Alternating Information Bottleneck Optimization for Weighted Sum Rate and Resource Allocation in the Uplink of C-RAN

Di Chen
Institute of Communications Engineering
University of Rostock
Email: di.chen@uni-rostock.de

Volker Kuehn
Institute of Communications Engineering
University of Rostock
Email: volker.kuehn@uni-rostock.de

***Abstract*—The quantizer design and resource allocation in the uplink of Cloud Radio Access Network (C-RAN) are studied. In C-RAN, multiple Radio Units (RUs) act as soft relays by compressing and forwarding the correlated received signals simultaneously to the Central Processor (CP) in the cloud, via the fronthaul links with finite capacities. Wyner-Ziv coding is utilized in order to exploit the correlation between the received signals at neighboring RUs. Thus, a joint optimization of the quantizers is necessary. Moreover, when the capacity resource is shared among fronthauls, the design of quantizers is closely related to the resource allocation. In this paper we aim to maximize the achievable weighted sum rate by joint optimizing all the quantizers and resource allocation. To make the problem more tractable, at first we assume that the resource allocation is predetermined, and perform a joint optimization of all quantizers, the optimization algorithm is a combination of the Alternating Information Bottleneck (AIB) method and the Alternating Bi-Section method, which is proposed in our previous work. We extend it to solve the problem in this work. Then we optimize the resource allocation based on it. The simulation results justify the correctness and effectiveness of our proposed algorithms.**

I. Introduction

It has been shown that a key challenge of C-RAN [1] is the transfer of baseband information to CP via the capacity-constrained fronthaul links [2]. Thus suitable compression strategies have to be developed in order to alleviate the requirements on the fronthaul links. Basically, two compression schemes can be exploited, e.g., point to point compression and distributed compression [2]. For point to point compression, each RU performs compression individually according to its allocated fronthaul capacity. While for distributed compression, the correlation between the signals received at neighboring RUs is exploited, in order to further utilize the capacity of the fronthaul. In this case, Wyner-Ziv coding [3] is used at the compression step. It has been shown that the distributed compression outperforms the point to point compression scheme [2], [4]. While it should be noted that this scheme has a higher complexity since a joint optimization among all RUs is required. While for the point to point compression scheme, the compression scheme can be optimized locally without the knowledge of neighboring RU. In C-RAN, the CP with access to all RUs and high computing capability makes this joint optimization possible. There has been already some papers considering the optimization of quantization noise levels, when Compress and Forward (CF) or Noisy Network Coding (NNC) [5] is performed at RUs, e.g. [4] and [6]. While these works consider only Gaussian codebooks, and treat the quantization as Gaussian test channels, the compression is modeled by adding Gaussian distributed quantization noise. Optimization of the quantization noise levels evaluates the performance only from the information theoretic perspective. In practice, the users might use arbitrary codebooks $\mathcal{X}$ with finite alphabet, and the received signal at RU is discretized and sampled firstly into finite alphabet $\mathcal{Y}$, then based on the compression scheme denoted by $P_{\hat{Y}|Y}$, it will be compressed into several quantization levels, with alphabet $\hat{\mathcal{Y}}$. Usually $|\hat{\mathcal{Y}}|$ is much smaller than $|\mathcal{Y}|$ due to the compression. In such scenarios, the Information Bottleneck (IB) method [7] is often used to optimize the quantizer $P_{\hat{Y}|Y}$ in order to maximize the objective mutual information. While in most of the previous works, only one quantizer is considered, the IB method is utilized for different optimization objectives, e.g., [8], [9]. When it comes to the multi-quantizer case where Wyner-Ziv coding is exploited, a joint optimization is required as we said above. The problem is whether the IB method can still be used. In our work [10], we propose the so-called Alternating Information Bottleneck (AIB) method and the Alternating Bi-Section method, which have been shown to be useful tools for the joint optimization of quantizers in C-RAN. While in that work, we aim to maximize the achievable sum rate in the uplink of C-RAN with fixed or predetermined individual fronthaul capacities, this may happen when the fronthaul are optical fiber links, the optimization of each quantizer depends on the fronthaul capacity allocated to that RU. While the fronthaul might also be wireless, where the fronthaul capacity can be dynamically shared among RUs. When the problem of resource allocation is considered, the joint optimization of all quantizers and resource allocation is necessary for different optimization objectives. In this paper, we address this resource allocation problem and propose an optimization scheme for it, which is based on our previously proposed AIB method and Alternating Bi-Section method and newly introduced Outer Linearization Method (OLM) [11]. Moreover, we consider the maximization of the weighted sum rate. We show that resource allocation is critical in this case.

The remainder of the paper is organized as follows: In Sec. II we introduce the channel model considered and state the problem mathematically. Our optimization algorithms of the quantizers are presented and explained in Sec. III. The

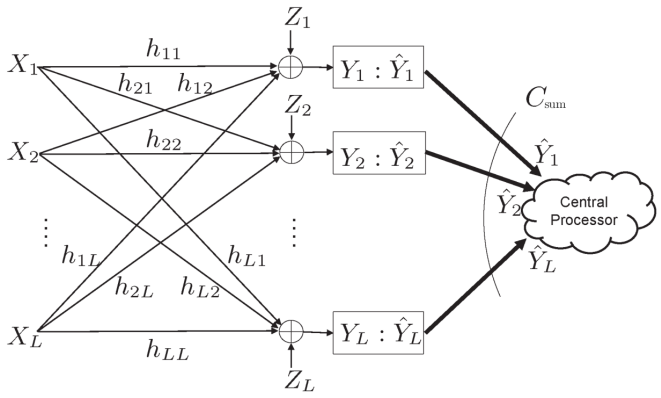

Fig. 1. The uplink of C-RAN with finite sum capacity fronthaul links [4].

optimization algorithm of capacity allocation is shown in Sec. IV. Simulation results and conclusions are provided in Sec. V and Sec. VI respectively.

