Generalized Compression Strategy for the Downlink Cloud Radio Access Network

Pratik Patil, Wei Yu, *Fellow, IEEE*

***Abstract*—This paper studies the downlink of a cloud radio access network (C-RAN) in which a centralized processor (CP) communicates with mobile users through base stations (BSs) that are connected to the CP via finite-capacity fronthaul links. Information theoretically, the downlink of a C-RAN is modeled as a two-hop broadcast-relay network. Among the various transmission and relaying strategies for such model, this paper focuses on the compression strategy, in which the CP centrally encodes the signals to be broadcast jointly by the BSs, then compresses and sends these signals to the BSs through the fronthaul links. We characterize an achievable rate region for a generalized compression strategy with Marton's multicoding for broadcasting and multivariate compression for fronthaul transmission. We then compare this rate region with the distributed decode-forward (DDF) scheme, which achieves the capacity of the general relay networks to within a constant gap, and show that the difference lies in that DDF performs Marton's multicoding and multivariate compression jointly as opposed to successively as in the compression strategy. A main result of this paper is that under the assumption that the fronthaul links are subject to a *sum* capacity constraint, this difference is immaterial; so, for the Gaussian network, the compression strategy based on successive encoding can already achieve the capacity region of the C-RAN to within a constant gap, where the gap is independent of the channel parameters and the power constraints at the BSs. As a further result, for C-RAN under individual fronthaul constraints, this paper also establishes that the compression strategy can achieve to within a constant gap to the *sum* capacity.**

***Index Terms*—Cloud radio access network (C-RAN), compression, distributed decode-forward, fronthaul, relay channel.**

I. Introduction

This paper studies the downlink of a cloud radio access network (C-RAN) in which the base stations (BSs) are connected to a centralized cloud-computing-enabled processor through wired or wireless fronthaul links [1]. Information theoretically, the downlink C-RAN can be modeled as a broadcast-relay channel: the CP broadcasts the user messages to the BSs via the fronthaul links and the BSs act as relays for the mobile users. This paper considers the C-RAN model where the BSs are connected to the CP through noiseless digital fronthaul links of finite capacities and there are no direct links between the CP and the mobile users. In the ideal case where the capacities of the fronthaul links are infinite, downlink C-RAN model reduces to a multi-antenna broadcast channel. The optimal transmission strategy in this case is cooperative beamforming combined with dirty-paper coding (DPC) [2]. For the practical situation where the fronthaul links have finite capacities, the optimal coding strategy must combine both broadcasting and relaying, and is highly non-trivial; the characterization of the capacity region is still an open problem. This paper makes progress in establishing the achievable rate region of a generalized compression strategy and in showing that it is approximately optimal for the downlink C-RAN under certain conditions.

Pratik Patil was with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering the University of Toronto; he is now with the Department of Statistics and Data Science and the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: pratik@cmu.edu).

Wei Yu is with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario M5S3G4, Canada (e-mail: weiyu@ece.utoronto.ca).

A. Coding Strategies

While the C-RAN architecture has been originally motivated by the radio-over-fiber concept [1], the information theoretical study of the downlink C-RAN model belongs to that of relay channels, and more specifically relates to the so-called diamond relay channels for which there is an extensive literature, e.g., [3], [4], [5], [6], [7]. In the C-RAN context, there are two main classes of transmission and relaying strategies available in the literature: the data-sharing and the compression strategies. In the data-sharing strategy, individual user messages are sent directly via the digital fronthaul to the BSs, which then perform cooperative beamforming to the users. The capacity constraints of the fronthaul links limit the number of users whose messages can be sent to each BS, hence limiting the cooperation BS cluster size for each user. Among the data-sharing schemes, joint encoding at the BSs can be done using linear beamforming with the sharing of the entire messages [8] or with message splitting [9]. Generalized versions of the data-sharing strategy using Marton's broadcast coding have been proposed for a 2-user 2-BS C-RAN in [10], and improved upon in [11], [12] by using a common message, and further generalized in [13] for arbitrary number of users and BSs. Although the data-sharing strategy does not necessarily achieve the capacity in general, there are some special cases for which it does. For example, the achievable rate based on Marton's coding proposed in [14] for a C-RAN with a single user (but any number of BSs) can be shown to achieve the capacity in some interesting regimes of operation. Upper bounds on the sum rate of some other specific cases of C-RAN model are studied in [15]. We also mention here that instead of sharing the individual user messages directly, the CP may send a function of user messages to the BSs. For example, in the reverse compute-forward strategy [16], a function of the messages is relayed to the BSs using lattice codes. As an alternative to the data-sharing strategy, the capacity limitation of the fronthaul links can also be dealt with using a compression strategy [17], in which the encoding is performed at the CP as a function of

the messages of all users, but in order to accommodate the capacity constraints of the fronthaul links, the encoded analog signals are compressed and sent to the BSs. The BSs then transmit the encoded signals to the users after decompressing the received compression bits. We note here that a hybrid scheme combining the data sharing and compression strategies is also possible [18].

This paper aims to understand the information theoretical optimality of the compression strategy for C-RAN. As pointed out earlier, if the fronthaul capacity is infinite, the downlink C-RAN reduces to the well-known vector Gaussian broadcast channel, for which DPC achieves the capacity region. For the finite fronthaul case, DPC and linear precoding schemes cannot be applied directly. A compressed version of DPC using independent compression across the BSs is introduced in [19] and the achievable user rates are derived for a simplified Wyner type model. The independent compression scheme can be further improved by using a multivariate compression strategy across all the BSs [20]. The idea is to correlate the quantization noises at the different BSs to better control the effect of quantization at the users. The achievable rate expressions under linear beamforming and multivariate compression for the Gaussian C-RAN model are given in [20] and the corresponding achievable rate region using dirty paper coding followed by multivariate compression is given in [1].

Can either the data-sharing or compression strategy approach the information theoretic capacity region of the C-RAN model? Toward answering this question, this paper draws inspiration from a new coding strategy named distributed decode-forward (DDF) [21] for broadcasting multiple messages over a general relay network, which has been shown to achieve the capacity region of the general Gaussian broadcast relay network to within a constant gap, which is linear in the number of nodes in the network but is independent of the channel parameters and the power constraints. We remark that when specialized to the downlink C-RAN model, the gap can be improved from linear to logarithmic in the number of users and BSs [22]. Further, it may be possible to further enlarge the rate region of the DDF strategy by incorporating a common codeword, as shown for a two-user two-BS C-RAN model with BS corporation in [11], [12].

B. Contributions

This paper makes an observation that when specialized to the C-RAN model, the DDF strategy resembles the compression strategy for C-RAN, but with a crucial difference that instead of performing the compression followed by Marton's multicoding, the DDF performs both the Marton's coding and multivariate compression jointly at the CP. As practical implementation for performing successive Marton's coding and multivariate compression would likely be easier, we ask in this paper whether there are conditions under which the difference is immaterial. One of the main results of this paper is that under a sum fronthaul constraint, this is indeed true. Thus, for the Gaussian C-RAN under the sum fronthaul constraint, the compression strategy can already achieve the capacity region to within a constant gap. As a further result, for the Gaussian C-RAN under individual fronthaul constraints, this paper also shows that Marton's encoding followed by multivariate compression can achieve the sum capacity to within a constant gap. More specifically, this paper makes the following contributions:

1) We provide the achievable rate region of a general form of the compression strategy that includes Marton's multicoding followed by multivariate compression for the C-RAN model with digital fronthaul in the first hop and a general discrete memoryless channel (DMC) in the second hop.
2) We specialize the DDF strategy to the C-RAN model and compare the coding strategies of the above generalized compression strategy and the DDF strategy. We observe that DDF is a further generalization in that the Marton's coding and multivariate compression are done jointly.
3) We analyze the conditions under which such a generalization of the compression strategy in the DDF strategy does not strictly enlarge the achievable rate region.
    a) With any DMC on the second hop, the generalized compression strategy and the DDF strategy achieve the same rate region under a sum fronthaul constraint.
    b) With a Gaussian network on the second hop, the sum rate achieved by the above general compression strategy is within a constant gap to the sum capacity of C-RAN, where the gap is independent of the network parameters.

C. Notation and organization

Random variables are denoted by uppercase letters, their realizations by lowercase letters, and the probability distributions by $p(\cdot)$. Sets are denoted by calligraphic letters, while $[1:n]$ denotes the set $\{1,\ldots,n\}$ for all natural numbers n. A subscript for a random variable and its realization denotes its node index. A superscript for a random variable or its realization is a time index that denotes a sequence of random variables or its realizations till that index (e.g., $X_l^n = (X_l^1,\ldots,X_l^n)$ or $x_l^n = (x_l^1,\ldots,x_l^n)$). Random variables can be indexed with sets (e.g., $X(\mathcal{S}) = (X_l : l \in \mathcal{S})$). Bold-face lower case letters are used to denote vectors and bold-face upper case letters are used to denote random vectors or matrices. The standard notations for entropy, $H(X)$, and mutual information, $I(X;Y)$, are used. Total correlation between a group of random variables is denoted by $T(\cdot)$ and is defined as

$$T(X(\mathcal{S})) = \sum_{l\in\mathcal{S}} H(X_l) - H(X(\mathcal{S})). \tag{1}$$

See [23] for motivation of such a definition and some of its properties. We follow the typicality notation of [24] and use $\mathcal{T}_\epsilon^{(n)}$ to denote the set of typical sequences of length n with parameter ϵ.

The rest of the paper is organized as follows. Section II provides a mathematical model for the downlink C-RAN. Section III provides the achievable rate region results of the generalized compression strategy. Section IV specializes the distributed decode-forward strategy to the downlink C-RAN

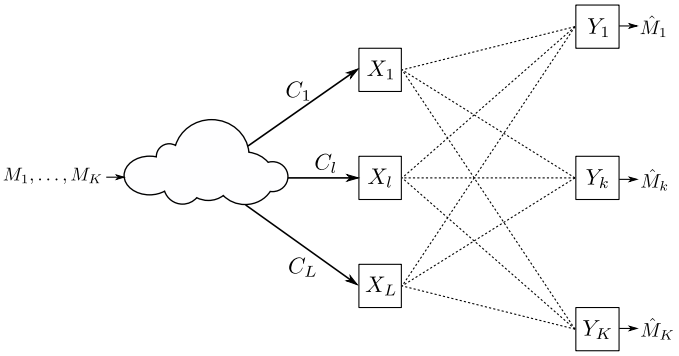

Fig. 1. Downlink C-RAN with L BSs, K users, and a channel $p(y_1, \ldots, y_K|x_1, \ldots, x_L)$ between the BSs and the users.

model under consideration. In Section V, we compare the rate regions achieved by the two strategies and provide conditions under which the two coincide. Section VI concludes the paper.

II. System Model

Consider the downlink of a C-RAN comprising of a CP and L BSs serving K users as shown in Fig. 1. The CP communicates with BSs through noiseless fronthaul links of finite capacities, denoted by C_l for BS l, $l \in \mathcal{L} \coloneqq [1:L]$. We assume a discrete memoryless channel $(\mathcal{X}_1 \times \cdots \times \mathcal{X}_L, p(y_1, \ldots, y_K|x_1, \ldots, x_L), \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K)$ between the BSs and the users. Let the intended message for user k be denoted by M_k, $k \in \mathcal{K} \coloneqq [1:K]$. A $(2^{nR_1}, \ldots, 2^{nR_K}, n)$ code for the downlink C-RAN consists of a mapping at the CP from the K user messages $(m_1, \ldots, m_K) \in [1:2^{nR_1}] \times \cdots \times [1:2^{nR_K}]$ to L indices $(t_1, \ldots, t_L) \in [1:2^{nC_1}] \times \cdots \times [1:2^{nC_L}]$, encoders at the L BSs that map the index t_l to a codeword $x_l^n(t_l)$, and decoders at the K users that estimate $\hat{m}_k$ based on the received signals y_k^n. The average probability of error is defined as $P_e^{(n)} = P\{\hat{m}_k \neq m_k \text{ for some } k \in \mathcal{K}\}$. A rate tuple $(R_1, \ldots, R_K)$ is achievable if there exists a sequence of codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$.

