1
Cloud Radio Access Networks: Optimal Design of Uplink Channel Estimation and Downlink Precoding

Osvaldo Simeone, Jinkyu Kang, Joonhyuk Kang and Shlomo Shamai (Shitz)

1.1
Abstract

The centralization gains afforded by cloud radio access network (C-RAN) in terms of capital/ operating expense savings, flexibility, interference management and network densification rely on the presence of high-capacity low-latency fronthaul connectivity between remote radio heads (RRHs) and the baseband unit (BBU). In light of the non-uniform and limited availability of fiber optics cables, the bandwidth constraints on the fronthaul network call, on the one hand, for the development of advanced baseband compression strategies and, on the other, for a closer investigation of the optimal functional split between RRHs and BBU. In this chapter, this optimal function split is studied at the PHY layer in terms of two key baseband signal processing steps, namely channel estimation in the uplink and channel encoding/ linear precoding in the downlink. Joint optimization of baseband fronthaul compression and of baseband signal processing is tackled under different PHY functional splits, whereby channel estimation and precoding are carried out either at the RRHs or at the BBU. The analysis and numerical results yield insight on the regimes in terms of network architecture and fronthaul capacities in which different functional splits are advantageous. The treatment also emphasizes the versatility of deterministic and stochastic successive convex approximation strategies for the optimization of C-RANs.

1.2
Introduction

In a Cloud Radio Access Network (C-RAN) architecture, the base station functionalities, from the physical layer to higher layers, are implemented in a virtualized fashion on centralized general-purpose processors rather than on the local hardware of the base stations or access points. This results in a novel cellular architecture in which low-cost wireless access points, which retain only radio functionalities and are known as Remote Radio Heads (RRHs), are centrally managed by a reconfigurable centralized "cloud", or baseband unit (BBU). At a high level, the C-RAN concept can be seen as an instance of network function virtualization techniques and hence as the RAN counterpart of the separation of control and data planes proposed for the core network in software-defined networking (see, e.g., [?]). The C-RAN architecture has the following key advantages:

- Reduced capital expense due to the possibility to substitute full-fledged base stations with RRHs with reduced space and energy requirements;
- Statistical multiplexing gain due to flexible allocation of radio and computing resources across all the connected RRHs;
- Easier implementation of coordinated and cooperative transmission/ reception strategies, such as eICIC and CoMP in LTE-A;
- Simplified network upgrades and maintenance due to the centralization of RAN functionalities.

In the uplink, the RRHs are required to convey their respective received signals, either in analog format or in the form of digitized baseband samples, to the BBU for processing. In a dual fashion, in a C-RAN downlink, each RU needs to receive from the BBU either directly the analog radio signal to be transmitted on the radio interface, or a digitized version of the corresponding baseband samples. The RU-CU bidirectional links that carry such information are referred to as *fronthaul* links, in contrast to the backhaul links connecting the CU to the core network. The analog transport solution is typically implemented on fronthaul links by means of radio-over-fiber (see, e.g., [?]). Instead, the digital transmission of baseband, or IQ, samples is currently carried out by following the CPRI standard [?], which most commonly requires fiber optic fronthaul links. The digital approach appears to be favored due to the traditional advantages of digital solutions, including resilience to noise and hardware impairments and flexibility in the transport options (see, e.g., [?]).

The main roadblock to the realization of the mentioned promises of C-RANs hinges on the inherent restriction on bandwidth and latency of the fronthaul links that may limit the advantages of centralized processing at the BBU. For example, implementing the CPRI standard, the bit rate required for an LTE base station that serves three cell sectors with carrier aggregation over five carriers and two receive antennas exceeds even the 10 Gbit/s provided by standard fiber optics links [?]. This problem is even more pronounced for networks in which fiber optic links are not available due to the large capital expense required for their deployment, as for heterogeneous networks with smaller RRHs.

