Joint Precoding and Fronthaul Optimization for C-RANs in Ergodic Fading Channels

Jinkyu Kang*, Osvaldo Simeone†, Joonhyuk Kang* and Shlomo Shamai (Shitz)‡
*Department of Electrical Engineering, KAIST, Daejeon, South Korea
†CWCSPR, NJIT, Newark, NJ, USA
‡Department of Electrical Engineering, Technion, Haifa, Israel
Email: kangjk@kaist.ac.kr

***Abstract*—This work investigates the joint design of fronthaul compression and precoding for the downlink of Cloud Radio Access Networks (C-RANs). The main goal is that of bringing insight into an aspect of the optimal functional split between Radio Units (RUs) and Central Unit (CU), namely: where should precoding be performed? Unlike previous works, we tackle this issue for a practical scenario with block-ergodic channels and either instantaneous or stochastic Channel State Information (CSI) at the CU. Optimization algorithms over fronthaul compression and precoding are proposed that are based on a stochastic successive upper-bound minimization approach. Via numerical results, the relative merits of two strategies, in which precoding is carried out at the CU or at the RUs, are evaluated as a function of system parameters such as fronthaul capacity and channel coherence time under either instantaneous or stochastic CSI at the CU.**

I. Introduction

As industry and academia reconsider conventional cellular systems in the face of unprecedented wireless traffic growth, the Cloud-Radio Access Network (C-RAN) architecture has emerged as a promising solution due to its potential to overcome the problems of cell association and interference management [1]. In the downlink, the standard C-RAN solution prescribes all baseband processing to be performed at the central unit (CU) on behalf of all connected radio units (RUs) via low-latency fronthaul links. Accordingly, the CU compresses the precoded baseband signals and forwards them on the fronthaul links to the corresponding RUs, which upconvert and transmit the compressed baseband signals to the mobile stations (MSs). This approach, which is referred to here as a *Compression-After-Precoding* (CAP), is studied in, e.g., [2]–[4]. According to an alternative strategy known as a *Compression-Before-Precoding* (CBP) [5], [6], the CU still calculates the precoding matrices, but it does not encode and precode the data streams; rather, it forwards the data streams and the precoding matrices to the RUs, which then perform encoding and precoding.

The comparison between CAP and CBP brings insight into an aspect of the optimal functional split between RUs and CU [7], namely whether precoding should be performed at the CU, as in CAP, or at the RUs, as in CBP. Unlike previous works [2]–[5], here we tackle this issue and the corresponding design of fronthaul compression and precoding, not under the assumption of static channels and full channel state information (CSI) at the CU, but under a block-ergodic fading model [8] under both instantaneous and stochastic CSI. Optimization algorithms are proposed that leverage successive convex optimization techniques [9] with the aim of maximizing the ergodic capacity. Via numerical results, we illustrate the relative merits of CAP and CBP as a function of system parameters such as fronthaul capacity and channel coherence time under both instantaneous and stochastic CSI.

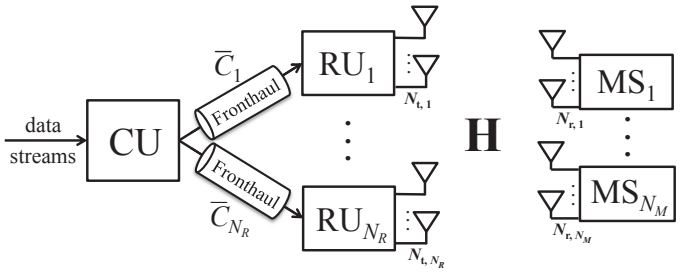

Fig. 1. Downlink of a C-RAN system in which a single cluster of RUs is connected to a CU via finite-capacity fronthaul links. The downlink channel matrix $\mathbf{H}$ varies in an ergodic fashion along the channel coherence blocks.

The rest of the paper is organized as follows. We describe the system model in Section II. In Section III, we study the CAP strategy, while the CBP approach is studied in IV. In Section III and Section IV, we concentrate on the stochastic CSI case and refer to [10] for a more thorough coverage that includes also the instantaneous CSI case. In Section V, numerical results are presented. Concluding remarks are summarized in Section VI.

II. System Model

We consider the downlink of a C-RAN in which a cluster of N_R RUs provides wireless service to N_M MSs as illustrated in Fig. 1. Part of the baseband processing for all the RUs in the cluster is carried out at a CU that is connected to each i-th RU via a fronthaul link of finite capacity, as further discussed below. Each i-th RU has $N_{t,i}$ transmit antennas and each j-th MS has $N_{r,j}$ receive antennas. We denote the set of all RUs as $\mathcal{N}_R = \{1, \ldots, N_R\}$ and of all MSs as $\mathcal{N}_M = \{1, \ldots, N_M\}$. We define the number of total transmit antennas as $N_t = \sum_{i=1}^{N_R} N_{t,i}$ and of total receive antennas as $N_r = \sum_{j=1}^{N_M} N_{r,j}$.

Each coded transmission block spans multiple coherence periods, e.g., multiple distinct resource blocks in an LTE

system, of the downlink channel. Specifically, we adopt a block-ergodic channel model, in which the fading channels are constant within a coherence period but vary in an ergodic fashion across a large number of coherence periods. Within each channel coherence period of duration T channel uses, the baseband signal transmitted by the i-th RU is given by a $N_{t,i} \times T$ complex matrix $\mathbf{X}_i$, where each column corresponds to the signal transmitted from the $N_{t,i}$ antennas in a channel use.