Of particular interest is the special case where the channel between the BSs and the users is a Gaussian channel such that

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}, \tag{2}$$

where $\mathbf{Y} = [Y_1, \ldots, Y_K]^T$ are the received signals at the K users, $\mathbf{X} = [X_1, \ldots, X_L]^T$ are the transmitted signals from the L BSs, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_k]^T$ is the $K \times L$ channel matrix consisting of channel vectors $\mathbf{h}_1$ to $\mathbf{h}_k$ for users 1 to K, respectively, and $\mathbf{Z} = [Z_1, \ldots, Z_K]^T \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is the additive white Gaussian noise. We assume all the BSs have an average power constraint of P without loss of generality. For simplicity, both the BSs and the users are assumed to be equipped with a single antenna in this paper.

III. Generalized Compression Strategy

The compression strategy has been extensively studied in the literature [17], [19], [20]. The coding strategy involves two steps. First, the CP jointly encodes the user messages. Second, the encoded signals are compressed in order to accommodate them through the fronthaul links. Different options for joint encoding include linear beamforming strategies such as zero-forcing or regularized zero-forcing, or non-linear beamforming strategy such as dirty paper coding. Different options for compression include independent compression or multivariate compression. The main point of this section is to show that these specific compression strategies previously studied in [17], [19], [20] are special forms of a generalized compression strategy in which joint encoding is performed via Marton's multicoding. The coding strategy proposed in this paper does not, however, incorporate the possibility of a common codeword, as done in [11], [12].

Theorem 1. *A rate tuple $(R_1, \ldots, R_K)$ is achievable for the downlink C-RAN using the compression strategy with Marton's multicoding followed by multivariate compression if*

$$\sum_{k\in\mathcal{D}} R_k < \sum_{k\in\mathcal{D}} I(U_k; Y_k) - T(U(\mathcal{D})) \tag{3}$$

for all $\mathcal{D} \subseteq \mathcal{K}$ such that

$$\sum_{l\in\mathcal{S}} C_l > I(U(\mathcal{K}); X(\mathcal{S})) + T(X(\mathcal{S})) \tag{4}$$

for all $\mathcal{S} \subseteq \mathcal{L}$ for some distribution $p(u_1, \ldots, u_K, x_1, \ldots, x_L)$.

The proof of achievibility is in Appendix A. The set of inequalities (3) represents the achievable user rates using Marton's multicoding for broadcast channels. In linear beamforming, the U's are just the messages and are thus independent of each other. The advantage of using Marton's multicoding is to introduce correlation among U's for the possibility of increased rates. But doing so incurs a penalty that depends on the total correlation present among U's. DPC is an example of such Marton's coding.

One way to implement Marton' coding is through successive encoding of user messages. Assuming without loss of generality that the encoding order is user $1, \ldots, K$. The achievable rate for user 1 is $I(U_1; Y_1)$. Treating user 1's message as known interference, user 2 achieves a rate of $I(U_2; Y_2) - I(U_1; U_2)$; and user k achieves a rate of $I(U_k; Y_k) - I(U_k; U_{k-1}, \ldots, U_1)$. We remark that, as pointed out in [25], there is a subtle issue that such successive encoding may not achieve the entire Marton's region. The reason is that even though the set function $\sum_{k\in\mathcal{D}} I(U_k; Y_k) - T(U(\mathcal{D}))$ satisfies the submodular property, it is not guaranteed that it satisfies the monotone property that the successive user rates $I(U_k; Y_k) - I(U_k; U_{k-1}, \ldots, U_1)$ are always non-negative. Hence, the Marton's region itself is not guaranteed to be a polymatroid. The rest of this paper ignores this subtlty and assumes that the Marton' rate region is polymatroid so that we can use successive encoding to achieve the corner points of the rate region.

The set of inequalities (4) represents the multivariate compression of $U(\mathcal{K})$ into X's that are transmitted by the BSs. If the BSs were co-located and can cooperate, the amount of quantization needed for compression is simply the first term $I(U(\mathcal{K}); X(\mathcal{S}))$. If the BSs are distributed and cannot cooperate, there is a penalty in terms of the correlation between the signals transmitted by the BSs.

Similar to the successive encoding for the Marton's region, the multivariate compression can also be implemented in a successive manner [20]. Without loss of generality, let's assume that the encoding order is BS $1,\ldots,L$. The fronthaul required to compress the signal for BS 1 is $I(U(\mathcal{K});X_1)$. After compressing the signal for BS 1, the fronthaul required to compress BS 2's signal is given by $I(U(\mathcal{K});X_2)+I(X_2;X_1)$; and for any BS l the fronthaul required is $I(U(\mathcal{K});X_l)+I(X_l;X_{l-1},\ldots,X_1)$. It can be verified that the fronthaul region in general is a contra-polymatroid [25].

The above achievability region has been presented at [26] and is subsequently generalized in [11] to the case with common information and BS cooperation where there are two BSs in the C-RAN. We now specialize the generalized compression strategy for Gaussian C-RAN (2) using various choices for the distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ and show how it results in the known compression strategies in the literature. We assume 2 BSs and 2 users for simplicity.

Consider the strategy of linear beamforming followed by compression. In this case, we choose the messages U's as independent Gaussian random variables, compute the beamformed signals to be transmitted by the BSs at the cloud, then compress using either independent compression or multivariate compression. Mathematically, we express the distribution $p(u_1,u_2,x_1,x_2)$ as

$$\mathbf{X}=\mathbf{W}\mathbf{U}+\mathbf{N}, \tag{5}$$

where $\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2]$ is a beamforming matrix with beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ for users 1 and 2, respectively, and $\mathbf{N}$ is the quantization noise, assumed to be a Gaussian vector $\mathcal{N}(0,\mathbf{Q})$. Here, $\mathbf{U}=[U_1,U_2]^T\sim\mathcal{N}(0,\mathbf{I})$ are the independent message signals for the two users. The achievable user rates of the generalized compression strategy with this choice of $\mathbf{U}$ are given by

$$R_1^{\text{linear}}=I(U_1;Y_1) \tag{6}$$

$$=\frac{1}{2}\log\left(1+\frac{\mathbf{h}_1^T\mathbf{w}_1\mathbf{w}_1^T\mathbf{h}_1}{\mathbf{h}_1^T\mathbf{w}_2\mathbf{w}_2^T\mathbf{h}_1+\mathbf{h}_1^T\mathbf{Q}\mathbf{h}_1+\sigma^2}\right) \tag{7}$$

for user 1, and similarly for user 2

$$R_2^{\text{linear}}=I(U_2;Y_2) \tag{8}$$

$$=\frac{1}{2}\log\left(1+\frac{\mathbf{h}_2^T\mathbf{w}_2\mathbf{w}_2^T\mathbf{h}_2}{\mathbf{h}_2^T\mathbf{w}_1\mathbf{w}_1^T\mathbf{h}_2+\mathbf{h}_2^T\mathbf{Q}\mathbf{h}_2+\sigma^2}\right). \tag{9}$$

Note that the covariance matrix of $\mathbf{N}$ enters the rate expression as an additional noise term. Depending on the compression strategy used, $\mathbf{Q}$ is either diagonal in case of independent compression owing to independent noise components among the compressed BS signals, or a full matrix in case of multivariate compression, due to the introduced correlation among the noise components of $\mathbf{N}$. In the independent compression case, let $\mathbf{Q}=\text{diag}(q_{11},q_{22})$, and $\mathbf{w}_1=[w_{11},w_{12}]^T$ and $\mathbf{w}_2=[w_{21},w_{22}]^T$. The amount of fronthaul needed to support compression at BS 1 is

$$C_1^{\text{linear,indep}}=I(X_1;U_1,U_2) \tag{10}$$

$$=\frac{1}{2}\log\left(1+\frac{w_{11}^2+w_{21}^2}{q_{11}}\right), \tag{11}$$

and similarly for BS 2,

$$C_2^{\text{linear,indep}}=I(X_2;U_1,U_2) \tag{12}$$

$$=\frac{1}{2}\log\left(1+\frac{w_{12}^2+w_{22}^2}{q_{22}}\right). \tag{13}$$

For the multivariate compression, the required fronthaul rates (C_1,C_2) must be inside a rate region. A corner point of the region assuming a successive compression strategy with the order of compression to be BS 1 followed by BS 2 is as following. For BS 1, $C_1^{\text{linear,multi}}$ is exactly the same as the independent compression case (11), but for BS 2, we have

$$C_2^{\text{linear,multi}}=I(X_2;U_1,U_2|X_1)+I(X_1;X_2) \tag{14}$$

$$=I(X_2;U_1,U_2)+I(X_1;X_2|U_1,U_2) \tag{15}$$

$$=\frac{1}{2}\log\left(1+\frac{w_{12}^2+w_{22}^2}{q_{22}}\right)+\frac{1}{2}\log\left(\frac{q_{22}}{q_{22}-q_{21}q_{11}^{-1}q_{12}}\right) \tag{16}$$

where $\mathbf{Q}=\begin{bmatrix}q_{11}&q_{12}\\q_{21}&q_{22}\end{bmatrix}$ is a full matrix whose correlation structure, although leading to higher (C_1,C_2), nevertheless allows possible reduction in the effective noise in the achievable rates (7) and (9), thereby potentially providing an overall benefit. These derived rate expressions can be shown to be equivalent to that in [20].

The linear beamforming strategy can be improved by introducing correlation between the U's. One example of using such correlation is DPC, which is capacity achieving for the Gaussian vector broadcast channel (i.e., with infinite C_1 and C_2). With DPC, the U's are now random vectors. Although using the U's designed for the broadcast channel for C-RAN is not necessarily optimal when C_1 and C_2 are finite, it is nevertheless instructive to write down the rate expressions to gain some insight. Assume an ordering of DPC with user 1 followed by user 2. The auxiliary random variables for DPC can be constructed as follows. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two independent Gaussian vectors with covariance matrices $\mathbf{K}_1$ and $\mathbf{K}_2$. Fix $\mathbf{N}\sim\mathcal{N}(0,\mathbf{Q})$. We choose

$$\mathbf{U}_1=\mathbf{S}_1,\mathbf{U}_2=\mathbf{S}_2+\mathbf{A}\mathbf{S}_1,\mathbf{X}=\mathbf{S}_1+\mathbf{S}_2+\mathbf{N}, \tag{17}$$

where $\mathbf{A}=\mathbf{K}_2\mathbf{h}_2\left(\mathbf{h}_2^T\left(\mathbf{K}_2+\mathbf{Q}\right)\mathbf{h}_2+\sigma^2\right)^{-1}\mathbf{h}_2^T$ and $\mathbf{N}$ is the quantization noise. This choice of the auxiliary variables allows the interference from user 1 to be completely presubtracted from user 2 [27], resulting in the following achievable user rates

$$R_1^{\text{DPC}}=I(\mathbf{U}_1;Y_1) \tag{18}$$

$$=\frac{1}{2}\log\left(1+\frac{\mathbf{h}_1\mathbf{K}_1\mathbf{h}_1^T}{\mathbf{h}_1\mathbf{K}_2\mathbf{h}_1^T+\mathbf{h}_1\mathbf{Q}\mathbf{h}_1^T+\sigma^2}\right) \tag{19}$$

for user 1 who sees user 2 as noise, and

$$R_2^{\text{DPC}}=I(\mathbf{U}_2;Y_2)-I(\mathbf{U}_1;\mathbf{U}_2) \tag{20}$$

$$=I(\mathbf{X};Y_2|\mathbf{S}_1) \tag{21}$$

$$=\frac{1}{2}\log\left(1+\frac{\mathbf{h}_2\mathbf{K}_2\mathbf{h}_2^T}{\mathbf{h}_2\mathbf{Q}\mathbf{h}_2^T+\sigma^2}\right) \tag{22}$$

for user 2, who no longer sees user 1 as interference. The required fronthaul rates depend on whether independent or multivariate compression is performed. Let the covariance matrix of $\mathbf{S}_1+\mathbf{S}_2$ be $\mathbf{K}_1+\mathbf{K}_2=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$ and the covariance matrix of the quantization noise $\mathbf{N}$ be $\mathbf{Q}=\begin{bmatrix}q_{11} & q_{12}\\ q_{21} & q_{22}\end{bmatrix}$. With independent compression, we have $q_{12}=q_{21}=0$ and

$$C_1^{\mathrm{DPC,indep}} = I(X_1;\mathbf{U}_1,\mathbf{U}_2) \tag{23}$$
$$= \frac{1}{2}\log\left(1+\frac{\Sigma_{11}}{q_{11}}\right) \tag{24}$$

$$C_2^{\mathrm{DPC,indep}} = I(X_2;\mathbf{U}_1,\mathbf{U}_2) \tag{25}$$
$$= \frac{1}{2}\log\left(1+\frac{\Sigma_{22}}{q_{22}}\right). \tag{26}$$