1.2.1 Overview

As reported in [? ?], the bottleneck on the performance of C-RANS due to the capacity limitations of the fronthaul links can be alleviated by implementing a more flexible separation of functionalities between RRHs and BBU rather than performing all baseband processing at the BBU. Examples of baseband operations that can be carried out at the RRH include FFT/IFFT, demapping, synchronization, channel estimation, precoding and channel encoding. Note that [?] also investigates the possibility to implement functions at the higher layers, such as error detection, at the RRHs. In this chapter, we explore the problem of optimal functional split between RRHs and BBU at the physical layer by focusing on the two key baseband operations of channel encoding and channel encoding/ precoding. The content of the chapter is summarized as follows.

- For the uplink, we compare the standard implementation in which all baseband processing, including channel estimation, is performed at the BBU with an alternative architecture in which channel estimation, along with the necessary frame synchronization and resource demapping, is instead implemented at the RRH[1] in Sec. 1.3.
- For the downlink, we contrast the standard C-RAN implementation with one in which channel encoding and precoding are applied at the RRH, while the BBU retains the function of designing the precoding matrices based on the available channel state information in Sec. 1.4.

Throughout, we take an information-theoretic approach in order to evaluate the analytical expressions for the achievable performance. The analysis is corroborated by extensive numerical results that provide insight into the performance comparisons highlighted above. The chapter is concluded in Sec. **??**.

1.3
Uplink: Where to Perform Channel Estimation?

In this section, we study the uplink and address the potential advantages that could be accrued by performing channel estimation at the RRH rather than at the BBU. The rationale for the exploration of this functional split is that communicating the digitized signal received within the training portion of the received signal, as done in the conventional implementation, may impose a more significant burden on the fronthaul network that communicating directly the estimated channel state information (CSI). It is also supported by the known information-theoretic optimality of separate estimation and compression [?]. In particular, we compare two different approaches: (*i*) *Conventional approach*: The RRHs compress the training signal and CSI estimation takes place at the BBU; and *ii Channel Estimation at the RRH*: The RRHs perform CSI estimation and forward a compressed version of the CSI to the BBU. Note that the conventional approach was studied in [2] and that this section is adapted from [?]. We start by discussing the system model in Sec. 1.3.2 and then elaborate on the two approaches in Sec. 1.3.3 and Sec. 1.3.4. Finally, we present some numerical results in Sec. 1.3.6.

[1] We focus on flat fading channels and hence implicitly assume that FFT/IFFT is implemented at the RRH so as to enable per-subcarrier procesing of an OFDM signal.

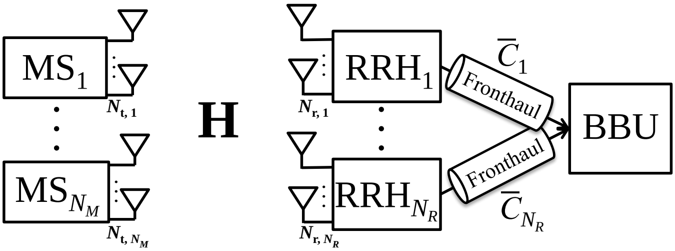

Figure 1.1 Uplink of a C-RAN system consisting of N_M MSs and N_R RRHs with the fronthaul links of capacity $\bar{C}_j$.

1.3.1 System Model

1.3.2 System Model

Consider the uplink of a cellular system consisting of N_M MSs, N_R RRHs and a BBU, as shown in Fig. 1.1. We denote the set of all MSs as $\mathcal{N}_M = \{1, \ldots, N_M\}$ and of all RRHs as $\mathcal{N}_R = \{1, \ldots, N_R\}$. The MSs, the i-th of which has $N_{t,i}$ transmit antennas, communicate in the uplink to the RRHs, where the j-th RRH is equipped with $N_{r,j}$ receive antennas. Each j-th RRH is connected to the BBU via a fronthaul link of capacity $\bar{C}_j$. All rates, including $\bar{C}_j$, are normalized to the bandwidth available on the uplink channel from MSs to RRHs and are measured in bits/s/Hz. More precisely, we assume that $\bar{C}_j TB$ bits can be transmitted on the fronthaul by any j-th RRH over an arbitrary number B of coherence blocks. Note that each j-th RRH can thus allocate its fronthaul bits across different coherence blocks. This is akin to the standard long-term power constraints considered in a large part of the literature on fading channels (see, e.g., [1]). We define $N_{\min} = \min(N_t, N_r)$ and $N_{\max} = \max(N_t, N_r)$ where N_t and N_r are the number of total transmit antennas and total receive antennas, that is $N_t = \sum_{i=1}^{N_M} N_{t,i}$ and $N_r = \sum_{j=1}^{N_R} N_{r,j}$, respectively.