II. System Model and Problem Statement

A. System Model

We consider the C-RAN model depicted in Fig. 1, where L single-antenna Mobile Users (MSs) send independent messages to L single-antenna RUs. All RUs connect to a CP in the cloud via fronthaul links with finite sum capacity denoted by C_{sum}, which will be shared among all fronthaul links. CP is expected to decode all messages. We consider only single antenna for simplicity, but the algorithm can be extended to the MIMO case, as shown in [10]. The channel model between the MSs and RUs is actually an $L \times L$ interference channel. The received analog signal at the i-th BS is

$$Y_{i,\text{analog}} = \sum_{j=1}^{L} h_{ij} X_j + Z_i, \ \ i \in \{1, 2, ..L\},$$

where $Z_i \sim \mathcal{CN}(0, \sigma_n^2)$ is the independent Gaussian noise with variance σ_n^2, and h_{ij} denotes the complex channel coefficient from j-th MS to i-th RU. X_j denotes the transmitted signal of j-th MS, arbitrary modulation scheme can be utilized. The available power of j-th MS is denoted as $P_j = \mathbb{E}\{|X_j|^2\}$. At i-the RU, the received analog signal $Y_{i,\text{analog}}$ is firstly sampled and discretized[1] into Y_i with finite alphabets $\mathcal{Y}_i$. Then each RU performs Compress and Forward (CF): Its quantizer compresses the signal Y_i into $\hat{Y}_i$ based on the compression scheme denoted by $P_{\hat{Y}_i|Y_i}$. $|\hat{\mathcal{Y}}_i|$ is assumed to be much smaller than $|\mathcal{Y}_i|$. At each RU, Wyner-Ziv coding is to be utilized by exploiting the correlation between the received signals of its neighboring RUs. Then RUs independently send compressed bits to the CP via the fronthaul links. The CP employs successive two-stage decoding: It first decodes all the compressed signals $\hat{\boldsymbol{Y}} = [\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_L]^T$ and then decodes MSs' messages $\boldsymbol{X} = [X_1, X_2, ..., X_L]^T$ based on the decoded compressed signals. Compared to NNC, where simultaneous joint decoding of compressed signals and the desired messages over all received blocks is required [5], the successive decoding nature of CF overcomes some difficulties in the practical implementation of NNC, such as long delay and high computational complexity. Moreover, we assume that the modulation scheme of each MS and CSI are known to the CP, and the design of the optimized quantizers can be feedbacked to the corresponding RU.

[1]This discretization is actually a pre-quantization, then the discretized signal should be further compressed by the quantizer due to limited fronthaul capacity. In this paper we address the optimization of the quantizer used for the compression.

B. Problem Statement

We aim to maximize the achievable weighted sum rate [2] in the uplink of C-RAN as follows.

$$\begin{aligned} &\max_{P_{\hat{\boldsymbol{Y}}|\boldsymbol{Y}}} \sum_{j=1}^{L} w_j R_j \\ &\text{Subject to } I(\boldsymbol{Y}; \hat{\boldsymbol{Y}}) \le C_{\text{sum}}, \end{aligned} \tag{1}$$

where $P_{\hat{\boldsymbol{Y}}|\boldsymbol{Y}} = \prod_{i=1}^{L} P_{\hat{Y}_i|Y_i}$. Naturally $\sum_{|\hat{y}_i|} P_{\hat{Y}_i|Y_i} = 1, \ \forall y_i$ and $P_{\hat{Y}_i|Y_i} \ge 0, \ \forall \hat{y}_i, y_i$ should be satisfied. R_j and w_j denote the achievable rate of j-th MS and its weight, respectively. $P_{\hat{Y}_i|Y_i}$ denotes the compression scheme of the quantizer at i-th RU, which should be jointly optimized at the CP. From CP's perspective, the network is actually MIMO-MAC, the capacity-achieving strategy in the MIMO-MAC is based on Successive Interference Cancellation (SIC). Moreover, according to [12], the solution of (1) is given by the decoding order π that sorts the weights in non-increasing order

$$w_{\pi_1} \ge w_{\pi_2} \ge \cdots \ge w_{\pi_L}.$$

With this decoding order, the resulting maximization problem is convex [12]. Since the decoding order is fixed solely by the weights, without loss of generality, we assume $w_L \ge w_{L-1} \ge \cdots \ge w_1$, i.e., x_1 is decoded first and x_L is decoded last. Thus we have

$$R_j = I(X_j; \hat{\boldsymbol{Y}} | X_1, X_2, ..., X_{j-1}) \ \forall j \in \{1, 2, ..., L\} \tag{2}$$

The constraint of (1) can also be expressed as

$$\begin{aligned} &I(Y_i; \hat{Y}_i | \hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_{i-1}) \le C_i \ \forall i \in \{1, 2, ..., L\}, \\ &\sum_{i=1}^{L} C_i \le C_{\text{sum}}. \end{aligned} \tag{3}$$

where C_i denote the capacity allocated to i-th RU. It can be predetermined or optimized by the CP. We see that when the modulations schemes, the capacity of each fronthaul link and all channel configurations are fixed, the weighted sum rate depends only on how RUs compress their received signals. While when the capacity of each fronthaul link is not predetermined, a simultaneous optimization of all quantizers and capacity allocation is necessary. In order to make the problem more tractable, we firstly assume that the capacity of each fronthaul link is predetermined, and propose algorithms to optimize all quantizers jointly, which are based on the AIB method and the Alternating Bi-Section method proposed in [10]. Then we adopt the Outer Linearization Method (OLM) and propose an algorithm for the optimization of capacity allocation.

III. Optimization Algorithm and Quantizer Design (With predetermined fronthaul capacity allocation)

In this section we assume the capacity of each fronthaul link is predetermined. Firstly, we give a brief introduction to the well-known Information Bottleneck (IB) method.

Since the capacity of each fronthaul is limited. On one hand, the quantization in the compression step cannot be too fine, such that the resultant compression rate exceeds the fronthaul capacity. The compressed signal cannot be decoded at CP. On the other hand, too coarse quantization cannot fully utilize the fronthaul capacity. The overall performance is limited. Hence, an optimal tradeoff between the compression rates that can be supported and the achievable sum rate must be found. The IB method is an effective tool to find this tradeoff as well as the corresponding optimized quantizer.