The $N_{r,j} \times T$ signal $\mathbf{Y}_j$ received by the j-th MS in a given channel coherence period, where each column corresponds to the signal received by the $N_{r,j}$ antennas in a channel use, is given by

$$\mathbf{Y}_j = \mathbf{H}_j \mathbf{X} + \mathbf{Z}_j, \tag{1}$$

where $\mathbf{Z}_j$ is the $N_{r,j} \times T$ noise matrix, which consist of i.i.d. $\mathcal{CN}(0,1)$ entries; $\mathbf{H}_j = [\mathbf{H}_{j1}, \ldots, \mathbf{H}_{j_{N_R}}]$ denotes the $N_{r,j} \times N_t$ channel matrix for j-th MS, where $\mathbf{H}_{ji}$ is the $N_{r,j} \times N_{t,i}$ channel matrix from the i-th RU to the j-th MS; and $\mathbf{X}$ is the collection of the signals transmitted by all the RUs, i.e., $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_{N_B}^T]^T$. As per the discussion above, the channel matrix $\mathbf{H}_j$ is assumed to be constant during each channel coherence block and to change according to a stationary ergodic process from block to block. We consider the scenario in which it is only aware of the distribution of the channel matrix $\mathbf{H}$, i.e., to have *stochastic CSI*. Instead, the MSs always have full CSI about their respective channel matrices, as we will state more precisely in the next sections. The transmit signal $\mathbf{X}_i$ has a power constraint given as $E[||\mathbf{X}_i||^2]/T \leq \bar{P}_i$.

A specific channel model of interest is the standard model with transmit-only correlation [11], where the channel matrix $\mathbf{H}_{ji}$ is written as $\mathbf{H}_{ji} = \widetilde{\mathbf{H}}_{ji} \mathbf{\Sigma}_{T,ji}^{1/2}$, where the $N_{t,i} \times N_{t,i}$ matrix $\mathbf{\Sigma}_{T,ji}$ accounts for transmit-side correlation matrices and the $N_{r,j} \times N_{t,i}$ random matrix $\widetilde{\mathbf{H}}_{ji}$ has i.i.d. $\mathcal{CN}(0,1)$ variables and accounts for the small-scale multipath fading [11]. With this model, stochastic CSI entails that the CU is only aware of the correlation matrix $\mathbf{\Sigma}_{T,ji}$. The correlation matrix depends on the geometry of the propagation environment and may be calculated as in [11, eq. (3)].

Each i-th fronthaul link has capacity $\bar{C}_i$, which is measured in bit/s/Hz, where the normalization is with respect to the bandwidth of the downlink channel. The fronthaul constraint will be further discussed in Section III and Section IV.

III. Compress-After-Precoding

In this section, we first describe the CAP strategy in Section III-A. Then, we propose an algorithm for the joint optimization of fronthaul compression and precoding under the assumption of stochastic CSI at the CU in Section III-B. The case of instantaneous CSI can be found in [10].

A. Precoding and Fronthaul Compression for CAP

With the CAP scheme as illustrated in Fig. 2, the CU performs channel coding and precoding, and then compresses the resulting baseband signals so that they can be forwarded

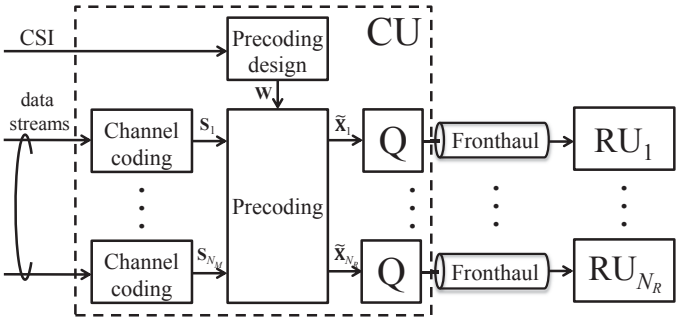

Fig. 2. Block diagram of the Compression-After-Precoding (CAP) scheme ("Q" represents fronthaul compression).

on the fronthaul links to the corresponding RUs. This strategy corresponds to the standard approach envisioned for C-RANs [2]–[4], [6]. Specifically, channel coding is performed separately for the information stream intended for each MS. This step produces the data signal $\mathbf{S} = [\mathbf{S}_1^\dagger, \ldots, \mathbf{S}_{N_M}^\dagger]^\dagger$ for each coherence block, where $\mathbf{S}_j$ is the $M_j \times T$ matrix containing, as rows, the $M_j \leq N_{r,j}$ encoded data streams for the j-th MS. We define the number of total data streams as $M = \sum_{j=1}^{N_M} M_j$ and assume the condition $M \leq N_t$. Following standard random coding arguments, we take all the entries of matrix $\mathbf{S}$ to be i.i.d. as $\mathcal{CN}(0,1)$. The encoded data $\mathbf{S}$ is further processed to obtain the transmitted signals $\mathbf{X}$ as detailed below.