The channel coherence block, of length T channel uses, is split it into a phase for channel training of length T_p channel uses and a phase for data transmission of length T_d channel uses, with

$$T_p + T_d = T, \tag{1.1}$$

as in [2, 3, 4, 5]. The signal transmitted by the i-th MS is given by a $N_{t,i} \times T$ complex matrix $\mathbf{X}_i$, where each column corresponds to the signal transmitted by the $N_{t,i}$ antennas in a channel use. This signal is divided into the $N_{t,i} \times T_p$ pilot signal $\mathbf{X}_{p,i}$ and the $N_{t,i} \times T_d$ data signal $\mathbf{X}_{d,i}$. We assume that the transmit signal $\mathbf{X}_i$ has a total per-block power constraint $\frac{1}{T} \left\|\mathbf{X}_i\right\|^2 = P_i$, and we define $\frac{1}{T_p} \left\|\mathbf{X}_{p,i}\right\|^2 = P_{p,i}$ and $\frac{1}{T_d} \left\|\mathbf{X}_{d,i}\right\|^2 = P_{d,i}$ as the powers used for training and data, respectively by the

i-th MS. In terms of pilot and data signal powers, then, the power constraint becomes

$$\frac{T_p}{T} P_{p,i} + \frac{T_d}{T} P_{d,i} = P_i. \tag{1.2}$$

For simplicity, we assume equal transmit power allocation for each antenna of all MSs, and hence we have $P_i = P$, $P_{d,i} = P_d$ and $P_{p,i} = P_p$ for all $i \in \mathcal{N}_M$. We define $\mathbf{X}_p$ and $\mathbf{X}_d$ as the overall pilot signal and the data signal transmitted by all MSs, respectively, i.e., $\mathbf{X}_p = [\mathbf{X}_{p,1}^T, \ldots, \mathbf{X}_{p,N_M}^T]^T$ and $\mathbf{X}_d = [\mathbf{X}_{d,1}^T, \ldots, \mathbf{X}_{d,N_M}^T]^T$.

As in [2, 3], we assume that coding is performed across multiple channel coherence blocks. This implies that the ergodic capacity describes the system performance in terms of achievable sum-rate. Moreover, the training signal is $\mathbf{X}_p = \sqrt{\frac{P_p}{N_t}}\mathbf{S}_p$ where $\mathbf{S}_p$ is a $N_t \times T_p$ matrix of i.i.d. $\mathcal{CN}(0,1)$ variables. This implies that an independently generated training sequence with power P_p/N_t is transmitted from each transmitting antenna across all MSs. Similarly, during the data phase, the MSs transmit independent streams with power P_d/N_t from its transmitting antennas using spatial multiplexing. As a result, we have $\mathbf{X}_d = \sqrt{\frac{P_d}{N_t}}\mathbf{S}_d$ where $\mathbf{S}_d$ is a $N_t \times T_d$ matrix of i.i.d. $\mathcal{CN}(0,1)$ variables.

The $N_{r,j} \times T$ signal $\mathbf{Y}_j$ received by the j-th RRH in a given coherence block, where each column corresponds to the signal received by the $N_{r,j}$ antennas in a channel use, can be split into the $N_{r,j} \times T_p$ received pilot signal $\mathbf{Y}_{p,j}$ and the $N_{r,j} \times T_d$ data signal $\mathbf{Y}_{d,j}$. The received signal at the j-th RRH is then given by

$$\mathbf{Y}_{p,j} = \sqrt{\frac{P_p}{N_t}}\mathbf{H}_j\mathbf{S}_p + \mathbf{Z}_{p,j} \tag{1.3a}$$

$$\text{and}\quad \mathbf{Y}_{d,j} = \sqrt{\frac{P_d}{N_t}}\mathbf{H}_j\mathbf{S}_d + \mathbf{Z}_{d,j}, \tag{1.3b}$$

where $\mathbf{Z}_{p,j}$ and $\mathbf{Z}_{d,j}$ are respectively the $N_{r,j} \times T_p$ and $N_{r,j} \times T_d$ matrices of independent and identically distributed (i.i.d.) complex Gaussian noise variables with zero-mean and unit variance, i.e, $\mathcal{CN}(0,1)$. The $N_{r,j} \times N_t$ channel matrix $\mathbf{H}_j$ collects all the $N_{r,j} \times N_{t,i}$ channel matrix $\mathbf{H}_{ji}$ from the i-th MS to the j-th RRH as $\mathbf{H}_j = [\mathbf{H}_{j1}, \ldots, \mathbf{H}_{jN_M}]$.