For multivariate compression, assuming the corner point of the fronthaul rate region with the ordering of compression to be BS 1 followed by BS 2, we have $C_1^{\mathrm{DPC,multi}}$ for BS 1 exactly the same as in the case of independent compression (23), but for BS 2, we need additional fronthaul capacity given by

$$C_2^{\mathrm{DPC,multi}} = I(X_2;\mathbf{U}_1,\mathbf{U}_2|X_1)+I(X_1;X_2) \tag{27}$$
$$= I(X_2;\mathbf{U}_1,\mathbf{U}_2)+I(X_1;X_2|\mathbf{U}_1,\mathbf{U}_2) \tag{28}$$
$$= \frac{1}{2}\log\left(1+\frac{\Sigma_{22}}{q_{22}}\right) + \frac{1}{2}\log\left(\frac{q_{22}}{q_{22}-q_{21}q_{11}^{-1}q_{12}}\right). \tag{29}$$

These rate expressions for DPC over C-RAN are equivalent to the ones given in [1]. They can be interpreted as the compression of $\mathbf{X}$ at the CP for transmission to the BSs. The above more rigorous derivation is based on transmitting $\mathbf{U}$ to the BSs via compression.

IV. Distributed Decode-Forward

The main objective of this paper is to understand whether the generalized compression strategy can approximately achieve the capacity region of the Gaussian C-RAN model. Toward this end, we examine the DDF strategy [21], which is a general coding scheme for broadcasting multiple messages over a general relay network that combines Marton's coding for the broadcast channel with partial decode-forward for the relay channel. The coding scheme involves using auxiliary random variables at each node in the network that implicitly carry information about the user messages. By specializing the DDF strategy to the C-RAN setup, we write down a succinct form of the achievable rate region using DDF and a simplified coding strategy that can be readily compared with the generalized compression strategy.

Theorem 2 ([21]). *A rate tuple $(R_1,\ldots,R_K)$ is achievable for the downlink C-RAN using the DDF strategy if*

$$\sum_{k\in\mathcal{D}} R_k < \sum_{k\in\mathcal{D}} I(U_k;Y_k)+\sum_{l\in\mathcal{S}} C_l - T(U(\mathcal{D}),X(\mathcal{S})) \tag{30}$$
$$= \sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})) + \sum_{l\in\mathcal{S}} C_l - I(U(\mathcal{D});X(\mathcal{S})) - T(X(\mathcal{S})) \tag{31}$$

for all $\mathcal{D}\subseteq\mathcal{K}$ and $\mathcal{S}\subseteq\mathcal{L}$ for some distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$.

The proof of achievibility is in Appendix B. Comparing the DDF coding strategy of Theorem 2 with that of the generalized compression strategy of Theorem 1, we observe that the DDF strategy generalizes the compression strategy by combining Marton's multicoding with multivariate compression and jointly encoding the Marton's and compression codewords. The key difference is that, in the compression strategy, Marton's codewords are formed first, then the multivariate compression codewords are computed in a sequential order. Note that the rate region in Theorem 1 is in general a subset of the rate region in Theorem 2 as any distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ satisfying the multivariate compression constraints (4) results in generalized compression rates (3) which are also achievable in the form (31) using the DDF strategy.

A key advantage of enlarging the allowable distributions to beyond the ones that explicitly satisfy the fronthaul constraints is that it permits a proof of the result that the DDF strategy can achieve to within a constant gap to the cut-set bound of the general Gaussian broadcast relay channel [21]. The ingenious choice of $p(u_1,\ldots,u_K|x_1,\ldots,x_L)$ proposed in [21] that accomplishes this task is a distribution for $p(u_1,\ldots,u_K|x_1,\ldots,x_L)$ that tries to mimic the Gaussian channel distribution $p(y_1,\ldots,y_K|x_1,\ldots,x_L)$. We now specialize the result of [21] to the C-RAN setup (2). The DDF strategy can be shown to achieve to within a constant gap to the cut-set outer bound by choosing $\mathbf{X}$ to be a vector of L independent Gaussian random variables $\mathcal{N}(0,P)$ and by choosing

$$\mathbf{U}=\mathbf{H}\mathbf{X}+\tilde{\mathbf{Z}}, \tag{32}$$

where $\tilde{\mathbf{Z}}\sim\mathcal{N}(0,\sigma^2 I)$ is independent of $\mathbf{Z}$. With this choice of $p(u_1,\ldots,u_K|x_1,\ldots,x_L)$, we have

Corollary 1 ([22]). *With Gaussian $p(y_1,\ldots,y_K|x_1,\ldots,x_L)$ on the second hop of the C-RAN model and individual fronthaul constraints $(C_1,\ldots,C_K)$, the DDF strategy achieves a rate region within a constant gap to the capacity region of C-RAN, where the gap is independent of the channel, the BS power constraints, and the fronthaul constraints, and only depends on the number of BSs and users.*

A natural question at this point is whether we can use the generalized compression strategy to accomplish the same. The next section gives some partial answers in the affirmative but under specific conditions.

V. Compression versus DDF

DDF generalizes the compression strategy, so the achievable rate region of the generalized compression strategy is a subset of the DDF region in general. This section asks the question of whether this subset inclusion is strict. The main result here is that, under certain conditions, the rate regions of the two strategies actually coincide. Specifically, we show that under a sum fronthaul constraint, the rate regions of the two strategies coincide for any discrete memory channel on the second hop of C-RAN. In other words, under a sum fronthaul constraint, performing Marton's coding and multivariate compression separately does not reduce the achievable user rates. As a second result of this section, we show that in the special case of Gaussian networks but under individual fronthaul constraint, the compression strategy achieves the sum capacity of C-RAN to within a constant gap. These results are useful, because successive Marton's coding and multivariate compression is likely easier to implement than the joint encoding for DDF. For example, an architecture based on successive estimation of minimum mean-squared error and per-BS compression to achieve the multivariate compression region is proposed in [17], while polar coding based scheme to achieve the general Marton's region for a 2-user broadcast channel is proposed in [28].

A. Rate Region Under Sum Fronthaul Constraint

Definition 1. *Consider the closure of the convex hull of achievable rate-fronthaul tuples $(R_1,\ldots,R_K,C_1,\ldots,C_L)$ using the generalized compression strategy satisfying (3)-(4) over all joint distributions $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ satisfying possibly input constraints on $(x_1,\ldots,x_L)$. Define $\mathcal{R}^s_{COM}(C)$ to be the projection of the above set along a sum fronthaul constraint C, i.e., the set of rate tuples $(R_1,\ldots,R_K)$ such that $C_l\geq 0$ and $\sum_l C_l\leq C$.*

Definition 2. *Consider the closure of the convex hull of achievable rate-fronthaul tuples $(R_1,\ldots,R_K,C_1,\ldots,C_L)$ using the DDF strategy satisfying (31) over all joint distributions $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ satisfying possibly input constraints on $(x_1,\ldots,x_L)$. Define $\mathcal{R}^s_{DDF}(C)$ to be the projection of the above set along a sum fronthaul constraint C, i.e., the set of rate tuples $(R_1,\ldots,R_K)$ such that $C_l\geq 0$ and $\sum_l C_l\leq C$.*

Let us write down the two rate regions $\mathcal{R}^{\rm s}_{\rm COM}(C)$ and $\mathcal{R}^{\rm s}_{\rm DDF}(C)$ defined above more explicitly. For the compression strategy, under a fixed joint distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ and a fixed sum fronthaul constraint C, only the constraint for $\mathcal{S}=\mathcal{L}$ is active in (4). Therefore, the set of $(R_1,\ldots,R_K)$ that satisfies the sum fronthaul constraint C under a fixed joint distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ is described by the constraints

$$\sum_{k\in\mathcal{D}} R_k < \sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})) \tag{33}$$

$$C > I(U(\mathcal{K});X(\mathcal{L})) + T(X(\mathcal{L})) \tag{34}$$

over all $\mathcal{D}\subseteq\mathcal{K}$. The set $\mathcal{R}^{\rm s}_{\rm COM}(C)$ is then the projection of the closure of the convex hull of these $(R_1,\ldots,R_K,C)$ tuples, where the convex hull is taken over both the distributions as well as C.

Similarly, for the DDF strategy, under a fixed distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ and a fixed sum fronthaul constraint C, the active constraints are those corresponding to $\mathcal{S}=\emptyset$ for the case when the sum fronthaul is large enough to accommodate the compression of all BS signals (i.e., $C > I(U(\mathcal{D});X(\mathcal{L})) + T(X(\mathcal{L}))$), which corresponds to the Marton's region, or $\mathcal{S}=\mathcal{L}$ for the case when the sum fronthaul is not large enough to accommodate the compression of all BS signals; see [29] for a similar result. The set of $(R_1,\ldots,R_K)$ that satisfies the sum fronthaul constraint C under a fixed joint distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ is thus described by the constraints

$$\sum_{k\in\mathcal{D}} R_k < \sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})) \tag{35}$$

$$\sum_{k\in\mathcal{D}} R_k < \sum_{k\in\mathcal{D}} I(U_k;Y_k) + C - T(U(\mathcal{D}),X(\mathcal{L})) \tag{36}$$

$$= \sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})) + C - I(U(\mathcal{D});X(\mathcal{L})) - T(X(\mathcal{L})) \tag{37}$$

over all $\mathcal{D}\subseteq\mathcal{K}$. The set $\mathcal{R}^{\rm s}_{\rm DDF}(C)$ is then the projection of the closure of the the convex hull of the above $(R_1,\ldots,R_K,C)$ tuples, where the convex hull (i.e., time-sharing) is taken over both the distributions as well as C.