The tradeoff problem described above can be mathematically formulated as follows. Consider three variables $X \rightarrow Y \rightarrow \hat{Y}$ forming a Markov chain, where $\hat{Y}$ is the compression of Y. We want the variable $\hat{Y}$ to compress Y as much as possible (smaller $I(Y;\hat{Y})$), while $\hat{Y}$ captures as much of the information about X as possible (larger $I(X;\hat{Y})$). The IB method proposed in [7] has been shown to be an useful tool to solve this problem. According to [7] and [9], the maximized $I(X;\hat{Y})$ as the function of the compression rate $I(Y;\hat{Y})$ can be numerically computed with the IB method, e.g.,

$$I(c) = \sup_{I(Y;\hat{Y})\leq c} I(X;\hat{Y})$$

can be computed and plotted. It has been proved to be a concave and increasing function for $c \in [0, H(\hat{Y})]$. The IB method is a *deterministic annealing* approach such that the whole curve $I(c)$ is obtained through a third parameter $\beta,\ \beta > 0$, where $1/\beta = \frac{\mathrm{d}I(c)}{\mathrm{d}c}$ corresponds to the slope of the curve at the point $(c, I(c))$. Actually β is the *Lagrange Multiplier* used in the IB method. We call it the tradeoff factor between the compression rate c and the objective mutual information. By choosing an arbitrary $\beta > 0$ as the input of the IB method, the point on the tradeoff curve with slope $1/\beta$ can be outputted. Since $I(c)$ is concave and increasing, the output compression rate c of the IB method increases with the input β, the whole tradeoff curve can be obtained by ranging the value of β from 0 to infinity and running the IB method repeatedly. After obtaining this tradeoff curve, we can use the Bi-Section method to find the specific value of β such that at this point the compression rate c can be supported and the objective mutual information is maximized.

In the uplink of C-RAN, a joint optimization among all the quantizers has to be performed. In our work [10], we proposed the Alternating Information Bottleneck (AIB) method and the Alternating Bi-Section method to achieve this goal. The basic idea of AIB method is to fix all the other quantizers, and adopt the IB method to optimize one quantizer. Then fix this newly optimized quantizer, and use the IB method to optimize the next quantizer. This procedure is carried on alternatively until reaching the convergence. The illustration for optimality and convergence is in [10]. It has been shown there, with AIB method, the trade-off between the compression rate vector $\Big(I(Y_1;\hat{Y}_1), I(Y_2;\hat{Y}_2|\hat{Y}_1), ..., I(Y_L;\hat{Y}_L|\hat{Y}_{L-1}, ..., \hat{Y}_1)\Big)$ of all RUs and the corresponding maximized sum rate $R_{\mathrm{sum}} = I(\boldsymbol{X};\hat{\boldsymbol{Y}})$ can be set up through the trade-off factor vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_L)$. For the specific predetermined fronthaul capacities, the proposed Alternating Bi-Section method is adopted to locate and compute the optimal trade-off factor vector $\boldsymbol{\beta}^{\mathrm{opt}}$ and optimized quantizers, such that its corresponding compression rate vector fulfills the predetermined fronthaul capacities simultaneously. More details can be found in [10].

Now we go back to the problem (1) addressed in this paper. In this subsection we assume that the each fronthaul capacity is predetermined, thus (1) becomes similarly to the problem we solved in [10]. For the ease of illustration, we start with the 2 MSs and 2 RUs case to show the optimization scheme, and at the end of this section, we will show our proposed optimization algorithms can be easily extended to the case with more devices. According to (1), the problem becomes

$$\begin{aligned} \max_{P_{\hat{Y}_1|Y_1} P_{\hat{Y}_2|Y_2}} \quad & w_1 I(X_1;\hat{Y}_1,\hat{Y}_2) + w_2 I(X_2;\hat{Y}_1,\hat{Y}_2|X_1), \\ \text{Subject to } & I(Y_1;\hat{Y}_1) \leq C_1, \\ & I(Y_2;\hat{Y}_2|\hat{Y}_1) \leq C_2. \end{aligned} \tag{4}$$

For arbitrary capacity allocation such that $C_1 + C_2 = C_{\mathrm{sum}}$ is fulfilled. Let $R_{\mathrm{wsum}} = w_1 I(X_1;\hat{Y}_1,\hat{Y}_2) + w_2 I(X_2;\hat{Y}_1,\hat{Y}_2|X_1)$. Similar to the steps in [10], with the AIB method, we firstly set up the trade-off between the compression rate pair of 2 RUs and the corresponding maximized weighted sum rate R_{wsum} in Sec. III-A. Then in Sec. III-B, with the Alternating Bi-Section method, we locate the specific point where the constraints in (4) is satisfied and the weighted sum rate is maximized.

A. The Alternating Information Bottleneck (AIB) method – Setting up the trade-off between the compression rate pair $\Big(I(Y_1;\hat{Y}_1), I(Y_2;\hat{Y}_2|\hat{Y}_1)\Big)$ and the maximized weighted sum rate R_{wsum}, through the trade-off factor pair (β_1, β_2)

In this subsection we try to set up the trade-off between the compression rate pair $\Big(I(Y_1;\hat{Y}_1), I(Y_2;\hat{Y}_2|\hat{Y}_1)\Big)$ and its corresponding maximized weighted sum rate R_{wsum}. Since the two quantizers need to be optimized jointly, the IB method cannot be used directly. However, as we have shown in [10], when one quantizer is fixed, the other quantizer can be readily optimized by the IB method.

1. Now we assume that the first quantizer $P_{\hat{Y}_1|Y_1}$ is fixed with an arbitrary valid distribution, then we need to find the optimal trade-off between the compression rate $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $\max_{P_{\hat{Y}_2|Y_2}} R_{\mathrm{wsum}}$. Because of the chain rule, we have

$$\begin{aligned} R_{\mathrm{wsum}} =& w_1 I(X_1;\hat{Y}_1,\hat{Y}_2) + w_2 I(X_2;\hat{Y}_1,\hat{Y}_2|X_1) \\ =& w_1 I(X_1;\hat{Y}_1) + w_2 I(X_2;\hat{Y}_1|X_1) \\ &+ w_1 I(X_1;\hat{Y}_2|\hat{Y}_1) + w_2 I(X_2;\hat{Y}_2|\hat{Y}_1, X_1). \end{aligned} \tag{5}$$

we see that when the quantizer $P_{\hat{Y}_1|Y_1}$ is fixed, it is sufficient to compute the trade-off between $I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and $\max_{P_{\hat{Y}_2|Y_2}} \Big(w_1 I(X_1;\hat{Y}_2|\hat{Y}_1) + w_2 I(X_2;\hat{Y}_2|\hat{Y}_1, X_1)\Big) = \max_{P_{\hat{Y}_2|Y_2}} \Big(w_1 I(X_1, X_2;\hat{Y}_2|\hat{Y}_1) + (w_2 - w_1) I(X_2;\hat{Y}_2|\hat{Y}_1, X_1)\Big)$.