Precoding: The precoded data signal computed by the CU for any given coherence time can be written as $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{S}$, where $\mathbf{W}$ is the $N_t \times M$ precoding matrix. Note that with instantaneous CSI a different precoding matrix $\mathbf{W}$ is used for different coherence times in the coding block, while, with stochastic CSI, the same precoding matrix $\mathbf{W}$ is used for all coherence times. In both cases, the precoded data signal $\widetilde{\mathbf{X}}$ can be divided into the $N_{t,i} \times T$ signals $\widetilde{\mathbf{X}}_i$ corresponding to i-th RU for all $i \in \mathcal{N}_R$ as $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_1^\dagger, \ldots, \widetilde{\mathbf{X}}_{N_R}^\dagger]^\dagger$. Specifically, the baseband signal $\widetilde{\mathbf{X}}_i$ for i-th RU is defined as $\widetilde{\mathbf{X}}_i = \mathbf{W}_i^r \mathbf{S}$, where $\mathbf{W}_i^r$ is the $N_{t,i} \times N_r$ precoding matrix for the i-th RU, which is obtained by properly selecting the rows of matrix $\mathbf{W}$ (as indicated by the superscript "r" for "rows"): the matrix $\mathbf{W}_i^r$ is given as $\mathbf{W}_i^r = \mathbf{D}_i^{rT} \mathbf{W}$, with the $N_t \times N_{t,i}$ matrix $\mathbf{D}_i^r$ having all zero elements except for the rows from $\sum_{k=1}^{i-1} N_{t,k} + 1$ to $\sum_{k=1}^{i} N_{t,k}$, that contain an $N_{t,i} \times N_{t,i}$ identity matrix.

Quantization: The CU quantizes each sequence of baseband signal $\widetilde{\mathbf{X}}_i$ for transmission on the i-th fronthaul link to the i-th RU. Leveraging random coding for quantization (see, e.g., [12]), we write the compressed signals $\mathbf{X}_i$ for i-th RU as

$$\mathbf{X}_i = \widetilde{\mathbf{X}}_i + \mathbf{Q}_{x,i}, \tag{2}$$

where the quantization noise matrix $\mathbf{Q}_{x,i}$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma_{x,i}^2)$ entries. The quantization noises $\mathbf{Q}_{x,i}$ are independent across the RU index i, as we assume separate quantizers for the signals of different RUs. Based on (2), the design of the fronthaul compression reduces to the optimization of the quantization noise variances $\sigma_{x,1}^2, \ldots, \sigma_{x,N_B}^2$. The

power transmitted by i-th RU is then computed as

$$P_i\left(\mathbf{W},\sigma_{x,i}^2\right)=\frac{1}{T}E[||\mathbf{X}_i||^2]=\text{tr}\left(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r+\sigma_{x,i}^2\mathbf{I}\right), \tag{3}$$

where we have emphasized the dependence of the power $P_i(\mathbf{W},\sigma_{x,i}^2)$ on the precoding matrix $\mathbf{W}$ and quantization noise variances $\sigma_{x,i}^2$. Moreover, using standard rate-distortion arguments, the rate required on the fronthaul between the CU and i-th RU in a given coherence interval can be quantified by $I(\widetilde{\mathbf{X}}_i;\mathbf{X}_i)/T$ (see, e.g., [12, Ch. 3]). Therefore, the rate allocated on the i-th fronthaul link is equal to

$$C_i\big(\mathbf{W},\sigma_{x,i}^2\big)=\log\det\big(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r+\sigma_{x,i}^2\mathbf{I}\big)-N_{t,i}\log\left(\sigma_{x,i}^2\right), \tag{4}$$

so that the fronthaul capacity constraint is $C_i(\mathbf{W},\sigma_{x,i}^2)\leq\bar{C}_i$.

Ergodic Achievable Rate: We assume that each j-th MS is aware of the effective receive channel matrices $\widetilde{\mathbf{H}}_{jk}=\mathbf{H}_j\mathbf{W}_k^c$ for all $k\in\mathcal{N}_M$ at all coherence times, where $\mathbf{W}_k^c$ is the $N_t\times N_{r,j}$ precoding matrix corresponding to k-th MS, which is obtained from the precoding matrix $\mathbf{W}$ by properly selecting the columns as $\mathbf{W}=[\mathbf{W}_1^c,\ldots,\mathbf{W}_{N_M}^c]$. We collect the effective channels in the matrix $\widetilde{\mathbf{H}}_j=[\widetilde{\mathbf{H}}_{j1},\ldots,\widetilde{\mathbf{H}}_{jN_M}]=\mathbf{H}_j\mathbf{W}$. The effective channel $\widetilde{\mathbf{H}}_j$ can be estimated at the MSs via downlink training. Under this assumption, the ergodic achievable rate for the j-th MS is computed as $E[R_j^{CAP}(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2)]$, with $R_j^{CAP}(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2)=I_{\mathbf{H}}(\mathbf{S}_j;\mathbf{Y}_j)/T$, where $I_{\mathbf{H}}(\widetilde{\mathbf{S}}_j;\mathbf{Y}_j)$ represents the mutual information conditioned on the value of channel matrix $\mathbf{H}$, the expectation is taken with respect to $\mathbf{H}$ and

$$\begin{aligned}R_j^{CAP}\big(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2\big)=&\log\det\left(\mathbf{I}+\mathbf{H}_j\left(\mathbf{W}\mathbf{W}^\dagger+\boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right) \\ &-\log\det\left(\mathbf{I}+\mathbf{H}_j\left(\sum_{k\in\mathcal{N}_M\setminus j}\mathbf{W}_k^c\mathbf{W}_k^{c\,\dagger}+\boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right),\end{aligned} \tag{5}$$

with the covariance matrix $\boldsymbol{\Omega}_x$ being a diagonal with diagonal blocks given as $\text{diag}([\sigma_{x,1}^2\mathbf{I},\ldots,\sigma_{x,N_B}^2\mathbf{I}])$ and $\boldsymbol{\sigma}_x^2=[\sigma_{x,1}^2,\ldots,\sigma_{x,N_B}^2]^T$.