Theorem 3. *For the downlink C-RAN with a general DMC $p(y_1,\ldots,y_K|x_1,\ldots,x_L)$ in the second hop and a sum fronthaul constraint C, we have $\mathcal{R}^s_{COM}(C)=\mathcal{R}^s_{DDF}(C)$.*

We briefly explain the main ideas of the proof using an illustrative 2-BS 2-user example. Consider any given channel $p(y_1,y_2|x_1,x_2)$ in the second hop of C-RAN and a sum fronthaul constraint C. The rate region using the generalized compression strategy is given by

$$R_1 < I(U_1;Y_1) \tag{38}$$

$$R_2 < I(U_2;Y_2) \tag{39}$$

$$R_1+R_2 < I(U_1;Y_1)+I(U_2;Y_2)-I(U_1;U_2) \tag{40}$$

under joint distributions $p(u_1,u_2,x_1,x_2)$ that satisfy $I(U_1,U_2;X_1,X_2)+I(X_1;X_2)<C$. For the DDF strategy, the rate region under any fixed distribution $p(u_1,u_2,x_1,x_2)$ can be expressed as

$$R_1 < I(U_1,Y_1) + \min\left\{\begin{matrix}0,\\ C-I(U_1;X_1,X_2)-I(X_1;X_2)\end{matrix}\right\} \tag{41}$$

$$R_2 < I(U_2,Y_2) + \min\left\{\begin{matrix}0,\\ C-I(U_2;X_1,X_2)-I(X_1;X_2)\end{matrix}\right\} \tag{42}$$

$$R_1+R_2 < I(U_1,Y_1)+I(U_2,Y_2)-I(U_1;U_2) + \min\left\{\begin{matrix}0,\\ C-I(U_1,U_2;X_1,X_2)-I(X_1;X_2)\end{matrix}\right\}. \tag{43}$$

To show that the generalized compression and DDF regions coincide (after convex hull), we start with the DDF region under some fixed distribution $p(u_1, u_2, x_1, x_2)$. If the sum fronthaul capacity is such that $C > I(U_1, U_2; X_1, X_2) + I(X_1; X_2)$ under a fixed distribution $p(u_1, u_2, x_1, x_2)$, then both rate regions are exactly the same. The interesting case is when the distribution $p(u_1, u_2, x_1, x_2)$ is such that $C < I(U_1, U_2; X_1, X_2) + I(X_1; X_2)$, which is allowed under the DDF strategy but not under the generalized compression strategy. But, we show that by time-sharing across varying $p(u_1, u_2, x_1, x_2)$, (specifically, the original $p(u_1, u_2, x_1, x_2)$ and one with either of the users shut off), the DDF achievable rate region can nevertheless be achieved using time-sharing of the generalized compression strategies while satisfying an average fronthaul constraint. Intuitively, the penalty that the DDF strategy pays to go beyond the fronthaul capacity is at least as large as the penalty for shutting off the appropriate users. The proof for the general case of arbitrary number of users and BSs makes use of the polymatroidal structure of the rate region to characterize all the corner points of the rate region achieved by the DDF strategy and constructs appropriate time-shared compression strategies to achieve all such corner points. The full proof is relegated to Appendix C.

Since the DDF strategy is known to achieve the rate region of the C-RAN to within a constant gap for the Gaussian network, having the generalized compression rate region coincide with the DDF region under the sum fronthaul constraint immediately gives us the following corollary.

Corollary 2. *With Gaussian $p(y_1, \ldots, y_K|x_1, \ldots, x_L)$ on the second hop of the C-RAN model and under a sum fronthaul constraint C, the compression strategy achieves a rate region to within a constant gap to the capacity region. The gap is independent of the channel, the BS power constraints, and the sum fronthaul constraint, and only depends on the number of BSs and users.*

As a remark, we wonder whether the generalized compression and DDF rate regions coincide not just under the sum fronthaul constraint, but also individual fronthaul constraints. While the answer to this question is not yet clear, we note here that the successive coding strategy of computing Marton's codewords $(U_1, \ldots, U_K)$ first, then forming the compression codewords $(X_1, \ldots, X_L)$ is not the only way to perform successive encoding. There is also the possibility of breaking the encoding into more than two steps. As an example, consider a 2-BS 2-user C-RAN. The compression encoding order that we consider in this paper encodes (U_1, U_2) jointly first, and then (X_1, X_2) is computed. But it is possible to encode (U_1, X_1) first, and then encode (U_2, X_2). Such a reordering can potentially help user 2 because knowing the exact signals to be transmitted to user 1 can benefit the search for U_2 to align its correlation with U_1 to appropriately cancel the interference at user 2. Thus, interleaving in the encoding of U's and X's is likely needed in order to achieve the same rate region as DDF under arbitrary fronthaul constraints. However, as shown in the next section, if we only consider the sum rate, the two-step encoding of the generalized compression strategy indeed achieves the sum capacity of C-RAN to within a constant gap, even under individual fronthaul constraints, if we assume a Gaussian channel $p(y_1, \ldots, y_K|x_1, \ldots, x_L)$ and use a Gaussian $p(u_1, \ldots, u_K, x_1, \ldots, x_L)$ in the encoding.

It is worth pointing out that similar results exist for the uplink C-RAN. In the uplink, by comparing the joint decoding of quantized BS signals and user messages using noisy network coding versus the successive decoding of quantized signals followed by decoding of user messages, it is possible to establish that the successive decoding of quantized signals and user messages achieves the same sum rate as the noisy network coding strategy (see [29], and also [30] under an "oblivious" assumption), while successive decoding (that allows for interleaving within successive quantized signal decoding and user message decoding) achieves the same rate region as noisy network coding under a sum fronthaul constraint [29]. Just as in the downlink, it is still an open question as to whether successive decoding can match the noisy network coding rate region under arbitrary individual fronthaul constraints by considering all possible interleaving combinations across quantization and user messages, In fact, there is a duality between uplink and downlink C-RAN. It can be shown under the assumption of independent compression that the uplink and downlink compression strategies achieve exactly the same rate region for the C-RAN model [31]; a similar result is expected to hold under the multivariate compression. This suggests an even stronger connection between the generalized compression strategies in the uplink and the downlink C-RAN.

B. Sum Rate Under Individual Fronthaul Constraints

In this section, we consider the general case of individual fronthaul constraints instead of restricting to the sum fronthaul constraint as in the previous section. However, we focus on the sum rate only, and aim to find the approximate sum capacity of C-RAN under arbitrary fronthaul constraints. The main result of this section is that under a Gaussian C-RAN model, the generalized compression strategy can achieve a sum rate which is within a constant gap to the cut-set bound of C-RAN under individual fronthaul constraints. More precisely, consider the Gaussian C-RAN model specified in (2). For the Gaussian channel, recall that if we set the distribution $p(u_1, \ldots, u_K, x_1, \ldots, x_L)$ according to (32), the DDF strategy can be shown to achieve to within a constant gap to the capacity region of the Gaussian C-RAN. For convenience, we call the distribution in (32) the *constant-gap distribution*. We show in this section that the sum rate achieved by the DDF strategy for the Gaussian C-RAN under the constant-gap distribution can also be achieved using the generalized compression strategy under the same set of fronthaul constraints.

For each fixed distribution, we can write down the achievable sum rate of the DDF and the generalized compression strategies explicitly. The sum rate achieved by the DDF strategy is given by R that satisfies

$$R < \sum_{k \in \mathcal{K}} I(U_k; Y_k) + \sum_{l \in \mathcal{S}} C_l - T(U(\mathcal{K}), X(\mathcal{S})), \qquad (44)$$

for all $\mathcal{S} \subseteq \mathcal{L}$ under some distribution $p(u_1, \ldots, u_K, x_1, \ldots, x_L)$. The sum rate achieved by

the generalized compression strategy is given by R that satisfies

$$R < \sum_{k\in\mathcal{K}} I(U_k;Y_k) - T(U(\mathcal{K})) \tag{45}$$

$$\sum_{l\in\mathcal{S}} C_l > I(U(\mathcal{K});X(\mathcal{S})) - T(X(\mathcal{S})), \tag{46}$$

for all $\mathcal{S} \subseteq \mathcal{L}$ under some joint distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$.

Definition 3. *Consider the closure of the convex hull of achievable sum-rate-fronthaul tuples $(R,C_1,\ldots,C_L)$ for the C-RAN with the Gaussian channel model (2) using the DDF strategy as expressed in (44) under the constant-gap distribution (32) with the BS powers constrained by the power constraint P. Define R^g_{DDF} to be the maximum sum rate under individual fronthaul constraints $(C_1,\ldots,C_L)$ in this set.*

Definition 4. *Consider the closure of the convex hull of achievable sum-rate-fronthaul tuples $(R,C_1,\ldots,C_L)$ of the C-RAN with the Gaussian channel model (2) using the generalized compression strategy (45)-(46) under the constant under the constant-gap distribution (32) with the BS powers constrained by the power constraint P. Define R^g_{COM} to be the maximum sum rate under individual fronthaul constraints $(C_1,\ldots,C_L)$ in this set.*

Comparing the sum rate of DDF in (44) with the sum rate of generalized compression in (45)-(46), we clearly have $R^{\rm g}_{\rm COM} \le R^{\rm g}_{\rm DDF}$. We show in this section that actually $R^{\rm g}_{\rm COM} = R^{\rm g}_{\rm DDF}$. As a consequence, we have the following main theorem of this section.

Theorem 4. *For the downlink C-RAN with a memoryless Gaussian channel on the second hop, the compression scheme achieves a sum rate to within a constant gap to the cut-set bound under individual fronthaul constraints $(C_1,\ldots,C_L)$. The gap is independent of the channel parameters, the BS power constraints, and the individual fronthaul constraints, and only depends on the number of BSs and users.*

We briefly explain the key ideas of the proof again using the illustrative 2-BS 2-user example. Under a fixed distribution $p(u_1,u_2,x_1,x_2)$, the sum rate achieved by the DDF strategy is given by

$$\begin{aligned} R < {} & I(U_1;Y_1) + I(U_2;Y_2) - I(U_1;U_2) \\ & + \min\left\{\begin{array}{l} 0, \\ C_1 - I(U_1,U_2;X_1), \\ C_2 - I(U_1,U_2;X_2), \\ C_1 + C_2 - I(U_1,U_2;X_1,X_2) - I(X_1;X_2) \end{array}\right\}. \end{aligned} \tag{47}$$

Likewise, for the generalized compression strategy, the sum rate is given by

$$R < I(U_1;Y_1) + I(U_2;Y_2) - I(U_1;U_2) \tag{48}$$

$$C_1 > I(U_1,U_2;X_1) \tag{49}$$

$$C_2 > I(U_1,U_2;X_2) \tag{50}$$

$$C_1 + C_2 > I(U_1,U_2;X_1,X_2) + I(X_1;X_2). \tag{51}$$

Clearly, if the fronthaul capacity constraints C_1 and C_2 are such that under the distribution $p(u_1,u_2,x_1,x_2)$, the fronthaul constraints (49)-(51) for the compression strategy are all satisfied, then the sum rate for the compression strategy is exactly equal to that of the DDF strategy. However, if either C_1 or C_2 or both are not large enough so that some of the fronthaul constraints are violated, then the DDF strategy can still provide an achievable rate-tuple, but the sum rate would be smaller than $I(U_1;Y_1)+I(U_2;Y_2)-I(U_1;U_2)$ by a penalty term equal to how large the maximum violation in the three fronthaul constraints is. For the compression strategy, however, whenever the fronthaul constraints are not satisfied, we can no longer use the distribution $p(u_1,u_2,x_1,x_2)$ directly. The idea of the proof is that we can modify the distribution (specifically, by time-sharing between the original $p(u_1,u_2,x_1,x_2)$ and that with one of the users turned off), so that under the new distribution, we stay within the allowed fronthaul constraint and achieve a sum rate $I(U_1;Y_1) + I(U_2;Y_2) - I(U_1;U_2)$ that is at least as large as the penalized sum rate of the DDF strategy. The proof for the general case of arbitrary number of users and BSs uses the contra-polymatroidal structure of the fronthaul region to characterize the corner points for each fixed sum rate R under the DDF strategy. Using appropriate time-sharing schemes in the generalized compression strategy, we show that each such corner point is achievable in the compression strategy with a sum rate at least as large as R. The complete proof is relegated to Appendix D.