We follow the similar procedure proposed in [9] to find this trade-off by the IB method. We have adopted this idea in [10], while in that paper, the sum rate is considered. The algorithm *Function IB2* of optimizing quantizer $P_{\hat{Y}_2|Y_2}$ conditioned upon fixed $P_{\hat{Y}_1|Y_1}$ is on the right part of this page. The fixed quantizer $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$ is the input and an local invariant. The *Lagrange Multiplier* $\beta_2 > 0$ is the trade-off factor between $I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and R_{wsum} and ϵ_1 denotes the tolerance and $D_{\text{KL}}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

2. Similarly, when the second quantizer $P_{\hat{Y}_2|Y_2}$ is fixed, the trade-off points $\left\{I(Y_1;\hat{Y}_1), \max\limits_{P_{\hat{Y}_1|Y_1}} R_{\text{wsum}}\right\}$ can also be obtained with the IB method, as summarized in *Function IB1*. More details of obtaining these two functions can be found in [9] and [10].

After obtaining these two functions, we alternatively run them in order to optimize both quantizers. The optimized quantizer obtained from one IB function is set as the input fixed quantizer of the other, until reaching the convergence. as shown in *Function AIB*, where ϵ_2 denotes the tolerance. With this AIB method, the trade-off between the compression rate pair $\left(I(Y_1;\hat{Y}_1), I(Y_2;\hat{Y}_2|\hat{Y}_1)\right)$ and the maximized weighted sum rate R_{wsum} can be set up, through the input trade-off factor pair (β_1,β_2). Since the problem is generally non-convex with respect to $P_{\hat{\boldsymbol{Y}}|\boldsymbol{Y}}$, similar to the IB method, we can try different valid initial distributions to get better results. By inputing different trade-off factor pair (β_1,β_2) to the AIB function, different trade-off between the compression rate pair and the maximized weighted sum rate can be obtained.

Function IB2($P^{\text{fixed}}_{\hat{Y}_1|Y_1},\ P^{\text{init}}_{\hat{Y}_2|Y_2},\ |\hat{\mathcal{Y}}_2|,\ \beta_2,\ \epsilon_1$)

Input : $P^{\text{fixed}}_{\hat{Y}_1|Y_1},\ P^{\text{init}}_{\hat{Y}_2|Y_2},\ |\hat{\mathcal{Y}}_2|,\ \beta_2,\ \epsilon_1$
Output : $[P^{\text{optimal}}_{\hat{Y}_2|Y_2},\ c_2,\ R_{\text{wsum}}]$

begin
  Initialization: $k \leftarrow 0$, set the initial mapping $P^{(k)}_{\hat{Y}_2|Y_2} \leftarrow P^{\text{init}}_{\hat{Y}_2|Y_2}$.
  do
    Based on $P^{\text{fixed}}_{\hat{Y}_1|Y_1}$ and newly generated $P^{(k)}_{\hat{Y}_2|Y_2}$,
    update $d^{(k)}(y_2,\hat{y}_2) \leftarrow$
    $w_1\beta_2 \sum\limits_{\hat{y}_1} P_{\hat{Y}_1|Y_2} D_{\text{KL}}\left(P_{X_1X_2|\hat{Y}_1Y_2}||P^{(k)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right) -$
    $\sum\limits_{\hat{y}_1} P_{\hat{Y}_1|Y_2} \log_2\left(P^{(k)}_{\hat{Y}_2|\hat{Y}_1}\right) + \log_2\left(P^{(k)}_{\hat{Y}_2}\right) + (w_2 -$
    $w_1)\beta_2 \sum\limits_{\hat{y}_1} P_{\hat{Y}_1X_2|Y_2} D_{\text{KL}}\left(P_{X_1|\hat{Y}_1Y_2X_2}||P^{(k)}_{X_1|\hat{Y}_1\hat{Y}_2X_2}\right)$
    Set $P^{(k+1)}_{\hat{Y}_2|Y_2} \leftarrow$
    $P^{(k)}_{\hat{Y}_2} 2^{-d^{(k)}(y_2,\hat{y}_2)} / \sum\limits_{\hat{y}_2} P^{(k)}_{\hat{Y}_2} 2^{-d^{(k)}(y_2,\hat{y}_2)}$
    Set $k \leftarrow k+1$
  while $\sum\limits_{y_2,\hat{y}_2} \left|P^{(k)}_{\hat{Y}_2|Y_2} - P^{(k-1)}_{\hat{Y}_2|Y_2}\right| / (|\mathcal{Y}_2| \cdot |\hat{\mathcal{Y}}_2|) \geq \epsilon_1$
  Set $P^{\text{optimal}}_{\hat{Y}_2|Y_2} \leftarrow P^{(k)}_{\hat{Y}_2|Y_2}$, based on $P^{\text{optimal}}_{\hat{Y}_2|Y_2}$, compute $c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and R_{wsum}

Function IB1($P^{\text{fixed}}_{\hat{Y}_2|Y_2},\ P^{\text{init}}_{\hat{Y}_1|Y_1},\ |\hat{\mathcal{Y}}_1|,\ \beta_1,\ \epsilon_1$)

Input : $P^{\text{fixed}}_{\hat{Y}_2|Y_2},\ P^{\text{init}}_{\hat{Y}_1|Y_1},\ |\hat{\mathcal{Y}}_1|,\ \beta_1,\ \epsilon_1$
Output : $[P^{\text{optimal}}_{\hat{Y}_1|Y_1},\ c_1,\ R_{\text{wsum}}]$

