largest \( d \) eigenvalues of \( A \).

**Scenario 2. Sum-unicast rate maximization including optimization of \( \delta \)**

If the scheduler does not care about the QoS of broadcast service, \( \delta \) also can be chosen in order to maximize the sum-unicast rate \( \sum_{j \in N_k} R_j^{EL}(W, \Omega_u, \delta) \). For instance, in case that the medium access control (MAC) excludes the users requesting broadcast programs to be served via unicast [29] and the broadcast service is regarded as supplementary, \( \delta \) can be optimized in unicast-oriented perspective. Therefore, the sum-unicast rate maximization problem can be extended from (8) as

\[
\max_{W, \Omega_u, \delta} \sum_{j \in N_k} R_j^{EL}(W, \Omega_u, \delta)
\]

s.t. \( \log |W_i^H W_i^{H} + \Omega_u| \leq C_i - R_b \),

\[
\text{Tr}[W_i^H W_i^{H} + \Omega_u] \leq \frac{\delta}{1 + \delta} P_i,
\]

\[
\delta_{\min} \leq \delta \leq \delta_{\max}.
\]

(11)

for \( \forall i \in \mathcal{N}_T \), which requires a line-search on \( \delta \in [\delta_{\min}, \delta_{\max}] \). In fact, the exhaustive search on \( \delta \) would be quite burdensome to the CU. However, the candidate set for the optimal \( \delta \) can be effectively reduced into a finite set through the following theorem.

**Theorem 1.** The sum-rate \( \sum_{j \in N_k} R_j^{EL}(W, \Omega_u, \delta) \) monotonically increases by \( \delta \geq \delta_0 \) for \( \forall \delta_0 \in (0, 1] \) until there exists \( \exists j \in N_k \) such that \( R_j^{CL}(\delta') \geq R_b^{ref} \) holds for \( \delta' \in [\delta_0, \delta] \) but yields \( R_j^{CL}(\delta) < R_b^{ref} \).
Algorithm 2 Joint design of $\mathbf{W}$ and $\mathbf{\Omega}_u$ and optimization of $\delta$ solving problem (11)

**Input:** Channel $\mathbf{H}$, fronthaul link capacity set $\{C_i\}_{i \in \mathcal{N}_I}$, broadcast program stream rate $R_b$, the maximum transmit power set $\{P_i\}_{i \in \mathcal{N}_I}$, required CL reception rate for broadcast service decoding $R_b^{\text{ref}}$, and the convergence threshold $\epsilon > 0$.

**Initialization:** Calculate $\delta_1 \leq \cdots \leq \delta_j |\mathcal{N}_N|$ such that $R_k^{\text{CL}}(\delta_j) = R_b^{\text{ref}}$. Define a $\delta$ candidate set $\Phi_\delta = \{\delta_{\text{min}}, \delta_{\text{max}}\} \cup \{\delta_j | \delta_j \in \{\delta_{\text{min}}, \cdots, \delta_j |\mathcal{N}_N\}\}$. Let $\Phi_\delta$ denote the power set $\mathcal{P}(\{\delta_{\text{min}}, \delta_{\text{max}}\} \cup \{\delta_j | \delta_j \in \{\delta_{\text{min}}, \cdots, \delta_j |\mathcal{N}_N\}\})$.

**Do:**
- For each $\delta_i \in \Phi_\delta$, execute Algorithm 1 with input injection level $\delta_i$ and obtain the corresponding $\max_{\mathbf{w}, \mathbf{\Omega}_u, \delta} \sum_{j \in \mathcal{N}_N} R_j^{\text{EL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$.
- $\delta \leftarrow \arg \max_{\delta \in \Phi_\delta} \max_{\mathbf{w}, \mathbf{\Omega}_u} \sum_{j \in \mathcal{N}_N} R_j^{\text{EL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$.

**Output:** Injection level $\delta$, the corresponding solution of compression noise variances $\{\Omega_{u,i}\}_{i \in \mathcal{N}_I}$ and precoding matrices $\{\mathbf{W}_j\}_{j \in \mathcal{N}_N}$.