As a final remark, we mention the work of [22], which shows that the gap between the achievable rate region and the cut-set bound for DDF can be refined, so that the gap is logarithmic in the number of BSs and users, instead of being linear as in [21]. The refinement uses a slightly modified form of the constant-gap distribution. The equivalence result shown in this section also works for this modified constant-gap distribution. Thus, a similar refinement can be used to conclude that the compression strategy can achieve the sum capacity of the C-RAN network to within a constant gap which is logarithmic in the number of BSs and users. A different improvement in the gap for the DDF strategy is proved in [11], and is also applicable to our result.

C. Sum Rate Under Sum Fronthaul Constraint

The previous two sections show that even though the DDF strategy allows for distributions $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ that can compress beyond the fronthaul constraints, under certain conditions, the compression strategy, which compresses within the fronthaul constraints, can achieve the same rate-region or the same sum rate if we allow time-sharing between different achievable rate tuples of the compression strategy. Applying this result to the Gaussian C-RAN gives us the conclusion that time-sharing of compression strategies can achieve the sum capacity of Gaussian C-RAN to within a constant gap. This section provides a slightly stronger statement. We show that for maximizing the *sum* rate under the *sum* fronthaul constraint, there exists a Gaussian compression strategy that achieves to within a constant gap of the cut-set bound even *without* time-sharing.

Let us first write down the achievable sum rates for the DDF and compression strategies under a sum fronthaul constraint C. From (35) and (37), we have that the achievable sum rate R^s_{DDF} for the DDF strategy under the sum fronthaul constraint is

$$R^s_{\text{DDF}} < \sum_{k\in\mathcal{K}} I(U_k;Y_k) - T(U(\mathcal{K})) + \min\left\{\begin{matrix}0,\\ C - I(U(\mathcal{K});X(\mathcal{L})) - T(X(\mathcal{L}))\end{matrix}\right\}, \quad (52)$$

for some distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$. Similarly, from (33), the achievable sum rate R^s_{COM} using the compression strategy under the sum fronthaul constraint is given by

$$R^s_{\text{COM}} < \sum_{k\in\mathcal{K}} I(U_k;Y_k) - T(U(\mathcal{K})), \quad (53)$$

for some distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ that satisfies

$$C > I(U(\mathcal{K});X(\mathcal{L})) + T(X(\mathcal{L})). \quad (54)$$

Consider the channel model (2). We know that the DDF strategy can achieve to within a constant gap to the capacity region (and hence the sum capacity) of this Gaussian C-RAN model under individual fronthaul constraints (and hence also the sum fronthaul constraint) by using the distribution given by (32). We now show that by using a (possibly) modified version of this distribution, the compression strategy can achieve the same sum rate under the sum fronthaul constraint.

We consider two cases. If under the distribution given in (32), we have

$$C \geq I(U(\mathcal{K});X(\mathcal{L})) + T(X(\mathcal{L})) \quad (55)$$

$$= \frac{1}{2}\log|\mathbf{I} + P\mathbf{H}\mathbf{H}^T|, \quad (56)$$

then we can simply use the same distribution in the compression strategy to achieve the same rate. If $C < \frac{1}{2}\log|\mathbf{I} + P\mathbf{H}\mathbf{H}^T|$, we propose to modify the distribution in (32) in such a way that when used in the compression strategy, the fronthaul constraint is satisfied, and further, it achieves a higher sum rate than the DDF strategy. The proposed modification is to reduce the power of X's by a factor $\gamma < 1$. We find γ such that

$$C = \frac{1}{2}\log|\mathbf{I} + \gamma P\mathbf{H}\mathbf{H}^T|. \quad (57)$$

This allows us to compress with the same sum fronthaul rate as the DDF strategy. To show that compression with this modified distribution actually improves upon the DDF sum rate, we compare the sum rate achieved by the compression strategy to that with DDF under the modified distribution as follows:

$$R^s_{\text{COM}} = \sum_{k\in\mathcal{K}} I(U'_k;Y'_k) - T(U'(\mathcal{K})) \quad (58)$$

$$\overset{(a)}{=} I(U'(\mathcal{K});X'(\mathcal{L})) - \sum_{k\in\mathcal{K}} I(U'_k;X'(\mathcal{L})|Y'_k) \quad (59)$$

$$= C - \sum_{k\in\mathcal{K}} \frac{1}{2}\log\left(1 + \frac{\sum_{l=1}^L h^2_{k,l}\gamma P}{\sum_{l=1}^L h^2_{k,l}\gamma P + \sigma^2}\right) \quad (60)$$

$$> C - \sum_{k\in\mathcal{K}} \frac{1}{2}\log\left(1 + \frac{\sum_{l=1}^L h^2_{k,l} P}{\sum_{l=1}^L h^2_{k,l} P + \sigma^2}\right) \quad (61)$$

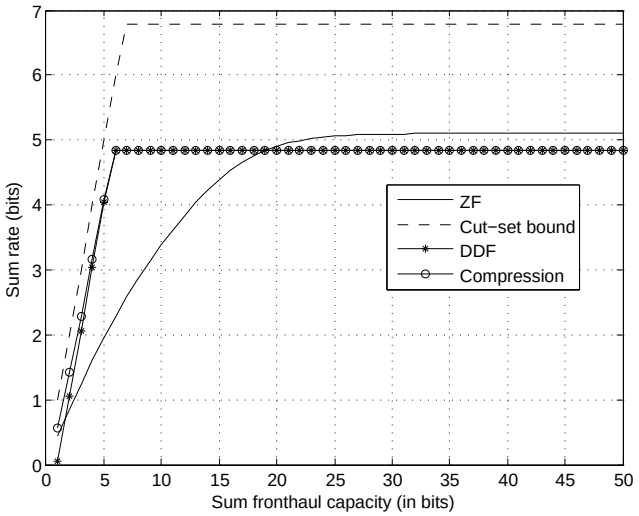

Fig. 2. Sum rate comparison of the zero-forcing, DDF, and compression strategies, along with the cut-set outer bound. Note that the DDF and generalized compression strategies are evaluated under their respective distributions that achieve to within a constant gap to capacity for each strategy. In particular, they are not evaluated under their respective optimal distributions, so the numerical result does not imply that the compression strategy outperforms DDF.

$$\overset{(b)}{\geq} I(U(\mathcal{K});X(\mathcal{L})) - \sum_{k\in\mathcal{K}} I(U_k;X(\mathcal{L})|Y_k) \quad (62)$$

$$\overset{(c)}{=} R^s_{\text{DDF}} \quad (63)$$

where (a) and (c) are due to Lemma 5 in Appendiex D, and (b) is due to (55). Since the DDF sum rate achieves to within a constant gap to the sum capacity of C-RAN, the above shows that this choice of (non-time-shared) distribution also achieves to within a constant gap to the sum capacity.

D. Numerical Example

We provide a numerical example that illustrates a performance comparison between the compression and DDF strategies, along with the more traditional beamforming based strategy that takes fronthaul into account.

Consider the Gaussian channel (2) with 2 BSs and 2 users. As a baseline, we consider beamforming followed by compression, where we use the zero-forcing beamformers in the directions of $\mathbf{W} = \mathbf{H}^{-1}$. We then allocate powers across the two normalized beams $\mathbf{w}_1$ and $\mathbf{w}_2$, and also find the quantization noise levels to maximize the sum rate by exhaustive search, while satisfying the sum fronthaul constraint. For the DDF strategy, the sum rate is calculated as given in (52). For the generalized compression scheme, we construct the explicit distribution that achieves the sum capacity to within a constant gap as explained in the previous section. The parameter γ is found using a line search between $[0,1]$.

Fig. 2 shows the sum rate achieved using these strategies as a function of the sum fronthaul capacity available for a fixed real-valued channel (generated at random according to a Rayleigh fading distribution). Individual BS power constraints of $P = 100$ and background noise $\sigma^2 = 1$ are assumed. For comparison, we also plot the cut-set bound. The figure shows

that the compression strategy performs nearly the same as the DDF strategy for most of the sum fronthaul capacity range. It performs slightly better than the DDF strategy at very low sum fronthaul capacities because of the improvement in the gap as a result of choosing $\gamma < 1$ to accommodate the fronthaul capacity as shown from (58) to (63). We note that both the compression and the DDF strategies are within a constant gap to the cut-set bound. When $C < \frac{1}{2}\log|\mathbf{I}+P\mathbf{H}\mathbf{H}^T|$, the gap is at most 1 bit, while for higher values of C, the gap is at most 2 bits. As compared to the zero-forcing strategy, we observe that the generalized compression strategy performs much better at lower sum fronthaul capacities, because the zero-forcing beam direction does not account for the quantization noise. At higher sum fronthaul capacities, all three strategies saturate. However, the zero-forcing strategy may achieve a higher sum rate than the compression and DDF strategies, because the latter does not explicitly null interference, but only aims to provide a universal strategy that approximately achieves the cut-set bound for all values of the sum fronthaul constraint. The choice of $\mathbf{U} = \mathbf{H}\mathbf{X} + \tilde{\mathbf{Z}}$ for the channel $\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}$ is in a sense trying to invert the channel, but not exactly.

Even though the generalized compression strategy achieves to within a constant gap under certain conditions, it is important to note that the gap depends on the number of users and BSs, at least logarithmically. Therefore, the data-sharing strategies might perform better than the generalized compression strategy, especially in the low-power regime or when the channel matrix is ill-conditioned; see [11] and [18] for some numerical evidence along these lines.

VI. Conclusion

This paper investigates the compression strategy for the downlink of a C-RAN from an information theoretic point of view. The paper first generalizes the existing compression strategies to include Marton's multicoding followed by multivariate compression, then analyzes the resulting rate region for a C-RAN with a general DMC between the BSs and the users. When compared with the DDF strategy specialized to the downlink C-RAN, it is pointed out that DDF is a generalization of the compression strategy where the Marton's multicoding and the multivariate compression are done jointly as opposed to successively in the compression strategy. The paper then shows that under a sum fronthaul constraint, such generalization does not lead to higher rates and the rate regions of the two strategies coincide. Thus, for the Gaussian C-RAN under a sum fronthaul constraint, the compression strategy already achieves the capacity region to within a constant gap. Furthermore, for the Gaussian C-RAN under individual fronthaul constraints, the paper shows that the two-phase compression strategy can achieve a sum rate that is within a constant gap to the cut-set bound. These results provide a justification for the practical choice of the two-phase compression strategy for the downlink C-RAN.

Appendix A
Proof of Theorem 1

We provide a proof sketch by first describing the coding scheme, then establishing the conditions on the achievable rates for vanishing probability of error $P_e^{(n)}$.

Fix the distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$. Let $\epsilon > 0$.

1) Codebook generation: We generate a random codebook for Marton's multicoding according to $u_k^n(m_k,l_k) \sim \prod_{i=1}^n p_{U_k}(u_{ki})$ for $(m_k,l_k) \in [1:2^{nR_k}] \times [1:2^{n\tilde{R}_k}]$, $k \in \mathcal{K}$. Similarly, we generate a random codebook for multivariate compression according to $x_l^n(t_l) \sim \prod_{i=1}^n p_{X_l}(x_{li})$ for $t_l \in [1:2^{nC_l}]$, $l \in \mathcal{L}$.
2) Encoding at the CP: To send $[m_1 : m_K]$, we find $[l_1 : l_K]$ such that $[u_1^n(m_1,l_1) : u_K^n(m_K,l_K)] \in \mathcal{T}_\epsilon^{(n)}$. Then, we find $[t_1 : t_L]$ such that $[u_1^n(m_1,l_1) : u_K^n(m_K,l_K), x_1^n(t_1) : x_L^n(t_L)] \in \mathcal{T}_\epsilon^{(n)}$. Finally, we forward t_l to BS l.
3) Mapping at the BSs: BSs transmit $x_l^n(t_l)$ to users.
4) Decoding at the users: User k finds $(\hat{m}_k,\hat{l}_k)$ such that $(u_k^n(\hat{m}_k,\hat{l}_k),y_k^n) \in \mathcal{T}_\epsilon^{(n)}$.