begin
  Initialization: $k \leftarrow 0$, set the initial mapping $P^{(k)}_{\hat{Y}_1|Y_1} \leftarrow P^{\text{init}}_{\hat{Y}_1|Y_1}$.
  do
    Based on $P^{\text{fixed}}_{\hat{Y}_2|Y_2}$ and newly generated $P^{(k)}_{\hat{Y}_1|Y_1}$,
    update $d^{(k)}(y_1,\hat{y}_1) \leftarrow$
    $w_1\beta_1 \sum\limits_{\hat{y}_2} P_{\hat{Y}_2|Y_1} D_{\text{KL}}\left(P_{X_1X_2|Y_1\hat{Y}_2}||P^{(k)}_{X_1X_2|\hat{Y}_1\hat{Y}_2}\right) +$
    $(w_2 -$
    $w_1)\beta_1 \sum\limits_{\hat{y}_2} P_{\hat{Y}_2X_2|Y_1} D_{\text{KL}}\left(P_{X_1|Y_1\hat{Y}_2X_2}||P^{(k)}_{X_1|\hat{Y}_1\hat{Y}_2X_2}\right)$
    Set $P^{(k+1)}_{\hat{Y}_1|Y_1} \leftarrow$
    $P^{(k)}_{\hat{Y}_1} 2^{-d^{(k)}(y_1,\hat{y}_1)} / \sum\limits_{\hat{y}_1} P^{(k)}_{\hat{Y}_1} 2^{-d^{(k)}(y_1,\hat{y}_1)}$
    Set $k \leftarrow k+1$
  while $\sum\limits_{y_1,\hat{y}_1} \left|P^{(k)}_{\hat{Y}_1|Y_1} - P^{(k-1)}_{\hat{Y}_1|Y_1}\right| / (|\mathcal{Y}_1| \cdot |\hat{\mathcal{Y}}_1|) \geq \epsilon_1$
  Set $P^{\text{optimal}}_{\hat{Y}_1|Y_1} \leftarrow P^{(k)}_{\hat{Y}_1|Y_1}$, based on $P^{\text{optimal}}_{\hat{Y}_1|Y_1}$, compute $c_1 = I(Y_1;\hat{Y}_1)$ and R_{wsum}

Function AIB($|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \beta_1,\ \beta_2,\ \epsilon_1,\epsilon_2$)

Input : $|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \beta_1,\ \beta_2,\ \epsilon_1,\epsilon_2$
Output : $[c_1,\ c_2,\ R_{\text{wsum}}]$
OptionalOutput: $[P^{\text{optimal}}_{\hat{Y}_1|Y_1},\ P^{\text{optimal}}_{\hat{Y}_2|Y_2}]$

begin
  Initialization : Randomly choose valid initial mappings $P^{(0)}_{\hat{Y}_1|Y_1}$ and $P^{(0)}_{\hat{Y}_2|Y_2}$
    Set $\ell \leftarrow 0$
  do
    Run function IB1 :
    $P^{(\ell+1)}_{\hat{Y}_1|Y_1} = \text{IB1}(P^{(\ell)}_{\hat{Y}_2|Y_2},\ P^{(\ell)}_{\hat{Y}_1|Y_1},\ |\hat{\mathcal{Y}}_1|,\ \beta_1,\ \epsilon_1)$
    Run function IB2 :
    $P^{(\ell+1)}_{\hat{Y}_2|Y_2} = \text{IB2}(P^{(\ell+1)}_{\hat{Y}_1|Y_1},\ P^{(\ell)}_{\hat{Y}_2|Y_2},\ |\hat{\mathcal{Y}}_2|,\ \beta_2,\ \epsilon_1)$
    Set $\ell \leftarrow \ell+1$
  while $\sum\limits_{y_1,\hat{y}_1} \left|P^{(\ell)}_{\hat{Y}_1|Y_1} - P^{(\ell-1)}_{\hat{Y}_1|Y_1}\right| / (|\mathcal{Y}_1| \cdot |\hat{\mathcal{Y}}_1|) +$
  $\sum\limits_{y_2,\hat{y}_2} \left|P^{(\ell)}_{\hat{Y}_2|Y_2} - P^{(\ell-1)}_{\hat{Y}_2|Y_2}\right| / (|\mathcal{Y}_2| \cdot |\hat{\mathcal{Y}}_2|) \geq \epsilon_2$
  Set $P^{\text{optimal}}_{\hat{Y}_1|Y_1} \leftarrow P^{(\ell)}_{\hat{Y}_1|Y_1},\ P^{\text{optimal}}_{\hat{Y}_2|Y_2} \leftarrow P^{(\ell)}_{\hat{Y}_2|Y_2}$, based on them, compute
  $c_1 = I(Y_1;\hat{Y}_1),\ c_2 = I(Y_2;\hat{Y}_2|\hat{Y}_1)$ and R_{wsum}

B. The Alternating Bi-Section method – Locating the optimal trade-off point

The AIB method shown in the last subsection can only be used to set up different trade-offs While in order to solve the problem in (4), only the trade-offs with resultant compression rate pair that simultaneously fulfill the conditions in (4) are valid. Moreover, in order to fully utilize the fronthaul links, the specific trade-off where $I(Y_1;\hat{Y}_1) = C_1$ and $I(Y_2;\hat{Y}_2|\hat{Y}_1) = C_2$ is what has to be located. If it can be located, its corresponding maximized weighted sum rate R_{wsum} and the quantizers $P_{\hat{Y}_1|Y_1},\ P_{\hat{Y}_2|Y_2}$ are the solution for (4). Naturally, this trade-off can be located by inserting different trade-off factor pairs (β_1, β_2) over a sufficiently fine grid of values until finding the point where $c_1 = C_1$ and $c_2 = C_2$. Obviously this approach is rather inefficient. In our work [10], we propose a so-called Alternating Bi-Section method for efficient location. This method can be readily used in the problem considered in this paper. This idea is somehow similar to the AIB method, we fix one trade-off factor $\beta_i,\ i \in \{1, 2\}$ and combine the AIB method and Bi-Section method to locate the other trade-off factor, such that its corresponding compression rate just satisfy its fronthaul constraint. Then we fix this newly located trade-off factor and locate the other. This procedure is also carried on alternatively, as shown in Algorithm AIB . $[\beta_{i\min}, \beta_{i\max}]$ denotes the searching range of β_i. C_i is the target compression rate of the i-th RU, which equals to the capacity of its fronthaul link, $\eta_1,\ \eta_2$ are the tolerance. More details can be found in our paper [10].