Let $\delta_k = 1, 2, \cdots, |\mathcal{N}_N|$ be the solution sets that satisfy $R_k^{\text{CL}}(\delta_k) = R_b^{\text{ref}}$, ordered in ascending order. Thereby, the search on optimal $\delta$ can be progressed in the reduced set $\{\delta_{\text{min}} \leq \delta_j \leq \delta_{\text{max}}\}_{j \in \mathcal{N}_N} \cup \{\delta_{\text{min}}, \delta_{\text{max}}\}$ as follows:

$$\max_{\mathbf{w}, \mathbf{\Omega}_u, \delta} \sum_{j \in \mathcal{N}_N} R_j^{\text{EL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$$

s.t. $\log |\mathbf{W}_j^t \mathbf{W}_j^H + \Omega_{u,i}| - \log |\Omega_{u,i}| \leq C_i - R_b$,

$$\text{Tr}[\mathbf{W}_j^t \mathbf{W}_j^H + \Omega_{u,i}] \leq \frac{\delta}{1 + \delta} \bar{P}_i,$$

$$\delta \in \{\delta_{\text{min}} \leq \delta_j \leq \delta_{\text{max}}\}_{j \in \mathcal{N}_N} \cup \{\delta_{\text{min}}, \delta_{\text{max}}\},$$

for $\forall i \in \mathcal{N}_I$, which can be solved via rank-relaxation and -reduction method likewise to scenario 1. The solution can be additionally conducted is to obtain solutions for $\delta_j \in \mathcal{N}_N$'s separately and compare the corresponding $\sum_{j \in \mathcal{N}_N} R_j^{\text{EL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$'s.

Algorithm 2 depicts the high-level procedure solving the problem (13) on the top of Algorithm 1.

B. Unicast in Core Layer and Broadcast in Enhanced Layer (UCBE)

If unicast services are allocated in CL during broadcast services are conveyed in EL, the unicast users do not need to perform SIC anymore. Thus, the unicast users only have to cope with ILI that can be simply treated as an arbitrary noise. Consequently, the unicast rate achievable at the $j$th Rx becomes

$$R_j^{\text{UCL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta) = \log |\mathbf{H}_j^H (\mathbf{W} \mathbf{W}_j^H + \Omega_{u,i} + \mathbf{D}^\dagger) \mathbf{H}_j + \mathbf{I}|$$

$$- \log |\mathbf{H}_j^H (\sum_{k \in \mathcal{N}_N \setminus j} \mathbf{W}_k \mathbf{W}_k^H + \mathbf{D}^\dagger \mathbf{H}_j + \mathbf{I})|,$$

for $\forall i \in \mathcal{N}_I$ and given $\delta$, $\delta$ should be determined in advance to guarantee the broadcast QoS at arbitrary users. It is obvious that the increase of $\delta$ does not benefit $\sum_{j \in \mathcal{N}_N} R_j^{\text{UCL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$. Unlike to Section III-A, the optimal $\delta$ that maximizes the sum-unicast rate therefore does not exist in this case, since $\sum_{j \in \mathcal{N}_N} R_j^{\text{UCL}}(\mathbf{w}, \mathbf{\Omega}_u, \delta)$ monotonically decreases by $\delta$. The problem (15) can be solved by an iterative algorithm similarly to Section III-A. A local maxima expression around the point $\{\mathbf{V}^{(n-1)}, \Omega_u^{(n-1)}\}$, $\mathbf{R}_j^{\text{UCL}}(\mathbf{V}, \mathbf{\Omega}_u, \delta |\mathbf{V}^{(n-1)}, \Omega_u^{(n-1)} = \log |\mathbf{H}_j^H (\mathbf{V} + \mathbf{W}_j + \mathbf{\Omega}_u + \mathbf{\Omega}_u) \mathbf{H}_j + \mathbf{I}) - f(\Psi_1(V_{N_{k,i}}), \Omega_u), \mathbf{\mathbf{\Psi}}_2(V_{N_{k,i}}, \Omega_u)), \text{ is defined for } \Psi_1(V_{N_{k,i}}), \Omega_u, \text{ and } \mathbf{\Psi}_2(V_{N_{k,i}}, \Omega_u). \text{ is defined for } \Psi_1(V_{N_{k,i}}, \Omega_u).$}
The received EL signal in (4) is rewritten by

\[ y_j^{EL} = (\mathbf{h}_j + \hat{\mathbf{h}}_j)^H \left( w_j s_j + \sum_{k \in N_k \setminus j} w_k^* s_k + q_u \right) + (\mathbf{h}_j + \hat{\mathbf{h}}_j)^H \mathbf{D}^2 (\sqrt{\alpha_b} s_b + \sqrt{1 - \alpha_b} q_u + \mathbf{z}_b, 1) (\mathcal{E}^{CL}) + n_j. \]  