In order to show that the average probability of error $P_e^{(n)}$ for the coding scheme vanishes as $n \to \infty$, we analyze three sources of error. For encoding at the central processor, we can find the incides $[l_1 : l_K]$ correctly with high probability if $\sum_{k\in\mathcal{D}} \tilde{R}_k > T(U(\mathcal{D}))$ due to the multivariate covering lemma [24, Lemma 14.1]. Similarly, we can find the indices $[t_1 : t_L]$ correctly with high probability if $\sum_{l\in\mathcal{S}} C_l > I(U(\mathcal{K});X(\mathcal{S})) + T(X(\mathcal{S}))$. Finally, the decoding at the user side is successful with high probability if $R_k + \tilde{R}_k < I(U_k;Y_k)$ due to the joint typicality lemma [24, p. 29]. Using the Fourier-Motzkin elimination, we project out the auxiliary rates $\tilde{R}_k$ to obtain required the rate region.

Appendix B
Proof of Theorem 2

We specialize the DDF coding scheme to the C-RAN model as follows.

Fix the distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$. Let $\epsilon > 0$.

1) Codebook generation: We generate a random codebook for Marton's multicoding according to $u_k^n(m_k,l_k) \sim \prod_{i=1}^n p_{U_k}(u_{ki})$ for $(m_k,l_k) \in [1:2^{nR_k}] \times [1:2^{n\tilde{R}_k}]$, $k \in \mathcal{K}$. Similarly, we generate a random codebook for multivariate compression according to $x_l^n(t_l) \sim \prod_{i=1}^n p_{X_l}(x_{li})$ for $t_l \in [1:2^{nC_l}]$, $l \in \mathcal{L}$.
2) Encoding at the CP: To send $[m_1 : m_K]$, we find $[l_1 : l_K, t_1 : t_L]$ such that $[u_1^n(m_1,l_1) : u_K^n(m_K,l_K), x_1^n(t_1) : x_L^n(t_L)] \in \mathcal{T}_\epsilon^{(n)}$.
3) Mapping at the BSs: BSs transmit $x_l^n(t_l)$ to users.
4) Decoding at the users: User k finds $(\hat{m}_k,\hat{l}_k)$ such that $(u_k^n(\hat{m}_k,\hat{l}_k),y_k^n) \in \mathcal{T}_\epsilon^{(n)}$.

Similar to the probability of error analysis in the compression strategy, to show that the average probability of error $P_e^{(n)}$ for the coding scheme vanishes as $n \to \infty$, we analyze two sources of error. For encoding at the CP, we can find the indices $[l_1 : l_K, t_1 : t_L]$ correctly with high probability if $\sum_{k\in\mathcal{D}} \tilde{R}_k + \sum_{l\in\mathcal{S}} C_l > T(U(\mathcal{D}),X(\mathcal{S}))$, due to the multivariate covering lemma [24, Lemma 14.1]. The decoding at the user side is successful with high probability if $R_k + \tilde{R}_k < I(U_k;Y_k)$ due to the joint typicality lemma [24, p. 29]. Combining the two, we obtain the required rate region.

APPENDIX C
PROOF OF THEOREM 3

We examine the set of achievable rate tuples $(R_1, \ldots, R_K)$ of the generalized compression and the DDF strategies under a sum fronthaul constraint C. Since the compression strategy is a special case of the DDF strategy, we have that $\mathcal{R}^{\text{s}}_{\text{COM}}(C) \subseteq \mathcal{R}^{\text{s}}_{\text{DDF}}(C)$. The main part of the proof is to show that $\mathcal{R}^{\text{s}}_{\text{DDF}}(C) \subseteq \mathcal{R}^{\text{s}}_{\text{COM}}(C)$. The proof uses properties of submodular optimization.

Take any achievable $(R_1, \ldots, R_K)$ using the DDF strategy under a fixed distribution $p(u_1, \ldots, u_K, x_1, \ldots, x_L)$ and under the fixed sum fronthaul constraint C. By definition, it must satisfy the inequalities (35) and (36). Define $\underline{\mathcal{P}}(C) \in \mathbb{R}^K$ to be the polytope formed by the inequalities (35) and (36). We show that each extreme point of $\underline{\mathcal{P}}(C)$ can be achieved using the time-sharing of rate tuples under the generalized compression strategy.

The inequalities (35)-(36) define $\underline{\mathcal{P}}(C)$ to be set of $(R_1, \ldots, R_K)$ for which

$$\sum_{k\in\mathcal{D}} R_k \leq \min \left\{ \begin{array}{l} \displaystyle\sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})), \\ \displaystyle\sum_{k\in\mathcal{D}} I(U_k;Y_k) + C - T(U(\mathcal{D}), X(\mathcal{L})) \end{array} \right\} \tag{64}$$

for all $\mathcal{D} \subseteq \mathcal{K}$. First, we show that we can alternatively write the above as

$$\sum_{k\in\mathcal{D}} R_k \leq \min \left\{ \begin{array}{l} \displaystyle\sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})), \\ \displaystyle\sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}), X(\mathcal{L})) \end{array} \right\} \tag{65}$$

for all $\mathcal{D} \subseteq \mathcal{K}$. The reason is that for any set $\mathcal{D} \subseteq \mathcal{K}$, we always have $\sum_{k\in\mathcal{D}} R_k \leq \sum_{k\in\mathcal{K}} R_k$, since the user rates are non-negative. But we already have the constraint

$$\sum_{k\in\mathcal{K}} R_k \leq \sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}), X(\mathcal{L})). \tag{66}$$

So, we can add the constraint

$$\sum_{k\in\mathcal{D}} R_k \leq \sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}), X(\mathcal{L})) \tag{67}$$

to (64) without affecting $\underline{\mathcal{P}}(C)$. Now, it turns out that

$$\sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}), X(\mathcal{L})) \leq \sum_{k\in\mathcal{D}} I(U_k;Y_k) + C - T(U(\mathcal{D}), X(\mathcal{L})), \tag{68}$$

so this new constraint is actually tighter than the second constraint in (64). Therefore, (64) can be equivalently written as (65).

To verify (68), we take the difference in summing over $\mathcal{D}$ versus summing over $\mathcal{K}$ in (68) as below:

$$\begin{aligned} &T(U(\mathcal{K}), X(\mathcal{L})) - T(U(\mathcal{D}), X(\mathcal{L})) - \sum_{k\in\mathcal{D}^c} I(U_k;Y_k) \\ &\quad = I(U(\mathcal{D}^c); U(\mathcal{D})) + T(U(\mathcal{D}^c)) \\ &\qquad + I(U(\mathcal{D}^c); X(\mathcal{L})|U(\mathcal{D})) - \sum_{k\in\mathcal{D}^c} I(U_k;Y_k), \end{aligned} \tag{69}$$

where $\mathcal{D}^c = \mathcal{K} \setminus \mathcal{D}$. This can be simplified as

$$\sum_{k\in\mathcal{D}^c} h(U_k|Y_k) - h(U(\mathcal{D}^c)|X(\mathcal{L}), U(\mathcal{D})) \tag{70}$$

$$\stackrel{(a)}{\geq} \sum_{k\in\mathcal{D}^c} h(U_k|Y_k) - \sum_{k\in\mathcal{D}^c} h(U_k|X(\mathcal{L})) \tag{71}$$

$$\stackrel{(b)}{\geq} \sum_{k\in\mathcal{D}^c} h(U_k|Y_k) - \sum_{k\in\mathcal{D}^c} h(U_k|X(\mathcal{L}), Y_k) \tag{72}$$

$$= \sum_{k\in\mathcal{D}^c} I(U_k, X(\mathcal{L})|Y_k) \tag{73}$$

$$\geq 0, \tag{74}$$

where (a) follows from the fact that conditioning reduces entropy and (b) follows since $U_k \to X(\mathcal{L}) \to Y_k$ form a Markov chain. This verifies (68), hence the equivalence between (64) and (65).

Let us now define a set function $f : 2^{\mathcal{K}} \to \mathbb{R}$ as

$$f(\mathcal{D}) := \min \left\{ \begin{array}{l} \displaystyle\sum_{k\in\mathcal{D}} I(U_k;Y_k) - T(U(\mathcal{D})), \\ \displaystyle\sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}), X(\mathcal{L})) \end{array} \right\} \tag{75}$$

for each $\mathcal{D} \subseteq \mathcal{K}$. By construction, $\underline{\mathcal{P}}(C)$ is the set of $(R_1, \ldots, R_K)$ that satisfies

$$\sum_{k\in\mathcal{D}} R_k \leq f(\mathcal{D}). \tag{76}$$

Since the second term in the $\min$ expression in (75) is a constant that does not depend of $\mathcal{D}$, it can be verified that the function f is a submodular function [29], if the Marton's region is a polymatroid (which we assume in this paper). This allows the rate region $\underline{\mathcal{P}}(C)$ to have a polymatroid structure. We remark that, although the Marton's region may not be polymatroid in general, for the constant gap Gaussian distribution, we can guarantee a certain monotone property of the Marton's rate expression by appropriate choice of the noise variance leading to a polymatroid rate region.

A result in submodular optimization [32] is that for a linear ordering $i_1 \prec i_2 \prec \ldots \prec i_K$ of $\{1, \ldots, K\}$, an extreme point of $\underline{\mathcal{P}}(C)$ can be greedily computed as $(R_1, \ldots, R_K)$ where

$$R_{i_j} = f(\{i_1, \ldots, i_j\}) - f(\{i_1, \ldots, i_{j-1}\}). \tag{77}$$

Moreover, all extreme points of $\underline{\mathcal{P}}(C)$ can be enumerated by considering all linear orderings. Since each ordering of $\{1, \ldots, K\}$ is analyzed in the same manner, for notational simplicity, we consider the natural ordering $i_j = j$.

Let j be the first index for which

$$\sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}),X(\mathcal{L})) < \sum_{k=1}^{j} I(U_k;Y_k) - T(U_1,\ldots,U_j). \quad (78)$$

Then, by construction, $\forall k < j$

$$\begin{aligned} R_k &= \sum_{i=1}^{k} I(U_i;Y_i) - T(U_1,\ldots,U_k) \\ &\quad - \sum_{i=1}^{k-1} I(U_i;Y_i) - T(U_1,\ldots,U_{k-1}) && (79)\\ &= I(U_k;Y_k) - I(U_k;U_{k-1},\ldots,U_1). && (80)\end{aligned}$$

Furthermore, using the fact that the second term of $f(\mathcal{D})$ does not depend on $\mathcal{D}$, so when the second term is the minimum, i.e., $\forall k > j$, we have

$$R_k = 0. \quad (81)$$

Finally, we express R_j as

$$\begin{aligned} R_j &= \sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}),X(\mathcal{L})) \\ &\quad - \sum_{k=1}^{j-1} I(U_k;Y_k) - T(U_1,\ldots,U_{j-1}) && (82)\\ &= I(U_j;Y_j) - I(U_j;U_{j-1},\ldots,U_1) \\ &\quad + \sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}),X(\mathcal{L})) \\ &\quad - \sum_{k=1}^{j} I(U_k;Y_k) + T(U_1,\ldots,U_j) && (83)\\ &= (1-\alpha)\left(I(U_j;Y_j) - I(U_j;U_{j-1},\ldots,U_1)\right), && (84)\end{aligned}$$

where α can be written explicitly as below

$$\frac{T(U(\mathcal{K}),X(\mathcal{L})) - T(U_1,\ldots,U_j) - C - \sum_{k=j+1}^{K} I(U_k;Y_k)}{I(U_j;Y_j) - I(U_j;U_{j-1},\ldots,1)}. \quad (85)$$

Note that $\alpha \in (0,1]$ due to (78) and the fact that $R_j \geq 0$.

