On Cloud Radio Access Networks With Cascade Oblivious Relaying

Mehrangiz Ensan*, Hamdi Joudeh*, Alex Alvarado*, Ulf Gustavsson†, Frans M. J. Willems*
*Information and Communication Theory Lab, Eindhoven University of Technology, The Netherlands
†Ericsson Research, Ericsson AB, Sweden
Email:*{m.ensan, h.joudeh, a.alvarado, f.m.j.willems}@tue.nl, †ulf.gustavsson@ericsson.com

***Abstract*—We consider a discrete memoryless cloud radio access network in which K users communicate with a remote destination through 2 relays in a cascade. The relays are oblivious in the sense that they operate without knowledge of the users' codebooks. We focus on a scenario where the first and second relays are connected through a finite-capacity error-free link, while the second relay is connected to the remote destination via an infinite-capacity link. We establish the capacity region in this case, and show that it is achieved via a compress-and-forward scheme with successive decoding. Finally, the extension to Gaussian networks is discussed.**

I. Introduction

In a cloud radio access networks (CRAN) operating in uplink, user equipments (UEs) send messages to a central processor (CP) via access points (APs) distributed over a service area. APs act as relays, and they are often connected to the CP through links of limited capacity. This architecture enables the joint processing of AP signals, which in turn helps alleviate the effects of inter-cell interference in traditional cellular architectures. Other benefits include the utilization of cost-effective APs with limited functionality, as most of the heavy-duty processing is carried out at the CP [1].

In this work, we are interested in the capacity of an elemental CRAN model, in which APs are connected to the CP through a line network, i.e., cascaded APs (see Fig. 1). We are especially interested in the case where APs are constrained to operate without knowledge of users' codebooks and are hence unable to decode user messages, which is known in the literature as *oblivious* relaying (or processing) [2]–[4]. This oblivious cascade model can perhaps be motivated by emerging cell-free networks, especially those utilizing the so-called radio stripe architecture, where small and low-cost APs are connected to a CP through a serial fronthaul [5], [6].

Related work: While the CRAN literature is too vast to survey here, most relevant to this paper are the works in [2]–[4]. Oblivious relaying or processing was introduced in [2], where randomized encoding was used to model the relays' lack of codebook knowledge. This was done in the context of a single-user multi-relay network, where each relay is connected to the CP through a dedicated error-free finite-capacity link. In [3], the concept of oblivious processing was extended to include time-sharing, where users are allowed to switch among different codebooks, and relays have knowledge of user schedules but not the codebooks themselves. [3] focused on the relay channel and the relay-aided interference channel settings. [4] extends the original setting in [2] by enabling time-sharing and incorporating multiple users (see also [7]). For settings with oblivious processing where the capacity has been established, e.g. [3] and [4], it has been shown that schemes based on compress-and-forward relaying are optimal.

Another relevant work is [8], where a Gaussian multiple-input multiple-output (MIMO) CRAN setting with cascaded relays was considered. This setting is very closely related to the one we consider here. However, oblivious processing is not explicitly treated in [8]. More importantly, the focus in [8] is on achievability schemes and evaluating the distortion resulting from relay processing and compression, and there are no information-theoretic capacity results.

Contribution: In this paper, we study the capacity of CRAN networks with oblivious cascaded relays. We consider an elemental discrete memoryless (DM) setting with K UEs and 2 APs (see Fig. 1). As an initial step, we focus on the case where AP-1 and AP-2 are connected through a finite-capacity error-free link, while AP-2 and the CP are connected via an infinite-capacity link. We establish the capacity region in this case, and show that it is achieved via a compress-and-forward scheme with separate and successive decompression and decoding. We then discuss the MIMO Gaussian setting, and we conclude the paper by highlighting a few interesting open problems.[1]

II. System model

We consider a DM-CRAN in which K UEs communicate with a CP through a cascade of 2 APs that act as relays—see Fig. 1. We define the sets of UEs and APs as $\mathcal{K} \coloneqq [K]$ and $\mathcal{L} \coloneqq [2]$, respectively. AP-1 is connected to AP-2 through an error-free fronthaul link of capacity C_1, and AP-2 is connected to the CP through an error-free fronthaul link of capacity C.

Discrete memoryless channel: Inputs and outputs to the discrete memoryless channel are denoted by X_k and Y_l, respectively, where $k \in \mathcal{K}$ and $l \in \mathcal{L}$, and are drawn from the alphabets $\mathcal{X}_k$ and $\mathcal{Y}_l$, respectively. We define $X_{\mathcal{K}} \coloneqq \{X_1, X_2, \cdots, X_K\}$ as the set of all UE inputs. The channel transition probabilities are given by $p_{Y_l|X_{\mathcal{K}}}(y_l|x_{\mathcal{K}})$, defined for all input-output pairs $(x_{\mathcal{K}}, y_l)$.

[1] We use standard notation. For a positive integer K, the set $\{1, 2, \ldots, K\}$ is denoted by $[K]$. $(\cdot)^{\mathsf{T}}$ and $(\cdot)^{\mathsf{H}}$ are, respectively, the transpose and conjugate-transpose operators. $\mathrm{diag}(\cdot)$ is a block diagonal matrix of its arguments. We sometimes write a probability mass function (pmf) $p_X(x)$ as $p(x)$.