This expression leads to the achievable unicast rate at the jth Rx

\[ \hat{R}_j^{CL}(\mathbf{V}, \Omega_u, \delta) = \min_{\Omega_u} \left\{ \begin{array}{l} \log (\text{Tr} \left[ (\hat{\mathbf{H}}_j + \Delta_j) (\mathbf{V} + \Omega_u + (1 - \alpha_b) \mathbf{D}) \right] + 1) \\ - \log \left( \text{Tr} \left[ (\mathbf{H}_j + \Delta_j) (\sum_{k \in N_k \setminus j} \mathbf{V}_k + \Omega_u + (1 - \alpha_b) \mathbf{D}) \right] + 1 \right) \end{array} \right\} \right. 

\[ \cdot 1(\hat{R}_j^{CL}(\delta) \geq R_{ref}^{CL}) + \left\{ \begin{array}{l} \log (\text{Tr} \left[ (\mathbf{H}_j + \Delta_j) (\mathbf{V} + \Omega_u + (1 + \alpha_b) \mathbf{D}) \right] + 1) \\ - \log \left( \text{Tr} \left[ (\mathbf{H}_j + \Delta_j) (\sum_{k \in N_k \setminus j} \mathbf{V}_k + \Omega_u + (1 + \alpha_b) \mathbf{D}) \right] + 1 \right) \end{array} \right\} \right. 

\[ \cdot 1(\hat{R}_j^{CL}(\delta) < R_{ref}^{CL}). \]  

where \( \hat{\mathbf{H}}_j \) is a covariance matrix of \( \mathbf{h}_j \), and a channel covariance uncertainty \( \Delta_j = \Delta_j^H + \hat{\mathbf{h}}_j^H \mathbf{H}_j \hat{\mathbf{h}}_j \) that satisfies \( \| \Delta_j \| \leq \theta_j^2 + 2 \theta_j \| \hat{\mathbf{h}}_j \| \leq \tau_j \) is defined for \( \forall j \in N_k \).

\[ \hat{R}_j^{CL}(\delta) = \log (\text{Tr} [\mathbf{H}_j (\mathbf{D} + \mathbf{V})] - \tau_j \| \mathbf{D} + \mathbf{V} \| + 1) - \log (\text{Tr} [\mathbf{H}_j (1 - \alpha_b) \mathbf{D} + \mathbf{V})] + \tau_j \| (1 - \alpha_b) \mathbf{D} + \mathbf{V} \| + 1 \] indicates the worst-case achievable CL rate that can be observed over \( \mathcal{H}_j \), whose calculation follows [47]. In consequence, the robust solution of \( (\mathbf{V}, \Omega_u) \) that copes with the worst-case channel uncertainty can be obtained from

\[
\max_{\mathbf{V}, \Omega_u} \left\{ \min_{j \in N_k} \sum_{j \in N_k} \hat{R}_j^{CL}(\mathbf{V}, \Omega_u, \delta) \right\}
\]

\[
s.t. \log |V_i^j + \Omega_u, \delta| - \log |\mathbf{V}_i, \delta| \leq \delta \left( \frac{1}{1 + \delta} \right) R_{ref},
\]

\[
\text{rank}(\mathbf{V}) = 1, \mathbf{V} \succeq 0,
\]

for \( \forall i \in N_T \) and \( \forall j \in N_k \). The term \( \min_{j \in N_k} \hat{R}_j^{CL}(\mathbf{V}, \Omega_u, \delta) \) in (19) can be replaced by a summation among \( \log (\text{Tr} [\mathbf{H}_j + \Delta_j^\text{min}]) (\mathbf{V} + \Omega_u + (1 - \alpha_b) \mathbf{D}) + 1 - \log (\text{Tr} [\mathbf{H}_j + \Delta_j^\text{max}]) (\mathbf{V} + \Omega_u + (1 + \alpha_b) \mathbf{D}) + 1) \) \( 1(\hat{R}_j^{CL}(\delta) \geq R_{ref}^{CL}) + \log (\text{Tr} [\hat{\mathbf{H}}_j + \Delta_j^\text{min}]) (\sum_{k \in N_k \setminus j} \mathbf{V}_k + \Omega_u + (1 + \alpha_b) \mathbf{D}) + 1) - \log (\text{Tr} [\hat{\mathbf{H}}_j + \Delta_j^\text{max}]) (\sum_{k \in N_k \setminus j} \mathbf{V}_k + \Omega_u + (1 + \alpha_b) \mathbf{D}) + 1) \) \( 1(\hat{R}_j^{CL}(\delta) < R_{ref}^{CL}) \) is, where \( \Delta_j^\text{min} \) and \( \Delta_j^\text{max} \) are the local extremums of \( \Delta_j \) value given by a Lagrangian method:

\[ \Delta_j^\text{min} = -\tau_j \frac{\log (\mathbf{V} + \Omega_u + (1 - \alpha_b) \mathbf{D})^H}{\| \mathbf{V} + \Omega_u + (1 - \alpha_b) \mathbf{D} \|} \left( 1(\hat{R}_j^{CL}(\delta) \geq R_{ref}^{CL}) \right) \]  

\[ + \tau_j \frac{\log (\mathbf{V} + \Omega_u + (1 + \alpha_b) \mathbf{D})^H}{\| \mathbf{V} + \Omega_u + (1 + \alpha_b) \mathbf{D} \|} \left( 1(\hat{R}_j^{CL}(\delta) < R_{ref}^{CL}) \right). \]  

Therefore, the problem (19) can be resolved by Algorithm 1 while the objective function at each iterate is altered from \( \sum_{j \in N_k} \hat{R}_j^{CL}(\mathbf{V}, \Omega_u, \delta) \| \mathbf{V}^{(n-1)}, \Omega_u^{(n-1)} \) to the lower bound of the achievable unicast-rate \( \sum_{j \in N_k} \hat{R}_j^{CL}(\mathbf{V}, \Omega_u, \delta) \| \mathbf{V}^{(n-1)}, \Omega_u^{(n-1)} \) in (16). Finally, the net precoding solution \( \mathbf{W} \) is extracted from the intermediate result of rank-relaxed problem, via rank-reduction. Note that \( \delta \) can also be optimized analogously to scenario 2 in Section III-A, through Algorithm 2.
B. Unicast in Core Layer and Broadcast in Enhanced Layer (UCBE)

UCBE case is re-explored under imperfect CSIT condition as well. Revisions on the derivations in Section IV-A, yield the achievable unicast rate at the $j$th user $R_j^{UCL} = \log(\text{Tr}([\mathbf{H}^j + \Delta_j]([\mathbf{V} + \Omega_a + \mathbf{D}''] + 1) - \log(\text{Tr}([\mathbf{H}^j + \Delta_j])(\sum_{k \in N_k \cup j} \mathbf{V}_k + \Omega_a + \mathbf{D}'') + 1))$. The lower bound of the sum-unicast rate, the worst-case within $\mathcal{H}_j$, is hence maximized by a max-min problem

$$\max_{\mathbf{V}, \Omega_a} \left( \min_{\|\Delta_j\| \leq \tau_j, j \in \mathcal{N}_j} \sum_{j \in N_k} R_j^{UCL}(\mathbf{V}, \Omega_a, \delta) \right)$$

s.t. \[ \log |\mathbf{V}_j + \Omega_a| - \log |\Omega_a| \leq C_i - R_b, \]
\[ \text{Tr}[\mathbf{V}_j + \Omega_a] \leq \frac{1}{1+\delta} P_t, \]
\[ \text{rank}(\mathbf{V}_j) = 1, \mathbf{V}_j \succeq 0, \]

for $\gamma_i \in \mathcal{N}_T$ and $\gamma_j \in \mathcal{N}_R$, where it can be resolved by a modified version of Algorithm 3. In the modified algorithm, the objective

$$\sum_{j \in N_k} R_j^{UCL}(\mathbf{V}, \Omega_a, \delta|\mathbf{V}^{(n-1)}, \Omega_a^{(n-1)}) \]
\[ \Delta \sum_{j \in N_k} \left\{ \log(\text{Tr}([\mathbf{H}^j + \mathbf{V} + \Omega_a + \mathbf{D}'']) + 1) \right. \]
\[ - f \left( \text{Tr} \left[ \mathbf{H}^j \left( \sum_{k \in N_k \cup j} \mathbf{V}_k^{(n-1)} + \Omega_a^{(n-1)} + \mathbf{D}'(n-1) \right) \right] \right) \]
\[ + \tau_j \left\| \sum_{k \in N_k \cup j} \mathbf{V}_k^{(n-1)} + \Omega_a^{(n-1)} + \mathbf{D}'(n-1) \right\| + 1, \]

is used instead of $\sum_{j \in N_k} R_j^{UCL}(\mathbf{V}, \Omega_a, \delta|\mathbf{V}^{(n-1)}, \Omega_a^{(n-1)})$.