The channel matrix $\mathbf{H}_{ji}$ is modeled as Rician fading with the line-of-sight (LOS) component $\bar{\mathbf{H}}_{ji}$, which is deterministic, and the scattered component $\mathbf{H}_{w,ji}$ with i.i.d. $\mathcal{CN}(0,1)$ entries. Overall, the channel matrix $\mathbf{H}_{ji}$ between the j-th RRH and the i-th MS is represented as

$$\mathbf{H}_{ji} = \sqrt{\alpha}_{ji}\left(\sqrt{\frac{K}{K+1}}\bar{\mathbf{H}}_{ji} + \sqrt{\frac{1}{K+1}}\mathbf{H}_{w,ji}\right), \tag{1.4}$$

where the Rician factor K defines the power ratio of the LOS component and the scattered component, and the parameter α_{ji} represents the power gain between the j-th RRH and the i-th MS. The channel matrix $\mathbf{H}_j$ is assumed to be constant during each channel coherence block and to change according to an ergodic process from block to block.

1.3.3 Conventional Approach

With the CFE scheme, the RRH compresses both its received pilot signal (1.3a) and its received data signal (1.3b), and forwards them to the BBU on the fronthaul link. The BBU then estimates the CSI based on the received compressed pilot signals and performs coherent decoding.

Training Phase During the training phase, the vector of received training signals $\mathbf{Y}_p$ (1.3a) across all coherence times is compressed as

$$\widehat{\mathbf{Y}}_p = \mathbf{Y}_p + \mathbf{Q}_p, \tag{1.5}$$

where the compression noise matrix $\mathbf{Q}_p$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma_p^2)$ entries (see Remark **??**). Based on (1.5), the channel matrix $\mathbf{H}_i$ from i-th MS to the RRH is estimated at the BBU by the minimum mean square error (MMSE) method. Hence, it can be expressed as

$$\mathbf{H}_i = \widehat{\mathbf{H}}_i + \mathbf{E}_i, \tag{1.6}$$

where the estimated channel $\widehat{\mathbf{H}}_i$ is a complex Gaussian matrix with mean matrix $\sqrt{\frac{\alpha_i K}{K+1}}\bar{\mathbf{H}}_i$ and covariance matrix $\sigma^2_{\widehat{h}_i}\mathbf{I}_{N_r N_{t,i}}$, and the estimation error $\mathbf{E}_i$ has i.i.d. $\mathcal{CN}(0, \sigma^2_{e_i})$ entries. The variances of the estimated channel and the estimation error can be calculated as $\sigma^2_{\widehat{h}_i} = \frac{\frac{\alpha_i}{K+1} T_p P_p}{T_p P_p + N_t(1+\sigma_p^2)(K+1)}$ and $\sigma^2_{e_i} = \frac{\alpha_i N_t(1+\sigma_p^2)}{T_p P_p + N_t(1+\sigma_p^2)(K+1)}$, respectively (see, e.g., [3, 9]).