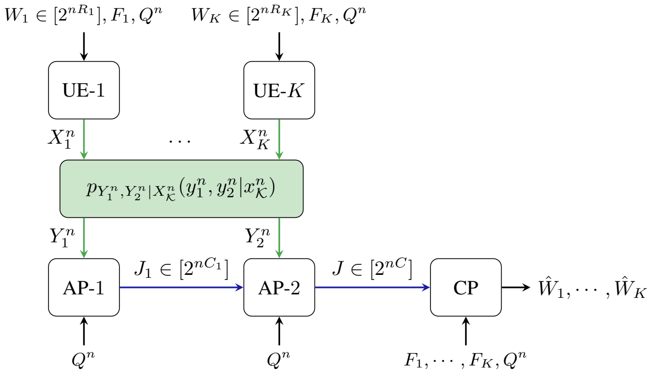

$W_1 \in [2^{nR_1}], F_1, Q^n$
$W_K \in [2^{nR_K}], F_K, Q^n$
UE-1
UE-K
X_1^n
$\cdots$
X_K^n
$p_{Y_1^n,Y_2^n|X_\mathcal{K}^n}(y_1^n,y_2^n|x_\mathcal{K}^n)$
Y_1^n
Y_2^n
AP-1
$J_1 \in [2^{nC_1}]$
AP-2
$J \in [2^{nC}]$
CP
$\hat{W}_1,\cdots,\hat{W}_K$
Q^n
Q^n
$F_1,\cdots,F_K,Q^n$

Fig. 1: DM-CRAN model with 2 oblivious relays (AP-1 and AP-2) in a cascade. The links from AP-1 to AP-2 and AP-2 to CP are assumed to be finite-capacity error-free links.

An input sequence when the channel is used n times is denoted by $X_k^n := (X_{k,1}, X_{k,2}, \ldots, X_{k,n})$. Similarly, an output sequences is given by $Y_l^n := (Y_{l,1}, Y_{l,2}, \ldots, Y_{l,n})$. We define $X_\mathcal{K}^n := \{X_1^n, X_2^n, \cdots, X_K^n\}$. The observations of the APs are conditionally independent on the transmitted sequences. Hence the multi-letter channel transition probabilities factorize as

$$\begin{aligned} p_{Y_1^n,Y_2^n|X_\mathcal{K}^n}(y_1^n,y_2^n|x_\mathcal{K}^n) &= p_{Y_1^n|X_\mathcal{K}^n}(y_1^n|x_\mathcal{K}^n)p_{Y_2^n|X_\mathcal{K}^n}(y_2^n|x_\mathcal{K}^n) \\ &= \prod_{i=1}^{n}\prod_{l=1}^{2} p_{Y_l|X_\mathcal{K}}(y_{l,i}|x_{\mathcal{K},i}). \end{aligned} \tag{1}$$

Oblivious processing with time-sharing: We assume that the APs are oblivious, i.e., they are constrained to operate without knowledge of the UEs' codebooks [2]. Nevertheless, all nodes in the network have access to a common time-sharing sequence Q^n, i.e., a sequence of time instances at which encoders may switch among different codebooks [3], [4]. The time-sharing sequence is i.i.d. with a distribution $p_{Q^n}(q^n) = \prod_{i=1}^n p_Q(q_i)$, where $p_Q(q)$ is some pmf defined on an alphabet $\mathcal{Q}$. As pointed out in [3], the presence of a common time-sharing sequence enables a degree of coordination, e.g. APs may know which UEs are active and which are not. However, APs cannot implement higher functionalities as decoding UE messages.

Encoding: Each UE-k has a message W_k of rate R_k, drawn uniformly at random from $[2^{nR_k}]$. Obliviousness of the APs is modeled through randomized encoding [2]. Specifically, each UE-k selects a codebook at random from the set of all possible codebooks of the given rate R_k. The index of the selected codebook is a random variable F_k defined on $[|\mathcal{X}_k|^{n2^{nR_k}}]$, and it is independent of W_k. Given the time sharing sequence Q^n, F_k has a conditional distribution of $p_{F_k|Q^n}(f_k|q^n)$. The codebook index F_k is revealed to the CP (to perform decoding), but not the APs (oblivious).

The encoding function for UE-k is defined by the pair $\left(\phi_k^n, \{p_{X_k|Q}\}\right)$. The first part is a mapping

$$\phi_k^n : [|\mathcal{X}_k|^{n2^{nR_k}}] \times [2^{nR_k}] \times \mathcal{Q}^n \to \mathcal{X}_k^n \tag{2}$$

which assigns an input sequence (i.e., a codeword) $X_k^n = \phi_k^n(F_k, W_k, Q^n)$ to each choice of codebook index F_k, UE message W_k, and time-sharing sequence Q^n. The second part $\{p_{X_k|Q}\}$ is a family of single-letter conditional pmfs $\{p_{X_k|Q}(x_k|q) : q \in \mathcal{Q}\}$, each defined on the alphabet $\mathcal{X}_k$. The probability of selecting a codebook with index $F_k = f_k$ is determined by $\left(\phi_k^n, \{p_{X_k|Q}\}\right)$ as

$$p_{F_k|Q^n}(f_k|q^n) = \prod_{w_k=1}^{2^{nR_k}} p_{X_k^n|Q^n}\left(\phi_k^n(f_k, q_k, q^n)|q^n\right) \tag{3}$$

where the distribution $p_{X_k^n|Q^n}(x_k^n|q^n) = \prod_{i=1}^n p_{X_k|Q}(x_{k,i}|q_i)$ is obtained from $\{p_{X_k|Q}\}$.