The ergodic achievable weighted sum-rate can be optimized over the precoding matrix $\mathbf{W}$ and the compression noise variances $\boldsymbol{\sigma}_x^2$ under fronthaul capacity and power constraints. In the next subsections, we consider the optimization problem with stochastic CSI.

B. Optimization Algorithm for Stochastic CSI

With only stochastic CSI at the CU, the same precoding matrix $\mathbf{W}$ and compression noise variances $\boldsymbol{\sigma}_x^2$ are used for all the coherence blocks. Accordingly, the problem of optimizing the ergodic weighted achievable sum-rate can be reformulated as follows

$$\begin{aligned}\underset{\mathbf{W},\boldsymbol{\sigma}_x^2}{\text{maximize}}\quad & \sum_{j\in\mathcal{N}_M}\mu_j E\left[R_j^{CAP}\left(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2\right)\right] && \text{(6a)}\\ \text{s.t.}\quad & C_i\left(\mathbf{W},\sigma_{x,i}^2\right)\leq\bar{C}_i, && \text{(6b)}\\ & P_i\left(\mathbf{W},\sigma_{x,i}^2\right)\leq\bar{P}_i, && \text{(6c)}\end{aligned}$$

where (6b)-(6c) apply to all $i\in\mathcal{N}_R$. In order to tackle this problem, we adopt the Stochastic Successive Upper-bound Minimization (SSUM) method [9], whereby, at each step, a stochastic lower bound of the objective function is maximized around the current iterate. To this end, similar to [4], we recast the optimization over the covariance matrices $\mathbf{V}_j=\mathbf{W}_j^c\mathbf{W}_j^{c\dagger}$ for all $j\in\mathcal{N}_M$, instead of the precoding matrices $\mathbf{W}_j^c$ for all $j\in\mathcal{N}_M$. We observe that, with this choice, the objective function is expressed as the average of DC functions, while the constraint (6b) is also a DC function, with respect to the covariance $\mathbf{V}=[\mathbf{V}_1\ldots\mathbf{V}_{N_M}]$ and the quantization noise variances $\boldsymbol{\sigma}_x^2$. As discussed above, the resulting problem is a rank-relaxation of the original problem (6).

The proposed algorithm to tackle the resulting problem is based on SSUM [9] and contains two nested loops. At each outer iteration n, a new channel matrix realization $\mathbf{H}^{(n)}=[\mathbf{H}_1^{T\,(n)},\ldots,\mathbf{H}_{N_M}^{T\,(n)}]$ is drawn based on the availability of stochastic CSI at the CU. For example, with the transmit-only correlation channel model, the channel matrices are generated based on the knowledge of the spatial correlation matrix. Following the SSUM scheme, the outer loop aims at maximizing a stochastic lower bound on the objective function, given as

$$\frac{1}{n}\sum_{l=1}^{n}\widetilde{R}_j^{CAP}\left(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)}\right), \tag{7}$$

where $\widetilde{R}_j^{CAP}(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)})$ is a locally tight convex lower bound on $R_j^{CAP}\left(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2\right)$ around solution $\mathbf{V}^{(l-1)}$, $\boldsymbol{\sigma}_x^{2\,(l-1)}$ obtained at the $(l-1)$ the outer iteration when the channel realization is $\mathbf{H}^{(l)}$. This can be calculated as (see [9])

$$\begin{aligned}&\widetilde{R}_j^{CAP}\left(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)}\right)\triangleq \\ &\log\det\left(\mathbf{I}+\mathbf{H}_j^{(l)}\left(\sum_{k=1}^{N_M}\mathbf{V}_k+\boldsymbol{\Omega}_x\right)\mathbf{H}_j^{(l)\,\dagger}\right) \\ &\qquad -f\left(\mathbf{I}+\mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j^{(l-1)}\mathbf{H}_j^{(l)\,\dagger},\mathbf{I}+\mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j\mathbf{H}_j^{(l)\,\dagger}\right),\end{aligned} \tag{8}$$

where $\boldsymbol{\Lambda}_j=\sum_{k=1,k\neq j}^{N_M}\mathbf{V}_k+\boldsymbol{\Omega}_x$, $\boldsymbol{\Lambda}_j^{(l-1)}=\sum_{k=1,k\neq j}^{N_M}\mathbf{V}_k^{(l-1)}+\boldsymbol{\Omega}_x$, and the covariance matrix $\boldsymbol{\Omega}_x^{(l)}$ is a diagonal matrix with diagonal blocks given as $\text{diag}([\sigma_{x,1}^{2\,(l)}\mathbf{I},\ldots,\sigma_{x,N_B}^{2\,(l)}\mathbf{I}])$ and the linearized function $f(\mathbf{A},\mathbf{B})$ is obtained from the first-order Taylor expansion of the log det function as

$$f(\mathbf{A},\mathbf{B})\triangleq\log\det\left(\mathbf{A}\right)+\frac{1}{\ln 2}\text{tr}\left(\mathbf{A}^{-1}\left(\mathbf{B}-\mathbf{A}\right)\right). \tag{9}$$