Now, we construct the time-sharing of two rate tuples corresponding to the generalized compression strategy (33) that achieves this above rate $(R_1,\ldots,R_K)$ as follows:

1) For $(1-\alpha)$ fraction of the time, transmit messages for users $1,\ldots,j$, only, i.e., set $p(u_1,\ldots,u_j,x_1,\ldots,x_L)$ to be the marginal distribution of the original distribution, but let $U_{j+1},\ldots,U_K$ be constants.
2) For the rest α fraction of time, transmit messages for users $1,\ldots,(j-1)$ only, i.e., set $p(u_1,\ldots,u_{j-1},x_1,\ldots,x_L)$ to be the marginal distribution of the original distribution, but let $U_j,\ldots,U_K$ be constants.

By construction, the time-sharing of these two compression schemes achieves the same rate tuple as the extreme point of $\underline{\mathcal{P}}(C)$, (80), (81), and (84).

To calculate the fronthaul capacity consumption of this time-sharing scheme, we have

$$\begin{aligned} \bar{C} &= (1-\alpha)\left(I(U_1,\ldots,U_j;X(\mathcal{L})) + T(X(\mathcal{L}))\right) \\ &\quad \alpha\left(I(U_1,\ldots,U_{j-1};X(\mathcal{L})) + T(X(\mathcal{L}))\right) && (86)\\ &= I(U_1,\ldots,U_j;X(\mathcal{L})) + T(X(\mathcal{L})) \\ &\quad - \alpha(I(U_1,\ldots,U_j;X(\mathcal{L})) - I(U_1,\ldots,U_{j-1};X(\mathcal{L}))) && (87)\\ &= I(U_1,\ldots,U_j;X(\mathcal{L})) + T(X(\mathcal{L})) - C \\ &\quad - \frac{I(U_1,\ldots,U_j;X(\mathcal{L})) - I(U_1,\ldots,U_{j-1};X(\mathcal{L}))}{I(U_j;Y_j) - I(U_j;U_{j-1},\ldots,U_1)} \\ &\quad \cdot \Bigg(T(U(\mathcal{K}),X(\mathcal{L})) - T(U_1,\ldots,U_j) - C \\ &\quad - \sum_{k=j+1}^{K} I(U_k;Y_k) \Bigg) + C && (88)\\ &\overset{(a)}{\leq} I(U_1,\ldots,U_j;X(\mathcal{L})) + T(X(\mathcal{L})) - C \\ &\quad - T(U(\mathcal{K}),X(\mathcal{L})) + T(U_1,\ldots,U_j) + C \\ &\quad + \sum_{k=j+1}^{K} I(U_k;Y_k) + C && (89)\\ &= \left(\sum_{k\in\mathcal{K}} I(U_k;Y_k) + C - T(U(\mathcal{K}),X(\mathcal{L}))\right) \\ &\quad - \left(\sum_{k=1}^{j} I(U_k;Y_k) + C - T(U_1,\ldots,U_j,X(\mathcal{L}))\right) \\ &\quad + C && (90)\\ &\overset{(b)}{\leq} C. && (91)\end{aligned}$$

The inequality (a) follows because

$$\begin{aligned} &I(U_1,\ldots,U_j;X(\mathcal{L})) - I(U_1,\ldots,U_{j-1};X(\mathcal{L})) \\ &\quad - I(U_j;Y_j) + I(U_j;U_{j-1},\ldots,U_1) && (92)\\ &= h(U_j|Y_j) - h(U_j|X(\mathcal{L}),U_{j-1},\ldots,U_1) && (93)\\ &\geq h(U_j|Y_j) - h(U_j|X(\mathcal{L})) && (94)\\ &= h(U_j|Y_j) - h(U_j|X(\mathcal{L}),Y_j) && (95)\\ &= I(U_j;X(\mathcal{L})|Y_j) && (96)\\ &\geq 0, && (97)\end{aligned}$$

where we used the fact that conditioning reduces entropy and that $U_j \to X(\mathcal{L}) \to Y_j$ forms a Markov chain. Intuitively, this holds because the contribution of U_j to the user rate is less than the fronthaul required to support U_j. Note that the term $T(U(\mathcal{K}),X(\mathcal{L})) - T(U_1,\ldots,U_j) - C - \sum_{k=j+1}^{K} I(U_k;Y_k)$ is positive from the assumption in (78). The inequality (b) follows from (68).

Therefore, every extreme point $(R_1,\ldots,R_K)$ of $\underline{\mathcal{P}}(C)$ is achievable using time-sharing of generalized compression strategies under the same average fronthaul constraint.

APPENDIX D
PROOF OF THEOREM 4

The proof is based on comparing the sum rate achieved by the compression strategy with that by the DDF strategy. Recall that, from the result in [21], for the DDF strategy the following choice of the distribution achieves to within a constant gap to the cut-set bound of a Gaussian relay broadcast network: Let $\mathbf{X}$ to be a vector of L i.i.d. $\mathcal{N}(0,P)$ random variables and $\mathbf{U}=\mathbf{H}\mathbf{X}+\tilde{\mathbf{Z}}$, where $\tilde{\mathbf{Z}}\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ is independent of $\mathbf{Z}$. We show that under such a choice of distribution $R^{\mathrm{g}}_{\mathrm{DDF}}=R^{\mathrm{g}}_{\mathrm{COM}}$. Then it follows that the compression strategy also achieves the sum rate to within a constant gap to the cut-set bound.

Consider the set of $(R,C_1,\ldots,C_L)$ achievable using the DDF strategy under such a constant-gap distribution. For fixed R, we define $\overline{\mathcal{P}}(R)\subseteq\mathbb{R}^L$ to be the polytope defined by inequalities (44) under the said distribution. We now show that each extreme point of $\overline{\mathcal{P}}(R)$ is dominated by some time-sharing of points in the compression region.

Let us define a set function $g:2^{\mathcal{L}}\to\mathbb{R}$ as

$$g(\mathcal{S}):=\max\left\{\begin{matrix}T(U(\mathcal{K}),X(\mathcal{S}))+R-\sum_{k\in\mathcal{K}}I(U_k;Y_k),\\ 0\end{matrix}\right\} \tag{98}$$

for each $\mathcal{S}\subseteq\mathcal{L}$. By construction, then $\overline{\mathcal{P}}(R)$ is equal to the set of $(C_1,\ldots,C_L)$ that satisfy

$$\sum_{l\in\mathcal{S}}C_l\geq g(\mathcal{S}). \tag{99}$$

Since the second term in the max expression in (98) is a constant, it can be verified that the function g is a supermodular function [33] and as a consequence the $\overline{\mathcal{P}}(R)$ region is a contra-polymatroid [25]. Similar to the case of submodular optimization, for a linear ordering $i_1\prec i_2\prec\ldots\prec i_K$ of $\{1,\ldots,K\}$, an extreme point of $\overline{\mathcal{P}}(R)$ can be greedily computed as

$$C_{i_j}=g(\{i_1,\ldots,i_j\})-g(\{i_1,\ldots,i_{j-1}\}). \tag{100}$$

Furthermore, all the extreme points of $\overline{\mathcal{P}}(R)$ can be computed by considering all linear orderings. Each ordering of $\{1,\ldots,K\}$ is analyzed in the same manner, hence for notational simplicity we consider the natural ordering $i_j=j$.

Let j be the first index for which $C_j>0$. Then, by construction,

$$C_l=0,\qquad\forall l<j \tag{101}$$

and

$$C_l=I(X_l;U(\mathcal{K})|X_{l-1},\ldots,X_1),\quad\forall l>j. \tag{102}$$

Note that the term $T(X(\mathcal{S}))$ vanishes because of the assumption of independence of X's in the constant-gap distribution. Finally, we express C_j as

$$\begin{aligned}C_j&=I(X_j,\ldots,X_1;U(\mathcal{K}))\\&\quad+R-\sum_{k\in\mathcal{K}}I(U_k;Y_k)+T(U(\mathcal{K}))&(103)\\&=I(X_j;U(\mathcal{K})|X_{j-1},\ldots,X_1)+I(X_{j-1},\ldots,X_1;U(\mathcal{K}))\\&\quad+R-\sum_{k\in\mathcal{K}}I(U_k;Y_k)+T(U(\mathcal{K}))&(104)\\&=(1-\beta)I(X_j;U(\mathcal{K})|X_{j-1},\ldots,X_1),&(105)\end{aligned}$$

where β is defined as

$$\frac{-\left(I(X_{j-1},\ldots,X_1;U(\mathcal{K}))+R-\sum_{k\in\mathcal{K}}I(U_k;Y_k)+T(U(\mathcal{K}))\right)}{I(X_j;U(\mathcal{K})|X_{j-1},\ldots,X_1)} \tag{106}$$

It is not difficult to see that $\beta\in(0,1]$. This is because j is the first index for which $C_j>0$, so $C_{j-1}=0$. By definition of C_j, it is easy to see that $g(\{1,\ldots,j-1\})=0$. Observe that the numerator in the expression for β is the negative of the first term in the definition of $g(\{1,\ldots,j-1\})$, so the numerator must be positive, hence $\beta>0$. Further, by (104) and the fact that $C_j\geq0$, we have $\beta\leq1$.

Now, consider the following time-sharing of two compression schemes. Starting with the fixed constant-gap distribution $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$, we modify the distribution as follows:

1) For $(1-\beta)$ fraction of the time, keep the BSs $j,\ldots,L$ active, i.e., for $(1-\beta)$ fraction of the time, keep $X_j,\ldots,X_L$ the same and set $X_1=\ldots=X_{j-1}=0$; denote this distribution as $p(u'_1,\ldots,u'_K,x'_1,\ldots,x'_L)$.
2) For the remaining β fraction of the time, keep the BSs $j+1,\ldots,L$ active, i.e., for β fraction of the time, keep $X_{j+1},\ldots,X_L$ the same and set $X_1=\ldots,X_j=0$; denote this distribution as $p(u''_1,\ldots,u''_K,x''_1,\ldots,x''_L)$.

We first verify that the average fronthaul capacities required for this time-sharing of two compression schemes, denoted here as $\bar{C}_1,\ldots,\bar{C}_L$, are exactly the same as the fronthaul capacities $C_1,\ldots,C_L$ under the DDF strategy. For the inactive BSs from 1 to $j-1$ the fronthaul capacities used is zero, i.e.,

$$\bar{C}_l=0=C_l,\qquad\forall l=1,\ldots,j-1. \tag{107}$$

We use the modified distributions under the compression strategy to calculate the fronthaul needed for the active BSs. Note that under the constant-gap distribution (or its modified form), a corner point of the fronthaul region (46) is just

$$C_l=I(X_l;U(\mathcal{K})|X_{l-1},\ldots,X_1) \tag{108}$$

where the term $T(X(\mathcal{S}))$ vanishes because of the assumed independence of X's in the constant-gap distribution.