Remark 1. Without knowledge of the selected codebook index F_k, and conditioned on the time-sharing sequence Q^n, the probability of drawing a codeword $x_k^n \in \mathcal{X}_k^n$ is given by

$$\Pr\{X_k^n(W_k, F_k, Q^n) = x_k^n | Q^n = q^n\} = \prod_{i=1}^{n} p_{X_k|Q}(x_{k,i}|q_i) \tag{4}$$

where probability is defined with respect to the joint pmf

$$p_{W_k,F_k,Q^n}(w_k, f_k, q^n) = 2^{-nR_k} p_{F_k|Q^n}(f_k|q^n) p_{Q^n}(q^n). \tag{5}$$

This follows from [2, Lemma 1] (see also [4, Lemma 1]). As pointed out in [3, Remark 5], the above implies that an oblivious node that is not informed about F_k, but has access to Q^n, sees the codeword X_k^n as an unstructured sequence with independent entries (although not necessarily i.i.d.). This prohibits the APs from decoding UE messages. On the other hand, a node which knows F_k sees a structured codeword X_k^n, where the structure is provided by the choice of codebook.

Relaying: With knowledge of Q^n, AP-1 maps its channel observation Y_1^n into an index J_1 which takes values on $[2^{nC_1}]$. In particular, we have $J_1 = \varphi_1^n(Y_1^n, Q^n)$, where $\varphi_1^n : \mathcal{Y}_1^n \times \mathcal{Q}^n \to [2^{nC_1}]$ is the corresponding relaying function. J_1 is sent to AP-2 over the fronthaul link of capacity C_1.

AP-2 maps the index J_1 and its own channel observation Y_2^n into a new index $J \in [2^{nC}]$, i.e., $J = \varphi^n(J_1, Y_2^n, Q^n)$, where $\varphi^n : [2^{nC_1}] \times \mathcal{Y}_2^n \times \mathcal{Q}^n \to [2^{nC}]$ is the relaying function of AP-2. The new index J is then sent to the CP over the fronthaul link of capacity C.

Decoding: Recall that the CP has access to Q^n and $F_\mathcal{K} := \{F_1, F_2, \ldots, F_K\}$. Upon receiving the index J, the CP estimates the UE messages as

$$(\hat{W}_1, \cdots, \hat{W}_K) = \psi^n(F_\mathcal{K}, J, Q^n) \tag{6}$$

where the decoding function ψ^n is a mapping such that

$$\begin{aligned} \psi^n : [|\mathcal{X}_1|^{n2^{nR_1}}] \times \cdots \times [|\mathcal{X}_K|^{n2^{nR_K}}] \times [2^{nC}] \times \mathcal{Q}^n \\ \to [2^{nR_1}] \times \cdots \times [2^{nR_K}]. \end{aligned} \tag{7}$$

Achievable rates and capacity: A $(n, R_1, R_2, \ldots, R_K)$ code for the above setting includes K UE encoding functions, 2 AP relaying functions, and a CP decoding function, all as specified above. For give fronthaul link capacities (C_1, C), a rate tuple $(R_1, R_2, \ldots, R_K)$ is said to be achievable if there exists a sequence of $(n, R_1, R_2, \ldots, R_K)$ codes with a vanishing probability of error, i.e.

$$\lim_{n\to\infty} \Pr\{(\hat{W}_1, \cdots, \hat{W}_K) \neq (W_1, \cdots, W_K)\} = 0. \quad (8)$$

Note that the probability of error is defined with respect to the joint distribution of $(W_{\mathcal{K}}, F_{\mathcal{K}})$, and is hence averaged over the distribution of codebooks. For given fronthaul capacities (C_1, C), the capacity region $\mathcal{C}(C_1, C)$ is the closure of the set of all achievable rate tuples.

III. Main Result and Insights

Here we present the main result of this paper, where we characterize the exact capacity region $\mathcal{C}(C_1, C)$ of the above described setting when the fronthaul capacity C of the link from AP-2 to the CP is unlimited. We denote the capacity region by $\mathcal{C}(C_1, \infty)$ in this case. Also, for any subset of UEs $\mathcal{S} \subseteq \mathcal{K}$, we use $R(\mathcal{S})$ to denote the sum rate $\sum_{k\in\mathcal{S}} R_k$.