Data Phase The compressed data signal received at the BBU in (**??**) can be written as the sum of a useful term $\widehat{\mathbf{H}}\mathbf{X}_d$ and of the equivalent noise $\mathbf{N}_d = \mathbf{E}\mathbf{X}_d + \mathbf{Z}_d + \mathbf{Q}_d$, namely

$$\widehat{\mathbf{Y}}_d = \widehat{\mathbf{H}}\mathbf{X}_d + \mathbf{N}_d, \tag{1.7}$$

where the equivalent noise $\mathbf{N}_d$ has zero-mean and covariance matrix

$$\begin{aligned} \mathbf{R}_N &= E[\text{vec}(\mathbf{N}_d)\text{vec}(\mathbf{N}_d)^\dagger] \\ &= \left(1 + \sigma_d^2 + \frac{P_d}{N_t}\sum_{i=1}^{N_M} N_{t,i}\sigma^2_{e_i}\right)\mathbf{I}_{N_r T_d}. \end{aligned} \tag{1.8}$$

Ergodic Achievable Rate The ergodic capacity that can be attained with the assumed Gaussian input distribution[2] is given by the mutual information $\frac{1}{T}I(\mathbf{X}_d; \widehat{\mathbf{Y}}_d|\widehat{\mathbf{H}})$ [bits/s/Hz] (see, e.g, [8, Ch. 3]), which is bounded in the next lemma.

[2] Given the presence of imperfect CSI at the receiver, a Gaussian input distribution is generally not optimal in terms of capacity (see, e.g., [10]).

Lemma 1 *Let C_p and C_d define the fronthaul rates allocated respectively to the compressed pilot and data signals on the fronthaul from the RRH to the BBU. The ergodic capacity for the CFE strategy can be bounded as $\frac{1}{T}I(\mathbf{X}_d;\widehat{\mathbf{Y}}_d|\widehat{\mathbf{H}}) \geq R$, where*

$$R = \frac{T_d}{T}E\left[\log_2 \det\left(\mathbf{I}_{N_r} + \rho_{\mathit{eff}}\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right)\right], \tag{1.9}$$

with $\rho_{\mathit{eff}} = \frac{P_d}{N_t\left(1+\sigma_d^2+\frac{P_d}{N_t}\sum_{i=1}^{N_M} N_{t,i}\sigma_{e_i}^2\right)}$, and $\widehat{\mathbf{H}}$ being distributed as in (1.6). Moreover, the quantization noise powers (σ_p^2, σ_d^2) must satisfy the fronthaul constraint $C_p + C_d = \bar{C}$, where

$$C_d = \frac{T_d}{T}\log_2\det\left(\mathbf{I}_{N_r} + \frac{\frac{P_d}{N_t}E[\mathbf{H}\mathbf{H}^\dagger] + \mathbf{I}_{N_r}}{\sigma_d^2}\right) \tag{1.10a}$$

$$\text{and } \; C_p = \frac{T_p}{T}\log_2\det\left(\mathbf{I}_{N_r} + \frac{\frac{P_p}{N_t}E[\mathbf{H}\mathbf{H}^\dagger] + \mathbf{I}_{N_r}}{\sigma_p^2}\right), \tag{1.10b}$$

with $E[\mathbf{H}\mathbf{H}^\dagger] = \left(\frac{K}{K+1}\bar{\mathbf{H}}\bar{\mathbf{H}}^\dagger + \frac{\sum_{i=1}^{N_M}\alpha_i N_{t,i}}{K+1}\mathbf{I}_{N_r}\right)$.

For the CFE scheme, the ergodic achievable sum-rate (1.9) can now be optimized over the fronthaul allocation (C_p, C_d) under the fronthaul constraint $\bar{C} = C_p + C_d$, with C_p and C_d in (1.10), by maximizing the effective SNR ρ_{eff} in (1.9). This non-convex problem can be tackled using a line search method [13] in a bounded interval (e.g., over C_p in the interval $[0, \bar{C}]$).

Remark 1 *The lower bound R on the ergodic capacity in (1.9), and related bounds in the next section, will be referred thereafter as the ergodic achievable rate.*

1.3.4 Channel Estimation at the RRHs

Here, we introduce the ECF approach. Accordingly, each RRH estimates the CSI based on its received pilot signal (1.3a), and then compresses both its estimated CSI and its received data signal (1.3b) for transmission on the fronthaul. In this section, we introduce the key common quantities that define the class of ECF schemes, which are then studied in Section 1.3.5 for the single RRH case and in Section **??** for the more general multiple RRHs case.