Since the maximization of (7) is subject to the non-convex DC constraint (6b), the inner loop tackles the problem via the MM algorithm i.e., by applying successive locally tight convex lower bounds to the left-hand side of the constraint (6b). Specifically, given the solution $\mathbf{V}^{(n,r-1)}$ and $\boldsymbol{\sigma}_x^{2\,(n,r-1)}$ at $(r-1)$-th inner iteration of the n-th outer iteration, the fronthaul constraint in (6b) at the r-th inner iteration can be

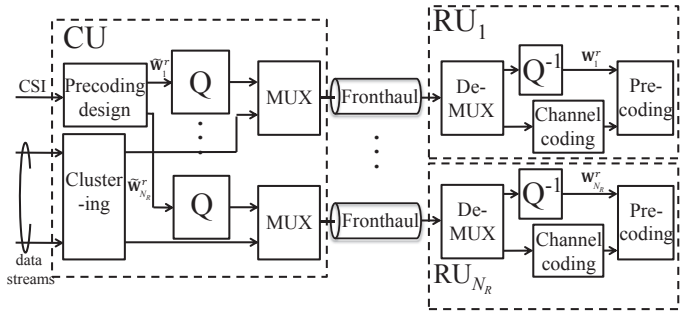

Fig. 3. Block diagram of the Compression-Before-Precoding (CBP) scheme ("Q" and "Q^{-1}" represents fronthaul compression and decompression, respectively).

locally approximated as

$$\widetilde{C}_i\left(\mathbf{V}, \sigma^2_{x,i}|\mathbf{V}^{(n,r-1)}, \sigma^{2\ (n,r-1)}_{x,i}\right) \triangleq \tag{10}$$
$$f\left(\sum_{k=1}^{N_M}\mathbf{D}_i^{rT}\mathbf{V}_k^{(n,r-1)}\mathbf{D}_i^r+\sigma^{2\ (n,r-1)}_{x,i}\mathbf{I}, \sum_{k=1}^{N_M}\mathbf{D}_i^{rT}\mathbf{V}_k\mathbf{D}_i^r+\sigma^2_{x,i}\mathbf{I}\right)$$
$$-N_{t,i}\log\left(\sigma^2_{x,i}\right).$$

The resulting combination of SSUM and MM algorithms for the solution of problem (6) is summarized in [10, Algorithm 1].

A few remarks are in place on the properties of the proposed algorithm. First, since the approximated functions (8) and (10) are local lower bounds, the algorithm provides a feasible solution of the relaxed problem at each inner and outer iteration (see, e.g., [9]). The second remark is that, from [9], [13], as long as a sufficient number of inner iterations is performed at each outer iteration, the algorithm is guaranteed to converge to stationary points of the relaxed problem. Third, from the obtained solution of the relaxed problem, a rank-constrained feasible solution of the original problem can be obtained via eigenvalue decomposition [10].

IV. Compression-Before-Precoding

With the Compression-Before-Precoding (CBP) scheme, the CU calculates the precoding matrices, but does not perform precoding. Instead, as illustrated in Fig. 3, it uses the fronthaul links to communicate the information messages of a given subset of MSs to each RU, along with the corresponding compressed precoding matrices. Each RU can then encode and precode the messages of the given MSs based on the information received from the fronthaul link. As it will be discussed, in the CBP scheme, unlike CAP, a preliminary clustering step is generally advantageous whereby each MS is assigned to a subset of RUs. In the following, we first describe the CBP strategy in Section IV-A and introduce an algorithm for the joint optimization of fronthaul compression and precoding with stochastic CSI at the CU. The instantaneous CSI case can be found in [10].

A. Precoding and Fronthaul Compression for CBP

As shown in Fig. 3, in the CBP method, the precoding matrix $\widetilde{\mathbf{W}}$ and the information streams are separately transmitted from the CU to the RUs, and the received information bits are encoded and precoded at each RU using the received precoding matrix. Note that, with this scheme, the transmission overhead over the fronthaul depends on the number of MSs supported by a RU, since the RUs should receive all the corresponding information streams.

Clustering: Given the above, with the CBP strategy, we allow for a preliminary clustering step at the CU whereby each RU is assigned by a subset of the MSs. We denote the set of MSs assigned by i-th RU as $\mathcal{M}_i \subseteq \mathcal{N}_M$ for all $i \in \mathcal{N}_B$. This implies that i-th RU only needs the information streams intended for the MSs in the set $\mathcal{M}_i$. We also denote the set of RUs that serve the j-th MS, as $\mathcal{B}_j = \{i|j \in \mathcal{M}_i\} \subseteq \mathcal{N}_B$ for all $j \in \mathcal{N}_M$. We use the notation $\mathcal{M}_i[k]$ and $\mathcal{B}_j[m]$ to respectively denote the k-th MS and m-th RU in the sets $\mathcal{M}_i$ and $\mathcal{B}_j$, respectively. We define the number of all transmit antennas for the RUs, which serve the j-th MS, as $N_{t,\mathcal{B}_j}$. We assume here that the sets of MSs assigned by i-th RU are given and not subject to optimization (see Section V for further details).