Theorem 1. *For the DM-CRAN setting with K UEs and 2 cascaded APs described in Section II, $\mathcal{C}(C_1, \infty)$ is given by all rate tuples $(R_1, R_2, \ldots, R_K)$ that satisfy*

$$R(\mathcal{S}) \leq I(X_{\mathcal{S}}; U_1, Y_2 | X_{\mathcal{S}^c}, Q), \forall \mathcal{S} \subseteq \mathcal{K} \quad (9)$$

$$C_1 \geq I(Y_1; U_1 | Y_2, Q) \quad (10)$$

for some $(Q, X_{\mathcal{K}}, Y_1, Y_2, U_1)$ with a joint pmf of

$$p(q) \prod_{k=1}^{K} p(x_k|q) p(y_1|x_{\mathcal{K}}) p(y_2|x_{\mathcal{K}}) p(u_1|y_1, q). \quad (11)$$

The achievability of Theorem 1, presented in Section IV-A, is based on classical coding for the discrete memoryless multiple access channel (DM-MAC), combined with compress-and-forward relaying. AP-1 applies Wyner-Ziv binning to its received observation Y_1^n and forwards the bin index J_1 to AP-2, which in turn forwards J_1 alongside an almost lossless representation of its own observation Y_2^n to the CP. This compress-and-forward approach is commonly employed in the CRAN literature, see, e.g. [2]–[4], [7]. We specialize this approach to the cascade setting of interest.

The converse proof, presented in Section IV-B, is less straightforward. In general, the proof follows along the same footsteps of previous proofs in the literature on oblivious relaying, specifically the proofs of [3, Proposition 1] and [4, Theorem 1]. Nevertheless, due to the distinct features of the multi-user cascade-relay setting under consideration, the converse proof does not follow directly from [3], [4]. A main challenge lies in identifying the right auxiliary random variables (Q, U_1) that work for this cascade setting, and satisfy the conditional independence relationships in (11).

Remark 2. The fact that a compress-and-forward based scheme is optimal for the setting considered in Theorem 1 is perhaps not surprising, given the obliviousness of the APs. A somewhat pleasant surprise, however, is that separate and successive decompression and decoding turn out to be optimal in this cascade setting. As shown in Section IV-A, to achieve all points in the capacity region it is sufficient for the CP to first retrieve Y_2^n, then retrieve the compressed version of Y_1^n from J_1 and Y_2^n, and then decode user messages in a successive fashion as done in the standard DM-MAC. This is in sharp contrast to the more common CRAN setting in [4], in which APs are connected to the CP through a fronthaul network with a star topology, and where joint decompression and decoding can strictly outperform separate and successive decompression and decoding in general [7]. The latter is, of course, much more desirable in practice due to lower complexity.

Remark 3. C need not be infinite for the capacity result in Theorem 1 to stand. In fact, the result holds as long as we have $C \geq C_1 + H(Y_2|Q)$, for all $p(q, x_{\mathcal{K}}) = p(q) \prod_{k=1}^{K} p(x_k|q)$ defined on $\mathcal{Q} \times \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$. In this case, AP-2 can forward both J_1 and Y_2^n almost losslessly to the CP for any choices of input distributions, and the capacity region remains unchanged.

IV. Proof of Theorem 1

A. Achievability

Encoding: We fix a pmf $p_Q(q)$, defined on $\mathcal{Q}$, and generate an i.i.d. time-sharing sequence q^n. This sequence is shared across all network nodes. For each UE-k, we fix a family of single-letter pmfs $\{p_{X_k|Q}(x_k|q) : q \in \mathcal{Q}\}$, defined on $\mathcal{X}_k$. A codebook of 2^{nR_k} sequences is generated, where each sequence $x_k^n(w_k, q^n)$ is independently drawn from $\prod_{i=1}^{n} p_{X_k|Q}(x_{k,i}|q_i)$. Note that this standard random codebook generation is equivalent to drawing a codebook index F_k from $p_{F_k|Q^n}(f_k|q^n)$ defined in (3). All UE codebooks are shared with the CP, but not the APs. To send the messages $(w_1, w_2, \ldots, w_k)$, each UE-k sends the corresponding codeword $x_k^n(w_k, q^n)$ over the channel.

Relaying: Each AP-l receives a sequence y_l^n, distributed as $p_{Y_l^n|Q^n}(y_l^n|q^n) = \prod_{i=1}^{n} p_{Y_l|Q}(y_{l,i}|q_i)$. AP-1 applies Wyner-Ziv binning to its observation y_1^n, see, e.g., [9, Ch. 15.9] and [10, Ch. 11.3]. We have a compression codebook of $2^{n\hat{C}_1}$ sequences, indexed by $[2^{n\hat{C}_1}]$, where each sequence is randomly and independently drawn from $\prod_{i=1}^{n} p_{U_1|Q}(u_{1,i}|q_i)$. The pmf $p_{U_1|Q}(u_{1,i}|q_i)$ is obtained from some fixed joint pmf $p_{Y_1,U_1|Q}(y_{1,i}, u_{1,i}|q_i)$, and we set $\hat{C}_1 \geq I(Y_1; U_1|Q)$. These sequences are randomly and independently assigned to 2^{nC_1} bins, indexed by the set $[2^{nC_1}]$. AP-1 finds a description $u_1^n(i_1)$, where $i_1 \in [2^{n\hat{C}_1}]$, which is strongly ϵ-jointly typical with y_1^n. AP-1 then finds the index $j_1 \in [2^{nC_1}]$ of the bin in which $u_1^n(i_1)$ falls. The index j_1 is forwarded to AP-2.