Algorithm: Alternating Bi-Section method

Input : $|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \beta_{1\max},\ \beta_{1\min},\ \beta_{2\max},\ \beta_{2\min}$
Input : $C_1,\ C_2,\ \epsilon_1, \epsilon_2,\ \eta_1,\ \eta_2$
Output : $R_{\text{wsum}}, P^{\text{optimal}}_{\hat{Y}_1|Y_1}, P^{\text{optimal}}_{\hat{Y}_2|Y_2}$
OptionalOutput: $c_1,\ c_2$

begin
  Set $\ell \leftarrow 0,\ \beta_1^{(0)} \leftarrow (\beta_{1\max} + \beta_{1\min})/2,$
  $\beta_2^{(0)} \leftarrow (\beta_{2\max} + \beta_{2\min})/2$
  do
    Set $\beta_{1\text{U}} \leftarrow \beta_{1\max},\ \beta_{1\text{L}} \leftarrow \beta_{1\min}$
    while $\beta_{1\text{U}} - \beta_{1\text{L}} > \eta_1$ **do**
      Set $\widetilde{\beta}_1 \leftarrow (\beta_{1\text{U}} + \beta_{1\text{L}})/2$
      Run Function AIB: $[c_1, \sim, \sim]$=AIB $(|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \widetilde{\beta}_1,\ \beta_2^{(\ell)},\ \epsilon_1, \epsilon_2)$
      if $c_1 < C_1$ **then**
        Set $\beta_{1\text{L}} \leftarrow \widetilde{\beta}_1$
      else
        Set $\beta_{1\text{U}} \leftarrow \widetilde{\beta}_1$
    Set $\beta_1^{(\ell+1)} \leftarrow (\beta_{1\text{U}} + \beta_{1\text{L}})/2$
    Set $\beta_{2\text{U}} \leftarrow \beta_{2\max},\ \beta_{2\text{L}} \leftarrow \beta_{2\min}$
    while $\beta_{2\text{U}} - \beta_{2\text{L}} > \eta_1$ **do**
      Set $\widetilde{\beta}_2 \leftarrow (\beta_{2\text{U}} + \beta_{2\text{L}})/2$
      Run Function AIB: $[\sim, c_2, \sim]$=AIB $(|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \beta_1^{(\ell+1)},\ \widetilde{\beta}_2,\ \epsilon_1, \epsilon_2)$
      if $c_2 < C_2$ **then**
        Set $\beta_{2\text{L}} \leftarrow \widetilde{\beta}_2$
      else
        Set $\beta_{2\text{U}} \leftarrow \widetilde{\beta}_2$
    Set $\beta_2^{(\ell+1)} \leftarrow (\beta_{2\text{U}} + \beta_{2\text{L}})/2$
    Set $\ell \leftarrow \ell + 1$
  while $|\beta_1^{(\ell)} - \beta_1^{(\ell-1)}| + |\beta_2^{(\ell)} - \beta_2^{(\ell-1)}| \geq \eta_2$
  Run Function AIB: $[c_1, c_2, R_{\text{wsum}}, P^{\text{optimal}}_{\hat{Y}_1|Y_1}, P^{\text{optimal}}_{\hat{Y}_2|Y_2}]$=AIB $(|\hat{\mathcal{Y}}_1|,\ |\hat{\mathcal{Y}}_2|,\ \beta_1^{(\ell)},\ \beta_2^{(\ell)},\ \epsilon_1, \epsilon_2)$

C. Extension to more MSs and RUs with multiple antennas

This alternative idea can be readily extended to the case where more MSs and RUs exist in C-RAN. A specific quantizer can be optimized with the IB method, while fix all the other quantizers. The optimized quantizer is set as the starting point to optimize the next quantizer. This procedure can be carried on alternatively until reaching convergence. For the multiple antenna case, Wyner-Ziv coding is also to be utilized since the received signals between different antennas of each RU are statistically correlated. For a specific RU, different quantizers will compress the received signals at different antennas by exploiting the correlation. The proposed algorithm can also be readily utilized in this case.

IV. Optimization Scheme for the capacity allocation

In the last section, we assume that the fronthaul capacity allocated to each RU is predetermined. We use the AIB method to set up the trade-off between the compression rate vector and its corresponding maximized weighted sum rate, through the help of trade-off factor vector. Then we exploit the Alternating Bi-Section method to locate the specific trade-off factor vector whose corresponding compression rate vector exactly fulfill the predetermined fronthaul capacity constraints simultaneously. In this section, we address the problem of resource allocation, where the total fronthaul capacity is to be allocated to each RU in order to maximized the achievable weighted sum rate in the uplink of C-RAN, by exploiting the algorithms proposed in the last section. We propose the algorithm by combining the AIB method, the Alternating Bi-Section method with the Outer Linearization Method (OLM) [11].

Note that in the original problem (1), the objective function is a concave function of the compression rates, and the sum of all compression rates is limited by the sum capacity C_{sum}, which is a linear inequality constraint. Thus the original problem is a convex optimization problem with respect to the compression rate vector $\boldsymbol{c}$. Since the AIB method and the Bi-Section method can be easily used to compute the value of the objective function for different $\boldsymbol{c}$, the original problem (1) can be solved by standard convex optimization methods. Here, similar to [9], we adopt the the Outer Linearization Method and list the procedures as follows.

1. Start with a random valid capacity allocation, $\boldsymbol{C}^{(0)} = \left(C_1^{(0)}, C_2^{(0)}, ..., C_L^{(0)}\right)$, such that $\sum_{i=1}^{L} C_i^{(0)} = C_{\text{sum}}$. Set $k = 0$, $f_{\text{LB}} = -1$ and f_{UB} to be large enough. Set δ be the desired tolerance.

At iteration k, repeat step 2 to step 4 until $f_{\text{UB}} - f_{\text{LB}} \leq \delta$.

2. Use the Alternating Bi-Section method to compute the

trade-off factor vector $\boldsymbol{\beta}^{(k)} = \left(\beta_1^{(k)}, \beta_2^{(k)}, ..., \beta_L^{(k)}\right)$ associated with the current capacity allocation $\boldsymbol{C}^{(k)}$.

3. Insert $\boldsymbol{\beta}^{(k)}$ to the AIB method, then compute current maximized weighted sum rate $R_{\text{wsum}}^{(k)}$. Set $f_{\text{LB}} = R_{\text{wsum}}^{(k)}$ and the sub-gradient $\boldsymbol{g}^{(k)} = \left(1/\beta_1^{(k)}, 1/\beta_2^{(k)}, ..., 1/\beta_L^{(k)}\right)$ and $b^{(k)} = R_{\text{wsum}}^{(k)} - \boldsymbol{C}^{(k)} \cdot (\boldsymbol{g}^{(k)})^T$.