Precoding: The precoding matrix $\widetilde{\mathbf{W}}$ is constrained to have zeros in the positions that correspond to RU-MS pairs such that the MS is not served by the given RU. This constraint can be represented as

$$\widetilde{\mathbf{W}} = \left[\mathbf{E}_1^c\widetilde{\mathbf{W}}_1^c, \ldots, \mathbf{E}_{N_M}^c\widetilde{\mathbf{W}}_{N_M}^c\right], \tag{11}$$

where $\widetilde{\mathbf{W}}_j^c$ is the $N_{t,\mathcal{B}_j} \times N_{r,j}$ precoding matrix intended for j-th MS and RUs in the cluster $\mathcal{B}_j$, and the $N_t \times N_{t,\mathcal{B}_j}$ constant matrix $\mathbf{E}_j^c$ ($\mathbf{E}_j^c$ only has either a 0 or 1 entries) defines the association between the RUs and the MSs as $\mathbf{E}_j^c = \left[\mathbf{D}^c_{\mathcal{B}_j[1]}, \ldots, \mathbf{D}^c_{\mathcal{B}_j[|\mathcal{B}_j|]}\right]$, with the $N_r \times N_{r,j}$ matrix $\mathbf{D}_j^c$ having all zero elements except for the rows from $\sum_{k=1}^{j-1} N_{r,k}+1$ to $\sum_{k=1}^{j} N_{r,j}$, which contain an $N_{r,j} \times N_{r,j}$ identity matrix.

Quantization: The sequence of the $N_{t,i} \times N_{r,\mathcal{M}_i}$ precoding matrices $\widetilde{\mathbf{W}}_i^r$ intended for each i-th RU for all coherence times in the coding block is compressed by the CU and forwarded over the fronthaul link to the i-th RU. The compressed precoding matrix $\mathbf{W}_i^r$ for i-th RU is given by

$$\mathbf{W}_i^r = \widetilde{\mathbf{W}}_i^r + \mathbf{Q}_{w,i}, \tag{12}$$

where the $N_{t,i} \times N_{r,\mathcal{M}_i}$ quantization noise matrix $\mathbf{Q}_{w,i}$ is assumed to have zero-mean i.i.d. $\mathcal{CN}(0, \sigma^2_{w,i})$ entries and to be independent across the index i. Overall, the $N_t \times N_r$ compressed precoding matrix $\mathbf{W}$ for all RUs is represented as

$$\mathbf{W} = \widetilde{\mathbf{W}} + \mathbf{Q}_w, \tag{13}$$

where $\mathbf{W} = [\mathbf{E}_1^{r\dagger}\mathbf{W}_{w,1}^{\dagger}, \ldots, \mathbf{E}_{N_B}^{r\dagger}\mathbf{W}_{w,N_B}^{\dagger}]^{\dagger}$, $\widetilde{\mathbf{W}}$ and $\mathbf{Q}_w$ are similarly defined. Note that we have $E[\mathrm{vec}(\mathbf{Q}_w)\ \mathrm{vec}(\mathbf{Q}_w)^{\dagger}] = \mathbf{\Omega}_w$, where $\mathbf{\Omega}_w$ is a diagonal matrix with diagonal blocks given by $[\sigma^2_{w,1}\mathbf{I}, \ldots, \sigma^2_{w,N_B}\mathbf{I}]$.

Ergodic Achievable Rate: The ergodic rate achievable for j-th MS can be written as $E[R_j^{CBP}(\mathbf{H}, \widetilde{\mathbf{W}}, \boldsymbol{\sigma}_w^2)]$, where

$$R_j^{CBP}\left(\mathbf{H}, \widetilde{\mathbf{W}}, \boldsymbol{\sigma}_w^2\right) = \frac{1}{T} I_{\mathbf{H}}\left(\mathbf{S}_j; \mathbf{Y}_j\right) = \tag{14}$$
$$\log\det\left(\mathbf{I} + \mathbf{H}_j\left(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\dagger} + \boldsymbol{\Omega}_w\right)\mathbf{H}_j^{\dagger}\right)$$
$$- \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\sum_{k\in\mathcal{N}_M\setminus j} \widetilde{\mathbf{W}}_k^c \widetilde{\mathbf{W}}_k^{c\dagger} + \boldsymbol{\Omega}_w\right)\mathbf{H}_j^{\dagger}\right).$$

B. Optimization Algorithm for Stochastic CSI

With stochastic CSI at the CU, the same precoding matrix is used for all the coherence blocks and hence the rate required to convey the precoding matrix $\widetilde{\mathbf{W}}_i^r$ to each i-th RU becomes negligible. As a result, we can set $\sigma_{w,i}^2 = 0$ for all $i \in \mathcal{N}_R$. Accordingly, the fronthaul capacity can be only used for transfer of the information stream as $\sum_{j\in\mathcal{M}_i} R_j \leq C_i$, for all $i \in \mathcal{N}_R$. Based on the above considerations, the optimization problem of interest is formulated as

$$\underset{\widetilde{\mathbf{W}},\{R_j\}}{\text{maximize}} \quad \sum_{j\in\mathcal{N}_M} \mu_j R_j \tag{15a}$$
$$s.t. \quad R_j \leq E\left[R_j^{CBP}\left(\mathbf{H}, \widetilde{\mathbf{W}}, \mathbf{0}\right)\right], \tag{15b}$$
$$\sum_{j\in\mathcal{M}_i} R_j \leq \bar{C}_i, \tag{15c}$$
$$P_i\left(\widetilde{\mathbf{W}}_i^r, 0\right) \leq \bar{P}_i, \tag{15d}$$

where (15b) applies to all $j \in \mathcal{N}_M$, (15c)-(15d) apply to all $i \in \mathcal{N}_R$ and the transmit power $P_i(\widetilde{\mathbf{W}}_i^r, \sigma_{w,i}^2)$ at i-th RU is defined in (3). In problem (15), the constraint (15b) is not only non-convex but also stochastic. Similar to Section III-B, the functions $R_j^{CBP}(\mathbf{H}, \widetilde{\mathbf{W}})$ can be seen to be DC functions of the covariance matrices $\widetilde{\mathbf{V}}_j = \widetilde{\mathbf{W}}_j^c \widetilde{\mathbf{W}}_j^{c\dagger}$ for all $j \in \mathcal{N}_M$, hence opening up the possibility to develop a solution based on SSUM. Similar to Section III-B, we refer to [10] for details.