V. NUMERICAL RESULTS

In this section, the achievable sum-unicast rates are shown with respect to the transmission power constraints and to the fronthaul capacity constraints among the CU-to-RRH links. Throughout the simulations, the C-RAN with $|\mathcal{N}_T| = 3$ RRHs and $|\mathcal{N}_R| = 3$ Rs were considered, where the fronthaul capacity constraints and transmission power constraints were assumed to be equal among the RRHs, i.e., $C_i = C_{FH}$ and $P_i = P$ for $\forall i \in \mathcal{N}_T$. In addition, the path-loss components were set to be $\beta_{j,k} = 1$, where each RRH and Rx were assumed to have three transmit antennas and single receive antenna, respectively. To evaluate the spectral-efficiency gain of the proposed compression/preceding designs, the simulations were conducted comparatively against the conventional decentralized architecture [48], which employs zero-forcing (ZF) precoding and uniform fronthaul compression. In the decentralized system we considered, each RRH designs its precoding matrix $\mathbf{W}_k \in \mathcal{N}_k$ regardless of other RRHs’ transmission. The precoding matrix of each RRH was optimized under ZF principle, by means of water-filling power allocation with Karush-Kuhn-Tucker (KKT) condition. Moreover, the proposed designs were also compared to an ideally centralized network, which assumes the RRHs to have internal servers for unicast services while the encoded stream for broadcast service is conveyed by the fronthaul link. In this ideally centralized case, the encoded streams for unicast services are assumed to be free from quantization noise (i.e., $\Omega_a = 0$), letting the fronthaul link to act like only a studio-to-transmitter link (STL) in broadcast transmission systems (i.e., $C_{FH} = R_b$).

The effect of per-RRH power constraint on

This procedure can be obtained similarly to [49]. Note that given the precoding matrix ZF-based and given the fronthaul quantization matrix fixed, the precoding optimization problem maximizing achievable EL rate in our NOMA system is equivalent to the problem in single layer transmission which is dealt in [49] and [50]. Such equivalence can be shown by simple calculation, while is omitted in this paper.

To avoid a confuse in terminology, we hereafter denote the decentralized ZF preceding-uniform fronthaul compression on both unicast and broadcast services as conventional decentralized scheme and denote the decentralized ZF preceding with no unicast stream compression during allocating entire fronthaul capacity solely to transporting the broadcast service stream as ideally decentralized network.
$\Sigma_{j \in N_k} P^j_{FH}(W, \Omega_y, \delta)$ is presented in Fig. 4. In Fig. 4, a drastic increase of sum-unicast rate in a certain $P_T$ region was found, particularly in $P_T \in [10, 15]$ dB for the proposed scheme with $\delta = -3$ dB. In those $P_T$ regions, the Rxs observe CL signal qualities comparable to the threshold SNR to attain a CL rate of $R_{b}^{\text{ref}}$. Therefore, in such regions, a notable portion of Rxs became able to cancel out CL successfully as $P_T$ increased. Since the higher $\delta$ degrades the CL signal quality more, the $P_T$ region that the Rxs transit into successful SIC appeared earlier at $\delta = -5$ dB case than at $\delta = -3$ dB case. On the other hand, the proposed scheme was shown to outperform the conventional decentralized scheme in entire area. With Algorithm 2 choosing the optimal $\delta$, the sum-unicast rate was found further enhanced, attaining about 5 bps/Hz enhancement compared to the conventional decentralized scheme at $P_T = 9$ dB. Furthermore, owing to optimized interference management, the proposed scheme was even found to outperform the ideally decentralized network in some regions, especially in low $P_T$ regime. However, the sum-unicast rate of the ideally decentralized network was shown to exceed that of the proposed scheme in high SNR regime, since it did not suffer from a compression noise for unicast signal. The no-compression assumption at unicast services (leading to $R_{b} = C_{FH}$) for the ideally decentralized network also allowed the Rxs to success at CL cancellation in lower $P_T$ region than the proposed scheme, while the assumption is too ideal to be considered in practice.