4. Solve the linear problem

$$\begin{aligned} &\max_{s,\boldsymbol{C}} s \\ &s.t.\ \boldsymbol{C} \cdot (\boldsymbol{g}^{(\ell)})^T + b^{(\ell)} \geq s,\ \ell = 0, 1, ..., k-1, \\ &\sum_{i=1}^{L} C_i = C_{\text{sum}} \end{aligned}$$

Let $(s^*, \boldsymbol{C}^*)$ be the maximizer, set $f_{\text{UB}} = s^*$ and $\boldsymbol{C}^{(k+1)} = \boldsymbol{C}^*$. Set $k = k + 1$.

V. Simulation Results

In this section, we evaluate the performance of the proposed algorithms and use them to study C-RAN.

General Setup: In the simulation we assume that all MSs use BPSK modulation for simplicity. All received signals at the RUs are sampled and discretized with 7 bits/sample, thus $|\mathcal{Y}_i| = 128$. The RUs will compress the signals into 8 quantization levels (at most 3 bits/sample), thus $|\hat{\mathcal{Y}}_i| = 8$. Moreover, we set $\beta_{1\max} = \beta_{2\max} = 260$, $\beta_{1\min} = \beta_{2\min} = 0.1$, $\epsilon_1 = 3 \times 10^{-4}$, $\epsilon_2 = 10^{-5}$, $\eta_1 = \eta_2 = 0.01, \delta = 0.01$.

We consider a 3MS - 3RU C-RAN and set $w_3 = 3$, $w_1 = w_2 = 1$. We assume that the channel is configured as $h_{11} = 1$, $h_{12} = 0.3$, $h_{13} = 0.2$, $h_{21} = 0.2$, $h_{22} = 1$, $h_{23} = 0.3$, $h_{31} = 0.2$, $h_{32} = 0.1$, $h_{33} = 0.5$, $\sigma_n^2 = 1$. At first we use the proposed algorithms to optimize the quantizers as well as the capacity allocation, in order to maximize the sum rate (case 1). Then we fix this capacity allocation, and use the AIB method and the Alternating Bi-Section method to optimize the quantizers only, so as to maximize the weighted sum rate, under the current capacity allocation (case 2). At last, we optimize both the quantizers and the capacity allocation for maximizing the weighted sum rate (case 3). The results is shown in Fig. 2 - 5. From the figures we see that when the quantizers and capacity allocation are optimized in order to maximize the sum rate, the individual rate of the third user R_3 is the smallest, while the achievable sum rate is maximized. Moreover, R_1 and R_2 are the largest in these 3 cases. While when we put more weight on R_3 by setting $w_3 = 3$, only optimizing the quantizers is not sufficient, the improvement of R_3 in case 2 compared to the former case is not significant. This is because the received signals at different RUs are the superposition of the signals from all users, only optimizing the quantization will not impose a prominent impact on individual rates. In order to further improve the individual rate with larger weight, it is necessary to consider a simultaneous optimization of both capacity allocation and compression. From Fig. 4, we see that by comparing with case 1, the improvement of R_3 in case 3 is much more significant than that of case 2. While this improvement is at the cost of a larger decrease of R_1, R_2 and sum rate R_{sum}.

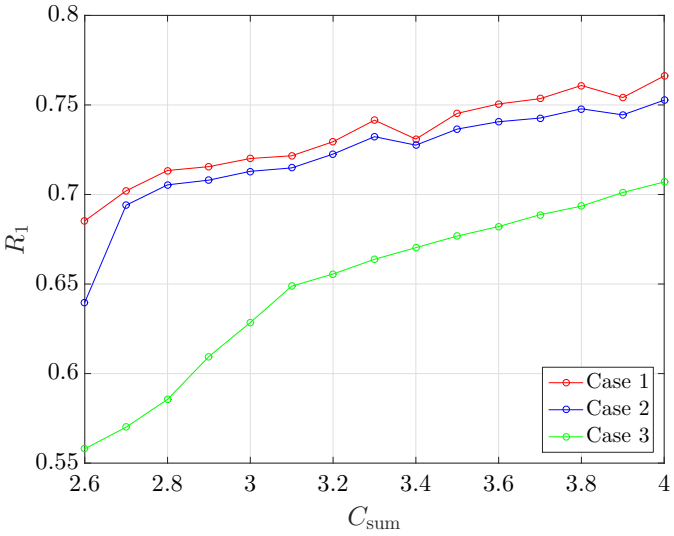

Fig. 2. R_1 with sum capacity of fronthaul.

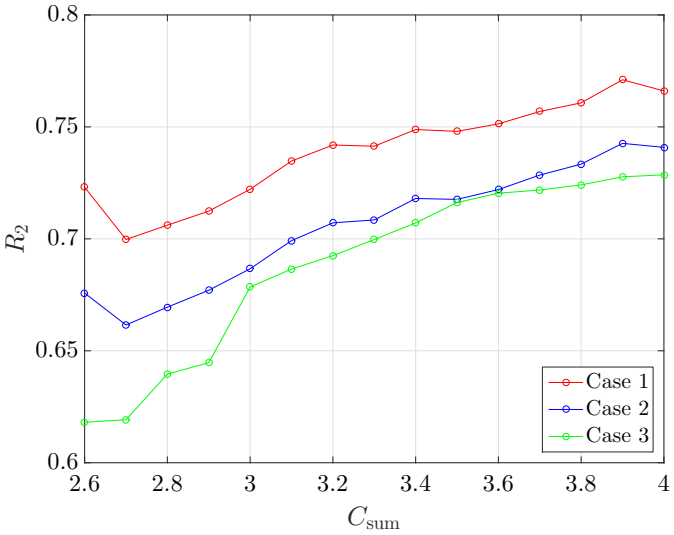

Fig. 3. R_2 with sum capacity of fronthaul.

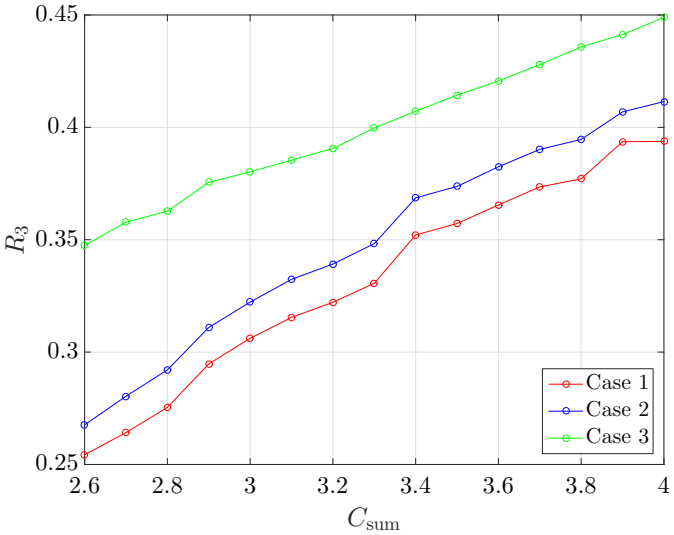

Fig. 4. R_3 with sum capacity of fronthaul.