AP-2 applies standard (almost) lossless compression to its observation y_2^n. Typical realizations of Y_2^n are indexed by the set $[2^{nC_2}]$, and AP-2 finds the index $j_2 \in [2^{C_2}]$ which corresponds to the observed y_2^n (non-typical observations are ignored). AP-2 forwards the tuple $j = (j_1, j_2)$ to the CP.

Decoding: The CP first reconstructs (u_1^n, y_2^n) in a successive fashion by first retrieving y_2^n from j_2, and then u_1^n from j_1 and the retrieved y_2^n. Using standard arguments, it follows that (u_1^n, y_2^n) can be reconstructed with a vanishing probability of error as $n \to \infty$ given that

$$C_2 \geq H(Y_2|Q) \quad (12)$$

$$C_1 \geq I(Y_1; U_1|Y_2, Q). \tag{13}$$

The CP then decodes the UE messages $(w_1, w_2, \ldots, w_K)$ from (u_1^n, y_2^n). Note that this can be viewed as a DM-MAC with inputs $X_1, X_2, \ldots, X_K$ and outputs U_1, Y_2. Therefore, UE messages are decoded with a vanishing probability of error as $n \to \infty$ provided that the rates satisfy

$$R(\mathcal{S}) \leq I(U_1, Y_2; X_{\mathcal{S}}|X_{\mathcal{S}^c}, Q), \forall \mathcal{S} \subseteq \mathcal{K}. \tag{14}$$

It is worthwhile highlighting that both successive decoding and joint decoding of UE messages from (u_1^n, y_2^n) lead to achieving the rates specified in (14), see, e.g., [10, Ch. 4.5].

B. Converse

Suppose that the rate tuple $(R_1, R_2, \ldots, R_K)$ is achievable. Then for any subset of UEs $\mathcal{S} \subseteq \mathcal{K}$, and starting from a standard application of Fano's inequality, we write

$$\begin{aligned}
nR(\mathcal{S}) &= H(W_{\mathcal{S}}) && (15)\\
&\leq I(W_{\mathcal{S}}; J, F_{\mathcal{K}}, Q^n) + n\epsilon_n && (16)\\
&= I(W_{\mathcal{S}}; J|F_{\mathcal{K}}, Q^n) + I(W_{\mathcal{S}}; F_{\mathcal{K}}, Q^n) + n\epsilon_n && (17)\\
&\leq I(W_{\mathcal{S}}, F_{\mathcal{S}}; J|F_{\mathcal{S}^c}, Q^n) + n\epsilon_n && (18)\\
&\leq I(W_{\mathcal{S}}, F_{\mathcal{S}}; J|W_{\mathcal{S}^c}, F_{\mathcal{S}^c}, Q^n) + n\epsilon_n && (19)\\
&\leq I(X_{\mathcal{S}}^n; J|W_{\mathcal{S}^c}, F_{\mathcal{S}^c}, Q^n) + n\epsilon_n && (20)\\
&= I(X_{\mathcal{S}}^n; J|X_{\mathcal{S}^c}^n, W_{\mathcal{S}^c}, F_{\mathcal{S}^c}, Q^n) + n\epsilon_n && (21)\\
&\leq I(X_{\mathcal{S}}^n, W_{\mathcal{S}^c}, F_{\mathcal{S}^c}; J|X_{\mathcal{S}^c}^n, Q^n) + n\epsilon_n && (22)\\
&= I(X_{\mathcal{S}}^n; J|X_{\mathcal{S}^c}^n, Q^n)\\
&\quad + I(W_{\mathcal{S}^c}, F_{\mathcal{S}^c}; J|X_{\mathcal{K}}^n, Q^n) + n\epsilon_n && (23)\\
&= I(X_{\mathcal{S}}^n; J|X_{\mathcal{S}^c}^n, Q^n) + n\epsilon_n && (24)\\
&\leq I(X_{\mathcal{S}}^n; J_1, Y_2^n|X_{\mathcal{S}^c}^n, Q^n) + n\epsilon_n. && (25)
\end{aligned}$$

In (17), $I(W_{\mathcal{S}}; F_{\mathcal{K}}, Q^n) = 0$ since $W_{\mathcal{S}}$ is independent from $(F_{\mathcal{K}}, Q^n)$. The inequality in (19) holds since $(W_{\mathcal{S}}, F_{\mathcal{S}})$ and $W_{\mathcal{S}^c}$ are independent. (20) follows from the data processing inequality. (21) is due to the fact that $X_{\mathcal{S}^c}^n$ is a function of $(W_{\mathcal{S}^c}, F_{\mathcal{S}^c}, Q^n)$. In going from (23) to (24), we used the Markov chain $(Q^n, W_{\mathcal{K}}, F_{\mathcal{K}}) \to X_{\mathcal{K}}^n \to J$, and (25) holds since J is a function of (J_1, Y_2^n).