The effect of fronthaul compression noise on achievable sum-unicast rate was verified through Fig. 5 and Fig. 6. As clearly shown in Fig. 5, sum-unicast rate increased by $C_{FH}$. Such tendency was also observed in the ideally decentralized network, since the ideally decentralized network still suffered from compression noise in broadcast service transportation. It was also found that the increase of achievable sum-unicast rates saturate at $C_{FH} \geq 24$ bps/Hz, where the compression noise was given sufficiently low. One notable point is that $\delta = -8$ dB injection came with a sum-unicast rate enhanced than that of $\delta = -5$ and $\delta = -9$ dB for the proposed scheme. It implies that although Algorithm 2 does not consider a fairness on broadcast service quality for optimizing $\delta$, it possibly guarantees the broadcast quality by a coincidence with pursuing the maximization of achievable sum-unicast rate. Can be straightforwardly inferred from (5), the unicast rate gain from guaranteeing broadcast reception rate to be $\geq R_{b}^{\text{ref}}$ increased as $R_b$ is given lower, where the penalty of SIC failure comes greater.

According to Fig. 6, increasing $R_b$ could rather degrade the achievable sum-unicast rate when the Rxs fail at CL cancellation (See $R_b \in [2, 4]$ bps/Hz region of $R_{b}^{\text{ref}} = 0.3$ bps/Hz plots and $R_b \in [2, 6]$ bps/Hz region of $R_{b}^{\text{ref}} = 0.4$ bps/Hz plots, particularly). Recall (5) and note that an increase of $R_b$ yields the higher $\alpha_b$. Until CL cancellation succeeds, increasing $R_b$ lays more additional noise due to an incorrect cancellation along with harsher unicast compression noise. However, if CL cancellation proceeds successfully, reducing compression noise at broadcast service transportation enhances the sum-unicast rate until the portion of $R_b$ within $C_{FH}$ becomes too large. As shown in $C_{FH} = 14$ bps/Hz plots, unicast compression noise became intolerable for excessively high $R_b$s. To be emphasized, it was also shown that the proposed scheme outperformed the ideally decentralized network when $C_{FH}$ was sufficiently high, owing to its fine interference management capability.

Fig. 7 presents the sum-unicast rate degradation followed by inaccurate CSIT. In case of $\tau = 0.015$ and $\tau = 0.025$ for proposed scheme, the degradation compared to the perfect CSIT case were found 0.8 bps/Hz and 1.5 bps/Hz, respectively at $P_T = 9$ dB. Note that the CSI uncertainty was more harmful at low SNR region since the CSI error yielded more CL detection failures.

In order to discuss a feasibility of unicast in CL/broadcast in EL scenario (scenario 3, in short), we presented per-user broadcast rates and sum-unicast rates in Fig. 8. Basically, it is impossible to match the broadcast service quality of scenario 3 with that of scenario 1, due to the definition of $\delta \in (0, 1)$. Instead, for Fig. 8, the broadcast rate at the unicast users in scenario 3 was aligned with the public.
broadcast rate in scenario 1. To elaborate, we matched a CL cancellation-assumed broadcast rate in scenario 3 to a broadcast rate without CL cancellation (public broadcast rate, in other words) in scenario 1 by setting $\delta = -7$ and $\delta = -5$ dB for scenario 3 and 1, respectively. In this circumstance, the sum-unicast rate of scenario 3 was observed higher than that of scenario 1; particularly 0.2 bps/Hz higher at $P_f = 15$ dB. Moreover, even the imperfect CSI solution in scenario 3 with $\tau = 0.025$ was shown to achieve higher sum-unicast rate compared to Algorithm 1. However, the public broadcast rate in scenario 3 was seen to be less than a quarter of that in scenario 1. Since it is difficult for the unaccessed users to decode the CL, which contains individual unicast streams, scenario 3 would therefore not be feasible in practice where it does not guarantee a sufficient broadcast service reliability.

VI. CONCLUSION

In this paper, a joint design of fronthaul compression and precoding was introduced for NOMA-enabled broadcast-unicast co-transmission in C-RAN. Broadcast and unicast services overlaid in power domain were simultaneously transmitted via RRHs, under the centralized management through finite capacity-fronthaul connections. In order to cope with the capacity constraints at the fronthaul links, a multivariate compression for CU-to-RRH delivery was used. The proposed design of fronthaul compression and precoding obtained a sum unicast-rate maximization as long as guaranteeing a predefined fronthaul transmission rate for broadcast service. The design problems were tackled separately for different NOMA configurations, where two combinatorial options at allocating broadcast and unicast services in CL and EL of NOMA were considered. Given the coordinated transmission system, DC programming-based solutions combined with rank-relaxation and -reduction techniques were proposed. Moreover, the proposed algorithms were extended to acquire a robust fronthaul compression/precoding solution to deal with CSIT inaccuracy. The presented numerical results verified that the proposed schemes always achieve enhanced sum-unicast rate performance against the disjoint fronthaul compression/precoding design.

REFERENCES