1.3.4.1 Training Phase

The MMSE estimate of $\mathbf{H}_j$ performed at the j-th RRH given the observation $\mathbf{Y}_{p,j}$ in (1.3a) is given by

$$\widetilde{\mathbf{H}}_j = \sqrt{\frac{N_t}{P_p}}\bar{\mathbf{Y}}_{p,j}\mathbf{S}_p^\dagger\left(\frac{N_t\,(K+1)}{P_p}\mathbf{I}_{N_r} + \mathbf{S}_p\mathbf{S}_p^\dagger\right)^{-1} + \sqrt{\frac{K}{K+1}}\bar{\mathbf{H}}_j, \tag{1.11}$$

where $\bar{\mathbf{Y}}_{p,j} = \mathbf{Y}_{p,j} - \sqrt{\frac{P_p}{N_t}\frac{K}{K+1}}\bar{\mathbf{H}}_j\mathbf{S}_p$ and $\bar{\mathbf{H}}_j = [\sqrt{\alpha}_{j1}\bar{\mathbf{H}}_{j1}, \ldots, \sqrt{\alpha}_{jN_M}\bar{\mathbf{H}}_{jN_M}]$ (see, e.g., [3, 9]). The estimated channel $\widetilde{\mathbf{H}}_j = [\widetilde{\mathbf{H}}_{j1}, \ldots, \widetilde{\mathbf{H}}_{jN_M}]$ in (1.11) is such that the estimated channel matrix $\widetilde{\mathbf{H}}_{ji}$ corresponding to the channel between the j-th RRH and i-th MS has a matrix-variate complex Gaussian distribution with mean matrix $\sqrt{\frac{\alpha_{ji}K}{K+1}}\bar{\mathbf{H}}_{ji}$ and covariance matrix $\sigma^2_{\tilde{h}_{ji}}\mathbf{I}_{N_{r,j}}$, where $\sigma^2_{\tilde{h}_{ji}} = \frac{\frac{\alpha_{ji}}{K+1}T_pP_p}{T_pP_p+N_t(K+1)}$. Moreover, we can decompose the channel matrix $\mathbf{H}_{ji}$ into the estimate $\widetilde{\mathbf{H}}_{ji}$ and the independent estimation error $\mathbf{E}_{ji}$, as

$$\mathbf{H}_{ji} = \widetilde{\mathbf{H}}_{ji} + \mathbf{E}_{ji}, \tag{1.12}$$

where the error $\mathbf{E}_{ji}$ has i.i.d. $\mathcal{CN}(0, \sigma^2_{e_{ji}})$ entries with $\sigma^2_{e_{ji}} = \frac{\alpha_{ji}N_t}{T_pP_p+N_t(K+1)}$.

The sequence of channel estimates $\widetilde{\mathbf{H}}_j$ for all coherence times in the coding block is compressed by the j-th RRH and forwarded to the BBU on the fronthaul link. The compressed channel $\widehat{\mathbf{H}}_j$ is related to the estimate $\widetilde{\mathbf{H}}_j$ as

$$\widetilde{\mathbf{H}}_j = \widehat{\mathbf{H}}_j + \mathbf{Q}_{p,j}, \tag{1.13}$$

where the $N_{r,j} \times N_t$ quantization noise matrix $\mathbf{Q}_{p,j}$ has zero-mean i.i.d. $\mathcal{CN}(0, \sigma^2_{p,j})$ entries (see Remark **??**) and the compressed estimate $\widehat{\mathbf{H}}_j$ is complex Gaussian with mean matrix $\sqrt{\frac{K}{K+1}}\bar{\mathbf{H}}_j$ and covariance matrix $\mathbf{R}_{\tilde{h}_j} - \sigma^2_{p,j}\mathbf{I}_{N_t}$, where $\mathbf{R}_{\tilde{h}_j}$ is diagonal matrix with main diagonals given by $[\sigma^2_{\tilde{h}_{j1}}\mathbf{I}_{N_{t,1}}, \ldots, \sigma^2_{\tilde{h}_{jN_M}}\mathbf{I}_{N_{t,N_M}}]$ (see, e.g., [8, Ch. 3]). We will discuss in Section 1.3.5 and Section **??** how to relate the quantization noise variance $\sigma^2_{p,j}$ to the fronthaul capacity $\bar{C}_j$.