Continuing from (25), we obtain

$$\begin{aligned}
&n(R(\mathcal{S}) - \epsilon_n)\\
&\leq \sum_{i=1}^{n} \Big[H(X_{\mathcal{S},i}|X_{\mathcal{S}}^{i-1}, X_{\mathcal{S}^c}^n, Q^n)\\
&\quad - H(X_{\mathcal{S},i}|J_1, Y_2^n, X_{\mathcal{S}}^{i-1}, X_{\mathcal{S}^c}^n, Q^n)\Big] && (26)\\
&\leq \sum_{i=1}^{n} \Big[H(X_{\mathcal{S},i}|X_{\mathcal{S}^c}^n, X_{\mathcal{S}}^{i-1}, X_{\mathcal{S},i+1}^n, Q^n)\\
&\quad - H(X_{\mathcal{S},i}|J_1, Y_2^n, X_{\mathcal{S}^c}^n, X_{\mathcal{S}}^{i-1}, X_{\mathcal{S},i+1}^n, Q^n)\Big] && (27)\\
&= \sum_{i=1}^{n} I(X_{\mathcal{S},i}; J_1, Y_2^n|X_{\mathcal{S}^c,i}, \tilde{Q}_i) && (28)\\
&\leq \sum_{i=1}^{n} I(X_{\mathcal{S},i}; J_1, Y_2^n, Y_1^{i-1}|X_{\mathcal{S}^c,i}, \tilde{Q}_i) && (29)\\
&= \sum_{i=1}^{n} I(X_{\mathcal{S},i}; U_{1,i}, Y_{2,i}|X_{\mathcal{S}^c,i}, \tilde{Q}_i). && (30)
\end{aligned}$$

$H(X_{\mathcal{S},i}|X_{\mathcal{S}}^{i-1}, X_{\mathcal{S}^c}^n, Q^n) = H(X_{\mathcal{S},i}|X_{\mathcal{S}^c}^n, X_{\mathcal{S}}^{i-1}, X_{\mathcal{S},i+1}^n, Q^n)$ in (27) due to (4), and we have used the fact that conditioning does not increase entropy for negative-signed terms in (27). In (28), we define $\tilde{Q}_i := (X_{\mathcal{K}}^{i-1}, X_{\mathcal{K},i+1}^n, Q^n)$. In (30), we define $U_{1,i} := (J_1, Y_1^{i-1}, Y_2^{i-1}, Y_{2,i+1}^n)$.

Next, we wish to verify that our choice of auxiliary random variables $\tilde{Q}_i$ and $U_{1,i}$ satisfy the relationship in (11). We observe that when conditioned on $\tilde{Q}_i$, the Markov chain

$$U_{1,i} \to Y_{1,i} \to X_{\mathcal{K},i} \to Y_{2,i} \tag{31}$$

holds. This can be shown using the d-separation graphical method, see, e.g. [11, Appendix A.9] and [3, Appendix C]. Therefore, $(\tilde{Q}_i, X_{\mathcal{K},i}, Y_{1,i}, Y_{2,i}, U_{1,i})$ have a joint distribution that takes the form of (11).

Next, we invoke a standard single-letterization argument. We define a time-sharing random variable Q', uniformly distributed on $[n]$. We also define $X_{\mathcal{K}} := X_{\mathcal{K},Q'}$, $Y_1 := Y_{1,Q'}$, $Y_2 := Y_{2,Q'}$, $U_1 := U_{1,Q'}$ and $Q := (\tilde{Q}_i, Q')$. Rewriting (30) in terms of the newly defined notation, we obtain

$$\begin{aligned}
n(R(\mathcal{S}) - \epsilon_n) &\leq nI(X_{\mathcal{S},Q'}; U_{1,Q'}, Y_{2,Q'}|X_{\mathcal{S}^c,Q'}, \tilde{Q}_{Q'}, Q')\\
&= nI(X_{\mathcal{S}}; U_1, Y_2|X_{\mathcal{S}^c}, Q). && (32)
\end{aligned}$$

Taking $n \to \infty$ leads to $\epsilon_n \to 0$, and we obtain the desired bound in (9). Next, we show that the bound in (10) is also a necessary condition. Since J_1 is defined on $[2^{nC_1}]$, we have

$$\begin{aligned}
nC_1 &\geq H(J_1) && (33)\\
&\geq I(X_{\mathcal{K}}^n, Y_1^n; J_1|Y_2^n, Q^n) && (34)\\
&= \sum_{i=1}^{n} I(X_{\mathcal{K}}^n, Y_{1,i}; J_1|Y_1^{i-1}, Y_2^n, Q^n) && (35)\\
&= \sum_{i=1}^{n} \Big[I(X_{\mathcal{K},i}, Y_{1,i}; J_1|Y_1^{i-1}, Y_2^n, X_{\mathcal{K}}^{i-1}, X_{\mathcal{K},i+1}^n, Q^n)\\
&\quad + I(X_{\mathcal{K}}^{i-1}, X_{\mathcal{K},i+1}^n, Y_{1,i}; J_1|Y_1^{i-1}, Y_2^n, Q^n)\Big] && (36)\\
&\geq \sum_{i=1}^{n} I(Y_{1,i}; J_1|Y_1^{i-1}, Y_2^n, \tilde{Q}_i) && (37)\\
&= \sum_{i=1}^{n} \Big[I(Y_{1,i}; J_1, Y_1^{i-1}, Y_2^{i-1}, Y_{2,i+1}^n|Y_{2,i}, \tilde{Q}_i)\\
&\quad - I(Y_{1,i}; Y_1^{i-1}, Y_2^{i-1}, Y_{2,i+1}^n|Y_{2,i}, \tilde{Q}_i)\Big] && (38)\\
&= \sum_{i=1}^{n} I(Y_{1,i}; U_{1,i}|Y_{2,i}, \tilde{Q}_i). && (39)
\end{aligned}$$