V. Numerical Results

In this section, we compare the performance of the CAP and CBP schemes in the set-up under study of block-ergodic channels. Note that, for the case of perfect CSI, a different precoder is designed for each coherence block by using a similar MM algorithm as detailed in [10]. We consider a system in which the RUs and the MSs are randomly located in a square area with side $\delta = 500m$. In [11, eq. (3)], we set the reference distance to $d_0 = 50m$, the path loss exponent to $\eta = 3$, angular spread $\Delta_{ji} = \arctan(r_s/d_{ji})$, and scattering radius $r_s = 10m$ with d_{ji} being the Euclidean distance between the i-th RU and the j-th MS. Throughout, we assume that the every RU is subject to the same power constraint $\bar{P}$ and has the same fronthaul capacity $\bar{C}$, that is $\bar{P}_i = \bar{P}$ and $\bar{C}_i = \bar{C}$ for $i \in \mathcal{N}_R$. Moreover, in the CBP scheme, the MS-to-RU assignment is carried out by choosing, for each RU, the N_c MSs that have the largest instantaneous channel norms for instantaneous CSI and the largest average channel matrix norms for stochastic CSI. Note that this assignment is done for each coherence block in the former case, while in the latter the same assignment holds for all coherence blocks.

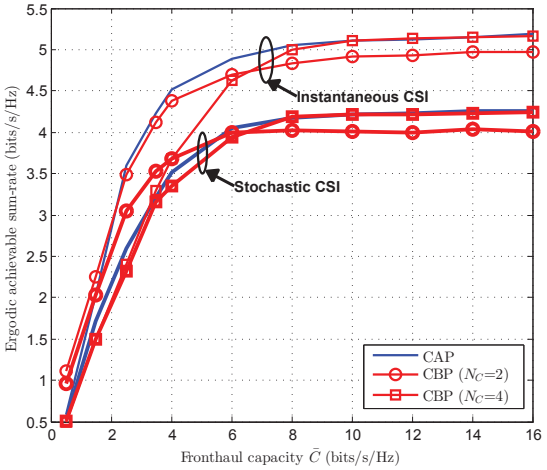

Fig. 4. Ergodic achievable sum-rate vs. the fronthaul capacity $\bar{C}$ ($N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{P} = 10$ dB, $T = 20$, and $\mu = 1$).

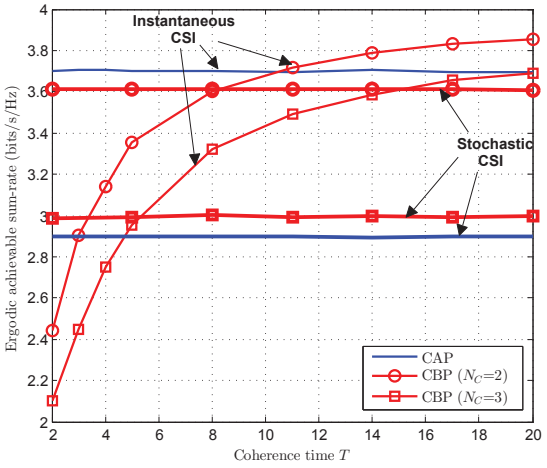

Fig. 5. Ergodic achievable sum-rate vs. the coherence time T ($N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 2$ bits/s/Hz, $\bar{P} = 20dB$, and $\mu = 1$).

The effect of the fronthaul capacity limitation on the ergodic achievable sum-rate is investigated in Fig. 4, where the number of RUs and MSs is $N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{P} = 10dB$, and $T = 20$. The CAP strategy is seen to perform CBP as long as the fronthaul capacity is sufficiently large, due to its capability to coordinate all the RUs via joint baseband processing without requiring the transmission of all messages on all fronthaul links. Moreover, for small $\bar{C}$, the CBP scheme, with progressively smaller N_c, has better performance thanks to the reduced fronthaul overhead, while, for large $\bar{C}$, CBP with $N_c = N_M$ approaches the performance of the CAP scheme.