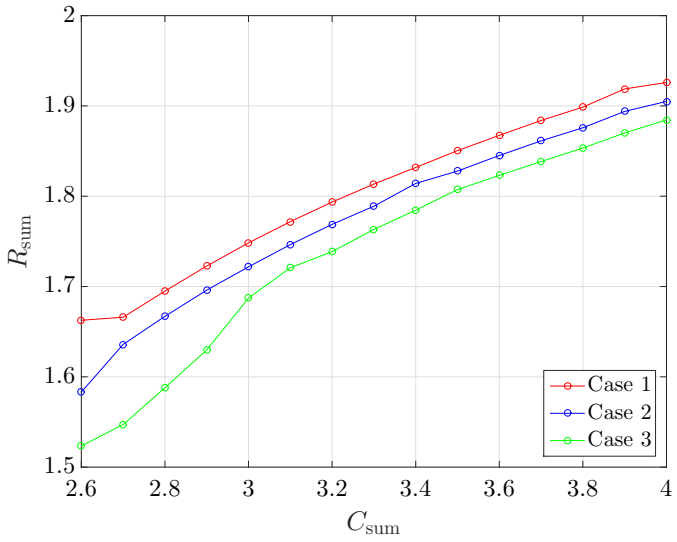

Fig. 5. Sum rate with sum capacity of fronthaul.

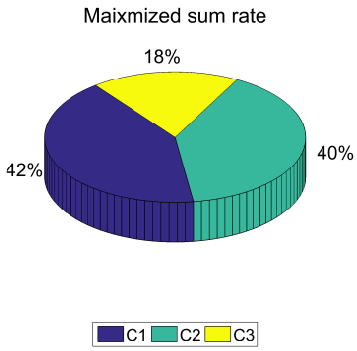

Fig. 6. Optimal capacity allocation for maximizing the sum rate.

Now we consider the same model as above, and assume the sum capacity available is 3 bits/cu, the optimal capacity allocations obtained from the proposed algorithm for different optimization objectives are shown in Fig. 6 and Fig. 7. We see that when the quantizers and the capacity allocation are optimized for maximizing the sum rate, only 18% of the capacity is allocated to the third RU. While when we want to maximize the weighted sum rate ($w_3 = 3,\ w_1 = w_2 = 1$), 38% of the capacity should be allocated to the third RU. The reason is that the signal from the third user at the third RU is the strongest, while at the first RU it is the weakest. Moreover, the observation of the superposed signal is more reliable at the first and second RU than that at the third RU. Thus, when the capacity allocation is optimized for maximizing the sum rate, the capacity allocated to the third RU should be the least, while it grows larger when the achievable rate of the third user has larger weight.

VI. Conclusion

In this paper we extend the Alternating Information Bottleneck (AIB) method and the Alternating Bi-Section method, which are proposed in our work [10], to maximize the weighted sum rate in the uplink of C-RAN, subjected to a sum fronthaul capacity. At first we suppose the capacity allocated to each fronthaul is predetermined, these two algorithms can

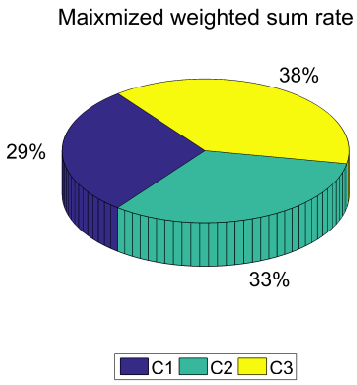

Fig. 7. Optimal capacity allocation for maximizing the weighted sum rate.

be readily used to find the optimal trade-off between the compression rates and the maximized weighted sum rate. Then based on these two algorithms, we propose the optimization for the capacity allocation. It has been shown that in order to maximize the weighted sum rate in the uplink of C-RAN, a joint optimization of all the quantizers and the resource allocation is necessary. The algorithms are suitable for the centralized optimization of the compression step in C-RAN.

Reference

[1] China Mobile. C-RAN: The road towards green ran, white paper, ver. 2.5. *China Mobile Research Institute*, Oct 2011.

[2] S. H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). Fronthaul compression for cloud radio access networks. *IEEE Signal Processing Magazine*, Nov 2014.

[3] A. D. Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inf. Theory*, 22(1), Jan 1976.

[4] Y. H. Zhou and W. Yu. Optimized backhaul compression for uplink cloud radio access network. *IEEE Journal on Selected Areas in Communications*, 32(6), Jun 2014.

[5] S. H. Lim, Y. H. Kim, A. E. Gamal, and S. Y. Chung. Noisy network coding. *IEEE Trans. Inf. Theory*, 57(5), May 2011.

[6] S. H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz). Joint decompression and decoding for cloud radio access networks. *IEEE Signal Processing Letters*, 20(5), May 2013.

[7] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *The 37th annual Allerton Conference on Communication, Control, and Computing*, Sep 1999.

[8] M. Heindlmaier, O. Iscan, and C. Rosanka. Scalar quantize-and-forward for symmetric half-duplex two-way relay channels. *IEEE International Symposium on Information Theory*, 2013.

[9] G. C. Zeitler. Low-precision quantizer design for communication problems. *http://mediatum.ub.tum.de/doc/1092069/1092069.pdf*, Nov 2011.

[10] D. Chen and V. Kuehn. Alternating information bottleneck optimization for the compression in the uplink of C-RAN. *IEEE International Conference on Communications (ICC)*, 2016.

[11] M. S. Bazaara, H. D. Sherali, and C. M. Shetty. Nonlinear programming. *Hoboken: John Wiley Sons, Inc.*, 2006.

[12] H. Boche and M. Wiczanowski. Stability optimal transmission policy for the multiple antenna multiple access channel in the geometric view. *EURASIP Signal Processing Journal*, Aug 2006.