1.3.4.2 **Data Phase**

During the data phase, the j-th RRH compresses the signal $\mathbf{Y}_{d,j}$ in (1.3b) and sends it to the BBU on the fronthaul link. The received signals at the BBU are related to $\mathbf{Y}_{d,j}$ as

$$\widehat{\mathbf{Y}}_{d,j} = \mathbf{Y}_{d,j} + \mathbf{Q}_{d,j}, \tag{1.14}$$

where $\mathbf{Q}_{d,j}$ is independent of $\mathbf{Y}_{d,j}$ and represents the quantization noise matrix[3]. This is assumed to be zero-mean complex Gaussian with covariance matrix $E[\text{vec}(\mathbf{Q}_{d,j})\text{vec}(\mathbf{Q}_{d,j})^\dagger] = \mathbf{R}_{d,j} \otimes \mathbf{I}_{T_d}$. By this definition, $\mathbf{R}_{d,j}$ is the covariance matrix of the $N_{r,j} \times 1$ compression noise vector for all the channel uses in a data transmission period. Following our design choices for the other quantization noises, we will mostly assume $\mathbf{R}_{d,j}$ to be a scaled identity matrix, namely $\mathbf{R}_{d,j} = \sigma^2_{d,j}\mathbf{I}_{N_{r,j}T_d}$ (see Remark **??**). However, we will allow this covariance

[3] Note that we use a different formulation for the quantization test channel (see, e.g., [8, Ch. 3]) in (1.14) with respect to (1.13). In (1.14) and similarly in (**??**) and (1.5), in fact, the quantization noise is added to the signal to be compressed. While the formulation in (1.13) is optimal from a rate-distortion point of view [8, Ch. 3], the test channel (1.14) is selected here for its analytical convenience. It is noted that this test channel is assumed in many previous studies, including [14, 15, 2, 16].

matrix to be arbitrary in Section 1.3.5.3 in order to illustrate the potential advantages of a system design that adapts the quantizers to the current channel conditions (see also Remark **??**). The relationship of matrix $\mathbf{R}_{d,j}$ with the fronthaul capacity will be clarified in the next sections.

We close this section by deriving a model for the received signals at the BBU that is akin to (1.7)-(1.8) for CFE. With ECF, the BBU recovers the sequence of quantized data signals $\widehat{\mathbf{Y}}_{d,j}$ in (1.14) and of quantized channel estimates $\widehat{\mathbf{H}}_j$ in (1.13) from the information received on the fronthaul link. Separating the desired signal and the noise in (1.14), the received signal $\widehat{\mathbf{Y}}_{d,j}$ from the j-th RRH can be expressed as

$$\widehat{\mathbf{Y}}_{d,j} = \widehat{\mathbf{H}}_j \mathbf{X}_d + \mathbf{N}_{d,j}, \tag{1.15}$$

where $\mathbf{N}_{d,j}$ denotes the equivalent noise $\mathbf{N}_{d,j} = (\mathbf{Q}_{p,j} + \mathbf{E}_j)\,\mathbf{X}_d + \mathbf{Z}_{d,j} + \mathbf{Q}_{d,j}$, which has zero-mean and covariance matrix

$$\mathbf{R}_{N_j}{=}E[\text{vec}(\mathbf{N}_{d,j})\text{vec}(\mathbf{N}_{d,j})^\dagger]{=}\mathbf{R}_{d,j} \otimes \mathbf{I}_{T_d}{+}\sigma^2_{pe_j}\mathbf{I}_{N_{r,j}T_d} \tag{1.16}$$

with

$$\sigma^2_{pe,j} = \left(1 + P_d\left(\sigma^2_{p,j} + \frac{\sum_{i=1}^{N_M} N_{t,i}\sigma^2_{e_{ji}}}{N_t}\right)\right), \tag{1.17}$$

where we have used the relations $E[\mathbf{Q}_{p,j}\mathbf{Q}^\dagger_{p,j}] = N_t\sigma^2_{p,j}\mathbf{I}_{N_{r,j}}$ and $E[\mathbf{E}_j\mathbf{E}^\dagger_j] = \sum_{i=1}^{N_M} N_{t,i}\sigma^2_{e_{ji}}\mathbf{I}_{N_{r,j}}$. We observe that, as in (1.7)-(1.8), $\mathbf{N}_{d,j}$ is not Gaussian distributed and is not independent of $\mathbf{X}_d$ (see also [3]).