Fig. 5 shows the ergodic achievable sum-rate as function of the coherence time T, with $N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 2$ bits/s/Hz, and $\bar{P} = 20$ dB. With instantaneous CSI, CBP is seen to benefit from a larger coherence time T, since the fronthaul overhead required to transmit precoding information gets amortized over a larger period. This is in contrast to CAP for which such overhead scales proportionally to the coherence time T and hence the CAP

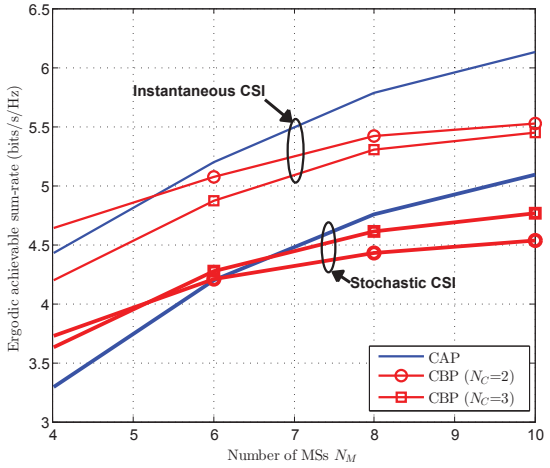

Fig. 6. Ergodic achievable sum-rate vs. the number of MSs N_M ($N_R = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 4$ bits/s/Hz, $\bar{P} = 10$ dB, $T = 10$, and $\mu = 1$).

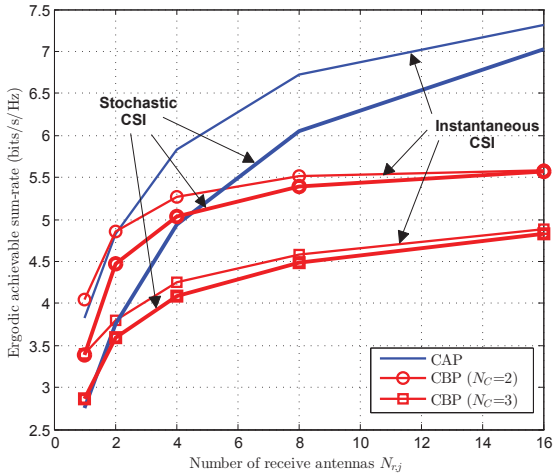

Fig. 7. Ergodic achievable sum-rate vs. the number of receive antennas N_r ($N_R = N_M = 4$, $N_{t,i} = 2$, $\bar{C} = 3$ bits/s/Hz, $\bar{P} = 10$ dB, $T = 10$, and $\mu = 1$).

scheme is not affected by the coherence time. As a result, CBP can outperform CAP for sufficiently large T in the presence of instantaneous CSI. Instead, with stochastic CSI, the effect is even more pronounced due to the additional advantage that is accrued by amortizing the precoding overhead over the entire coding block.

In Fig. 6, the ergodic achievable sum-rate is plotted versus the number of MSs N_M for $N_R = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 4$, $\bar{P} = 10dB$ and $T = 10$. It is observed that the enhanced interference mitigation capabilities of CAP without the overhead associated to the transmission of all messages on the fronthaul links yield performance gains for denser C-RANs, i.e., for larger values of N_M. This remains true for both instantaneous and stochastic CSI cases.

Finally, in Fig. 7, the ergodic achievable sum-rate is plotted versus the number of each receive antennas $N_{r,j}$ for $N_R = N_M = 4$, $N_{t,i} = 2$, $\bar{C} = 3$ bits/s/Hz, $\bar{P} = 10$ dB and $T = 10$. Although the achievable rate of each MS is increased by using a large number of MS antennas, the achievable sum-rate with the CBP approach is restricted due to the limited number of cooperative RUs as dictated by the fronthaul capacity requirements for the transmission of the data streams. Hence, it is shown that the CAP approach provides significant advantages in the presence of a large number of antennas at MS for both instantaneous and stochastic CSI. Moreover, we observe that the performance advantages of having instantaneous CSI as compared to stochastic CSI decrease in the regime of the large number of MS antenna. This is because, in this regime, serving only one MS entails only a minor loss in capacity, hence not requiring sophisticated precoding operations.

VI. Conclusion

In this paper, we have investigated the joint design of fronthaul compression and precoding for the downlink of C-RANs in the practically relevant scenario of block-ergodic fading with stochastic CSI. The study compares Compress-After-Precoding (CAP), in which precoding is carried out at the CU, and Compress-Before-Precoding (CBP), in which precoding takes place at the RUs. From numerical results, we have observed that the relative merits of the two techniques depend on the interplay between the enhanced interference management abilities of CAP, particularly for dense networks, and the lower fronthaul requirements of CBP in terms of precoding information overhead, especially for small number of MSs, large coherence periods and with stochastic, rather than instantaneous, CSI.

References

[1] China Mobile, "C-RAN: the road towards green RAN," White Paper, ver. 2.5, China mobile Research Institute, Oct. 2011.

[2] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP Jour. Adv. Sig. Proc.*, 2009.

[3] P. Marsch and G. Fettweis, "On downlink network MIMO under a constrained backhaul and imperfect channel knowledge," *Proc. IEEE Glob. Comm. Conf.*, Nov. 2009.

[4] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks," *IEEE Sig. Proc. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[5] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoded massive MIMO in cloud radio access networks," *Proc. IEEE Int. Conf. on Comm.*, Jun. 2013.

[6] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," *Proc. of IEEE Info. Th. and Application Workshop*, San Diego, CA, USA, Feb. 2014.

[7] P. Rost, C. j. Bernardos, A. D. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Comm. Mag.*, vol. 52, no. 5, pp. 68 – 76, May 2014.

[8] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Info. Th.*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[9] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *arXiv:1307.4457*.

[10] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channel," *arXiv:1412.7713*.

[11] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Trans. Info. Th.*, vol. 59, no. 10, pp. 6441–6463, Oct. 2014.

[12] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[13] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.