Now for BS j, since $X'_1=\ldots=X'_{j-1}=0$, the fronthaul used by the compression strategy is just

$$\begin{aligned}\bar{C}_j&=(1-\beta)I(X'_j;U'(\mathcal{K}))&(109)\\&=(1-\beta)I(X_j;U(\mathcal{K})|X_{j-1},\ldots,X_1)=C_j,&(110)\end{aligned}$$

where the equality is due to the form of the Gaussian $p(u_1,\ldots,u_K,x_1,\ldots,x_L)$ in which conditioning on $X_{j-1},\ldots,X_1$ is the same as setting them to be zero.

For BS $l=(j+1),\dots,L$, the fronthaul capacity used by the generalized compression strategy is given by

$$\begin{aligned}\bar{C}_j &= (1-\beta)I(X_l';U'(\mathcal{K})|X_{l-1}',\dots,X_j') \\ &\quad + \beta I(X_l'';U''(\mathcal{K})|X_{l-1}'',\dots,X_{j+1}'') \qquad (111)\\ &= (1-\beta)I(X_l;U(\mathcal{K})|X_{l-1},\dots,X_j,X_{j-1},\dots,X_1) \\ &\quad + \beta I(X_l;U(\mathcal{K})|X_{l-1},\dots,X_{j+1},X_j,\dots,X_1) \qquad (112)\\ &= I(X_l;U(\mathcal{K})|X_{l-1},\dots,X_1)=C_j, \qquad (113)\end{aligned}$$

This verifies that the time-sharing strategy uses the same amount of fronthaul as DDF.

As a final step, we show that the time-sharing of the two compression schemes achieves a sum rate no less than the DDF strategy. First, we re-write the sum rate expression under the constant-gap distribution (or its modified version) in a form that shows explicit dependence on the X variables.

Lemma 5. *Suppose that $\mathbf{X}$ is a vector of indepedent variables, and $\mathbf{Y}=\mathbf{H}\mathbf{X}+\mathbf{Z}$, $\mathbf{Y}=\mathbf{G}\mathbf{Y}+\tilde{\mathbf{Z}}$, where $\mathbf{H}$ and $\mathbf{G}$ are fixed matrices and $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are vectors of independent variables that are also independent of each other and of $\mathbf{X}$. Then,*

$$\sum_{k\in\mathcal{K}} I(U_k;Y_k)-T(U(\mathcal{K}))= I(U(\mathcal{K});X(\mathcal{L}))-\sum_{k\in\mathcal{K}} I(U_k;X(\mathcal{L})|Y_k) \qquad (114)$$

Proof.

$$\begin{aligned}&\sum_{k\in\mathcal{K}} I(U_k;Y_k)-T(U(\mathcal{K})) \qquad (115)\\ &= h(U(\mathcal{K}))-\sum_{k\in\mathcal{K}} h(U_k|Y_k) \qquad (116)\\ &= h(U(\mathcal{K}))-h(U(\mathcal{K})|X(\mathcal{L}))+h(U(\mathcal{K})|X(\mathcal{L})) \\ &\quad -\sum_{k\in\mathcal{K}} h(U_k|Y_k) \qquad (117)\\ &\overset{(a)}{=} I(U(\mathcal{K});X(\mathcal{L}))+\sum_{k\in\mathcal{K}}\left(h(U_k|X(\mathcal{L}),Y_k)-h(U_k|Y_k)\right) \qquad (118)\\ &= I(U(\mathcal{K});X(\mathcal{L}))-\sum_{k\in\mathcal{K}} I(U_k;X(\mathcal{L})|Y_k), \qquad (119)\end{aligned}$$

where in (a) we used the fact that $U(\mathcal{K})\to X(\mathcal{K})\to Y(\mathcal{K})$ forms a Markov chain and that conditioned on $X(\mathcal{L})$ the U's are independent. □

Based on the Lemma, the sum rate achieved using the time-sharing of the two generalized compression schemes with the modified constant-gap distributions can be written as

$$\begin{aligned}\bar{R} &= (1-\beta)\big(I(X_j',\dots,X_L';U'(\mathcal{K})) \\ &\quad -\sum_{k\in\mathcal{K}} I(U_k';X_j',\dots,X_L'|Y_k')\big) \\ &\quad +\beta\big(I(X_{j+1}'',\dots,X_L'';U''(\mathcal{K})) \\ &\quad -\sum_{k\in\mathcal{K}} I(U_k'';X_{j+1}'',\dots,X_L''|Y_k'')\big) \qquad (120)\\ &\overset{(a)}{=} (1-\beta)I(X_j,\dots,X_L;U(\mathcal{K})|X_{j-1},\dots,X_1) \\ &\quad +\beta I(X_{j+1},\dots,X_L;U(\mathcal{K})|X_j,\dots,X_1) \\ &\quad -(1-\beta)\sum_{k\in\mathcal{K}} I(U_k';X_j',\dots,X_L'|Y_k') \\ &\quad -\beta\sum_{k\in\mathcal{K}} I(U_k'';X_{j+1}'',\dots,X_L''|Y_k'') \qquad (121)\\ &= (1-\beta)\big(I(U(\mathcal{K});X_j|X_{j-1},\dots,X_1) \\ &\quad +I(U(\mathcal{K});X_{j+1},\dots,X_L|X_j,\dots,X_1)\big) \\ &\quad +\beta I(U(\mathcal{K});X_{j+1},\dots,X_L|X_j,\dots,X_1) \\ &\quad -(1-\beta)\sum_{k\in\mathcal{K}} I(U_k';X_j',\dots,X_L'|Y_k') \\ &\quad -\beta\sum_{k\in\mathcal{K}} I(U_k'';X_{j+1}'',\dots,X_L''|Y_k'') \qquad (122)\\ &\overset{(b)}{=} I(U(\mathcal{K});X_1,\dots,X_L)+R-\sum_{k\in\mathcal{K}} I(U_k;Y_k)+T(U(\mathcal{K})) \\ &\quad -(1-\beta)\sum_{k\in\mathcal{K}} I(U_k';X_j',\dots,X_L'|Y_k') \\ &\quad -\beta\sum_{k\in\mathcal{K}} I(U_k'';X_{j+1}'',\dots,X_L''|Y_k'') \qquad (123)\\ &\overset{(c)}{\geq} I(U(\mathcal{K});X_1,\dots,X_L)+R-\sum_{k\in\mathcal{K}} I(U_k;Y_k)+T(U(\mathcal{K})) \\ &\quad -\sum_{k\in\mathcal{K}} I(U_k;X(\mathcal{L})|Y_k) \qquad (124)\\ &\overset{(d)}{=} R. \qquad (125)\end{aligned}$$

The equality (a) holds, because as mentioned before, for the constant-gap distribution, shutting down a BS is the same as conditioning on the corresponding random variable. For the first $(1-\beta)$ fraction of time, we condition on $X_1,\dots,X_{j-1}$, and for the rest β fraction of the time, we condition on $X_1,\dots,X_j$. The equality (b) holds from the relation (105). The inequality (c) follows because under the modified constant-gap distribution,

$$\begin{aligned}&I(U_k';X_j',\dots,X_L'|Y_k') \\ &\quad = h(U_k'|Y_k')-h(U_k'|X_j',\dots,X_L',Y_k') \qquad (126)\\ &\quad = h(U_k',Y_k')-h(Y_k')-h(\tilde{Z}_k) \qquad (127)\\ &\quad = \frac{1}{2}\log\left(1+\frac{\sum_{l=j}^{L} h_{k,l}^2 P}{\sum_{l=j}^{L} h_{k,l}^2 P+\sigma^2}\right) \qquad (128)\\ &\quad < \frac{1}{2}\log\left(1+\frac{\sum_{l=1}^{L} h_{k,l}^2 P}{\sum_{l=1}^{L} h_{k,l}^2 P+\sigma^2}\right) \qquad (129)\\ &\quad = I(U_k;X(\mathcal{L})|Y_k), \qquad (130)\end{aligned}$$

and similarly $I(U_k'';X_{j+1}'',\dots,X_L''|Y_k'')<I(U_k;X(\mathcal{L})|Y_k)$. Finally, the equality (d) follows from the equivalent way of writing the sum rate as shown in Lemma 5.

Therefore, for every extreme point $(C_1,\dots,C_L)$ of $\overline{\mathcal{P}}(R)$, the time-shared compression strategy achieves a sum rate at least as large as the DDF strategy. This completes the proof.

REFERENCES

[1] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[2] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[3] B. E. Schein, "Distributed coordination in network information theory," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[4] D. Traskov and G. Kramer, "Reliable communication in networks with multi-access interference," in *IEEE Inf. Theory Workshop (ITW)*, Tahoe City, USA, Sep. 2007, pp. 343–348.

[5] W. Kang and S. Ulukus, "Capacity of a class of diamond channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4955–4960, Aug. 2011.

[6] B. Chern and A. Ozgur, "Achieving the capacity of the n-relay Gaussian diamond network within log n bits," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7708–7718, Dec. 2014.

[7] W. Kang, N. Liu, and W. Chong, "The Gaussian multiple access diamond channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6049–6059, Nov. 2015.

[8] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

[9] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6102–6111, Dec. 2011.

[10] N. Liu and W. Kang, "A new achievability scheme for downlink multicell processing with finite backhaul capacity," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, USA, Jun.-Jul. 2014, pp. 1006–1010.

[11] C. Wang, M. Wigger, and A. Zaidi, "On achievability for downlink cloud radio access networks with base station cooperation," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5726–5742, Aug. 2018.

[12] ——, "On achievability for downlink cloud radio access networks with base station cooperation," in *IEEE Wireless Commun. Netw. Conf.*, San Francisco, USA, Mar. 2017.

[13] X. Yi and N. Liu, "An achievability scheme for downlink multicell processing with finite backhaul capacity: The general case," in *Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2015.

[14] S. S. Bidokhti, G. Kramer, and S. S. Shitz, "Capacity bounds on the downlink of symmetric, multi-relay, single receiver C-RAN networks," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2058–2062.

[15] T. Yang, N. Liu, W. Kang, and S. S. Shitz, "An upper bound on the sum capacity of the downlink multicell processing with finite backhaul capacity," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2053–2057.

[16] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[17] S. H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[18] P. Patil, B. Dai, and W. Yu, "Hybrid data-sharing and compression strategy for downlink cloud radio access network," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5370–5384, Nov. 2018.

[19] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Advances Signal Process.*, Jun. 2009.

[20] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[21] S. H. Lim, K. T. Kim, and Y. H. Kim, "Distributed decode-forward for relay networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4103–4118, Jul. 2017.

[22] S. Ganguly and Y.-H. Kim, "On the capacity of cloud radio access networks," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2063–2067.

[23] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Research Develop.*, vol. 4, no. 1, pp. 66–82, Jan. 1960.

[24] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.

[25] X. Zhang, J. Chen, S. B. Wicker, and T. Berger, "Successive coding in multiuser information theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2246–2254, Jun. 2007.

[26] W. Yu, "Cloud radio access networks: coding strategies, capacity analysis, and optimization techniques," Presented at *IEEE Commun. Theory Workshop (CTW)*, Nafplio, Greece, May 2016.

[27] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, "Writing on colored paper," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Washington, DC, USA, Jun. 2001, p. 302.

[28] M. Mondelli, S. H. Hassani, I. Sason, and R. L. Urbanke, "Achieving Marton's region for broadcast channels using polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 783–800, Feb. 2015.

[29] Y. Zhou, Y. Xu, W. Yu, and J. Chen, "On the optimal fronthaul compression and decoding strategies for uplink cloud radio access networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7402–7418, Dec. 2016.

[30] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, Jul. 2019.

[31] L. Liu, P. Patil, and W. Yu, "An uplink-downlink duality for cloud radio access network," in *IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1606–1610.

[32] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*. Springer-Verlag, 2003.

[33] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.