In the above, (35) and (36) follow from the chain rule, while in (37) we used the non-negativity of mutual information and the definition of $\tilde{Q}_i$. The negative-signed mutual information terms in (38) are equal to zero due to the Markov chain $(Y_{1,i}, Y_{2,i}) \to X_{\mathcal{K},i} \to (X_{\mathcal{K}}^{i-1}, X_{\mathcal{K},i+1}^n) \to (Y_1^{i-1}, Y_2^{i-1}, Y_{2,i+1}^n)$, which holds due to (1). Now continuing from (39) using the time-sharing argument, we obtain

$$nC_1 \geq nI(Y_{1,Q'}; U_{1,Q'}|Y_{2,Q'}, \tilde{Q}_{Q'}, Q') \tag{40}$$

$$= nI(Y_1; U_1|Y_2, Q). \tag{41}$$

This concludes the proof.

V. Gaussian Network

In this section, we build on the result for the DM setting in Theorem 1 and investigate a memoryless Gaussian MIMO CRAN with K UEs, equipped with M antennas each, and 2 cascaded APs, equipped with N antennas each. Ignoring the channel use index, the input-output relationship is given by

$$\boldsymbol{Y}_l = \sum_{k=1}^{K} \mathbf{H}_{lk} \boldsymbol{X}_k + \boldsymbol{Z}_l, \ l \in \mathcal{L} \tag{42}$$

where $\boldsymbol{Y}_l$ is the $N \times 1$ signal received by AP-l, $\mathbf{H}_{lk}$ is the $N \times M$ complex channel matrix between UE-k and AP-l, $\boldsymbol{X}_k$ is the $M \times 1$ complex signal transmitted by UE-k, and $\boldsymbol{Z}_l \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_l)$ is the $N \times 1$ additive noise vector at AP-l, with zero mean and covariance matrix $\mathbf{\Sigma}_l$. Channel matrices are fixed, and each $\boldsymbol{X}_k$ is subject to a covariance power constraint $\mathbb{E}[\boldsymbol{X}_k \boldsymbol{X}_k^{\mathsf{H}}] \preceq \mathbf{K}_k$, for some covariance matrix $\mathbf{K}_k$. As in Theorem 1, we assume that C_1 is finite and C is infinite.

Next, we restrict to Gaussian signalling at the UEs, Gaussian quantization at AP-1, and no time-sharing. Here we have $\boldsymbol{X}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{K}_k)$ for all $k \in \mathcal{K}$, and $p(\boldsymbol{u}_1|\boldsymbol{y}_1) \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda}_1)$, where $\boldsymbol{u}_1$ is a quantized version of $\boldsymbol{y}_1$ and $\mathbf{\Lambda}_1$ is the Gaussian quantization covariance matrix at AP-1.

Lemma 1. Under the above Gaussian signalling and quantization restrictions, an achievable rate region is given by all rate tuple $(R_1, R_2, \ldots, R_K)$ that satisfy

$$R(\mathcal{S}) \leq \log \frac{|\mathbf{H}_{\mathcal{LS}} \mathbf{K}_{\mathcal{S}} \mathbf{H}_{\mathcal{LS}}^{\mathsf{H}} + \mathrm{diag}(\mathbf{\Sigma}_1 + \mathbf{\Lambda}_1, \mathbf{\Sigma}_2)|}{|\mathrm{diag}(\mathbf{\Sigma}_1 + \mathbf{\Lambda}_1, \mathbf{\Sigma}_2)|}, \tag{43}$$

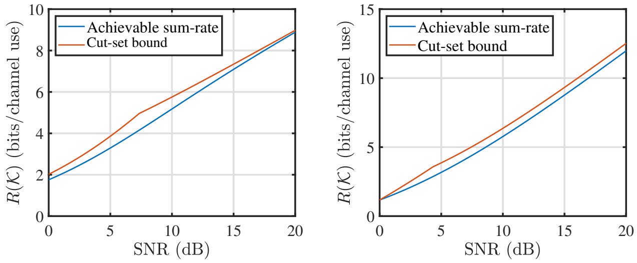
$$C_1 \geq \log \frac{|\mathbf{H}_{\mathcal{LK}} \mathbf{K}_{\mathcal{K}} \mathbf{H}_{\mathcal{LK}}^{\mathsf{H}} + \mathrm{diag}(\mathbf{\Sigma}_1 + \mathbf{\Lambda}_1, \mathbf{\Sigma}_2)|}{|\mathbf{H}_{2\mathcal{K}} \mathbf{K}_{\mathcal{K}} \mathbf{H}_{2\mathcal{K}}^{\mathsf{H}} + \mathbf{\Sigma}_2||\mathbf{\Lambda}_1|} \tag{44}$$

where $\mathbf{H}_{l\mathcal{S}} := [\mathbf{H}_{ls_1} \ \cdots \ \mathbf{H}_{ls_{|\mathcal{S}|}}]$, $\mathbf{H}_{\mathcal{LS}} := [\mathbf{H}_{1\mathcal{S}}^{\mathsf{T}} \ \mathbf{H}_{2\mathcal{S}}^{\mathsf{T}}]^{\mathsf{T}}$ and $\mathbf{K}_{\mathcal{S}} := \mathrm{diag}(\mathbf{K}_{s_1}, \ldots, \mathbf{K}_{s_{|\mathcal{S}|}})$, for any subset of users given by $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\} \subseteq \mathcal{K}$.

Lemma 1 is obtained from the achievability part of Theorem 1 in combination with calculations similar to the ones in [7, Sec. IV]. The proof is omitted for the sake of brevity.

It has been observed in the oblivious relaying literature that restricting to Gaussian signalling can be sub-optimal, as in some cases non-Gaussian (e.g. discrete) signalling enables the oblivious relay to perform some useful processing (e.g. demodulation), see [2, Sec. VI.B]. Nevertheless, it remains interesting (and perhaps useful) to evaluate how close we can get to the capacity using only Gaussian codebooks. To this end, we present a couple of numerical examples in which we compare the achievable sum-rate obtained from Lemma 1 to the cut-set bound. Note that the cut-set bound here is obtained by adapting the bound for the relay channel with orthogonal receiver components [10, Sec. 16.7.3], from which we have

$$R(\mathcal{K}) \leq \min \left\{ C_1 + \log \frac{|\mathbf{H}_{2,\mathcal{K}} \mathbf{K}_{\mathcal{K}} \mathbf{H}_{2,\mathcal{K}}^{\mathsf{H}} + \mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_2|}, \right. \\ \left. \log \frac{|\mathbf{H}_{\mathcal{L},\mathcal{K}} \mathbf{K}_{\mathcal{K}} \mathbf{H}_{\mathcal{L},\mathcal{K}}^{\mathsf{H}} + \mathrm{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)|}{|\mathrm{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)|} \right\} \tag{45}$$



(a) (b)

Fig. 2: Achievable sum-rate in Lemma 1 vs. cut-set bound in (45) for (a) Fixed fronthaul: $C_1 = 2$, (b) Scaling fronthaul: $C_1 = \log_2(\text{SNR})$.

The comparisons between the achievable sum-rate and the cut-set bound are shown in Fig. 2 for a simple 2-UE, 2-AP network, where nodes have a single-antenna each. Each UE has a power constraint equal to SNR, while a unity noise variance is assumed at the APs. Channel vectors of UE-1 and UE-2 are given by $[1 \ 0.5]^{\mathsf{T}}$ and $[0.5 \ 1]^{\mathsf{T}}$, respectively. In Fig. 2a we consider a fixed fronthaul capacity $C_1 = 2$, and in Fig. 2b we consider a scaling fronthaul capacity $C_1 = \log(\text{SNR})$, both in bits per channel use. In both cases, the gap between the achievable sum-rate and the cut-set bound never exceeds 1 bit per channel use for any SNR, and is often much less. This suggests that Gaussian signaling may be sufficient to achieve rates within a small constant gap from the capacity.

VI. Conclusion

The results of this paper open the door for a number of extensions. For instance, it is of interest to find the capacity region $\mathcal{C}(C_1, C)$ in the regime where C is not large enough for AP-2 to forward (J_1, Y_2^n) to the CP without loss. Perhaps the most straightforward approach here is to separately apply binning to Y_2^n at AP-2, and forward the corresponding bin index J_2 to the CP alongside AP-1's bin index J_1. Alternatively, AP-2 can decode J_1 and retrieve the compressed version of Y_1^n, denoted by U_1^n, and then jointly compress U_1^n and Y_2^n into a new index J, which is then forwarded to the CP. In a related cascade multi-terminal source coding problem [12], it has been observed that both of the above approaches can be sub-optimal in general. Whether a same observation holds in the cascade CRAN setting at hand remains unknown. Another interesting direction is to generalize the results to $L > 2$ relays. In this context, variants of the schemes in [8] can be useful.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860023.

References

[1] T. Q. Quek, O. Simeone, and W. Yu, *Cloud radio access networks: Principles, technologies, and applications*. Cambridge University Press, 2017.

[2] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer, "Communication via decentralized processing," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3008–3023, 2008.

[3] O. Simeone, E. Erkip, and S. Shamai, "On codebook information for interference relay channels with out-of-band relaying," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2880–2888, 2011.

[4] I. Estella Aguerri, A. Zaidi, G. Caire, and S. Shamai Shitz, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4575–4596, 2019.

[5] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Network.*, vol. 2019, no. 1, pp. 1–13, 2019.

[6] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes," *arXiv:2012.13928*, 2020.

[7] 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, 2016.

[8] I. Estella Aguerri and A. Zaidi, "In-network compression for multiterminal cascade MIMO systems," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4176–4187, 2017.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006.

[10] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[11] G. Kramer, "Topics in multi-user information theory," *Found. Trends Commun. Inf. Theory*, vol. 4, no. 4–5, pp. 265–444, 2008.

[12] P. Cuff, H.-I. Su, and A. El Gamal, "Cascade multiterminal source coding," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 1199–1203.