1.3.5 Analysis of ECF : The Single Base Station Case

In this section, we discuss how to calculate the compression noises statistics, namely $\sigma^2_{p,j}$ for the estimated CSI (see (1.13)) and $\mathbf{R}_{d,j}$ for the data (see (1.14)). We consider three different strategies in order of complexity, namely separate compression, joint compression and joint adaptive compression of estimated CSI and received data signal. Specifically, here, we first consider the single base station case, i.e., $N_R = 1$. The more complex scenario with multiple RRHs will be studied in Section **??** by building on the analysis in this section. For simplicity of notation, we drop the RRH index in this section.

1.3.5.1 Separate Compression of Channel and Received Data Signal

Here, we consider the conventional option of compressing separately the sequence of the estimated channels $\widetilde{\mathbf{H}}$ and of the received data signals $\mathbf{Y}_d$. For simplicity, and due to the identical distribution of the entries of $\mathbf{Y}_d$, here we choose $\mathbf{R}_d = \sigma^2_d\mathbf{I}_{N_r}$ (see Remark 1).

Proposition 1 *Let C_p and C_d denote respectively the fronthaul rates allocated for the transmission of the compressed channel estimates (1.13) and of the compressed* *received signals (1.14) on the fronthaul link from the RRH to the BBU. The ergodic achievable sum-rate for separate compression strategy is given as*

$$R = \frac{T_d}{T} E\left[\log_2 \det\left(\mathbf{I}_{N_r} + \rho_{\mathit{eff}} \widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right)\right], \tag{1.18}$$

with

$$\rho_{\mathit{eff}} = \frac{P_d}{N_t\left(1 + \sigma_d^2 + P_d\left(\sigma_p^2 + \sum_{i=1}^{N_M} N_{t,i}\sigma_{e_i}^2/N_t\right)\right)}, \tag{1.19}$$

with $\widehat{\mathbf{H}}$ *being distributed as in (1.13), and with* $\sigma_{e_i}^2$ *in (1.12). Moreover, the quantization noise powers* (σ_p^2, σ_d^2) *must satisfy the fronthaul constraint* $C_p + C_d = \bar{C}$*, where*

$$C_p = \frac{N_r}{T}\log_2\left(\frac{\prod_{i=1}^{N_M}\left(\sigma_{\tilde{h}_i}^2\right)^{N_{t,i}}}{(\sigma_p^2)^{N_t}}\right) \tag{1.20a}$$

$$\text{and} \quad C_d = \frac{T_d}{T}\log_2\det\left(\mathbf{I}_{N_r} + \frac{\frac{P_d}{N_t}E[\mathbf{H}\mathbf{H}^\dagger] + \mathbf{I}_{N_r}}{\sigma_d^2}\right), \tag{1.20b}$$

with $\sigma_{\tilde{h}_i}^2$ *being given in (1.11).*

As for CFE, the ergodic achievable sum-rate (1.18) can now be optimized over the fronthaul allocation (C_p, C_d) under the fronthaul constraint $\bar{C} = C_p + C_d$, with C_p and C_d in (1.20), by maximizing the effective SNR ρ_{eff} in (1.19) using a line search [13] in a bounded interval.

Remark 2 *If we consider the special case of a Rayleigh fading channel, that is* $K = 0$*, the ergodic achievable sum-rate (1.18) can be evaluated explicitly following [17]. Moreover, by imposing equality in (1.20b), we can easily calculate the quantization variance* σ_d^2 *as*

$$\sigma_d^2 = \frac{\frac{P_d}{N_t}\sum_{i=1}^{N_M}\alpha_i N_{t,i} + 1}{2^{TC_d/(N_r T_d)} - 1}. \tag{1.21}$$

Remark 3 *For Rayleigh fading (*$K = 0$*) and* $N_r = N_t = 1$*, the ergodic achievable sum-rate (1.9) obtained with CFE equals the ergodic achievable sum-rate (1.18) with ECF based on separate compression. Further comparisons among the discussed methods will be presented in Section* **??** *via numerical results.*

Remark 4 *In the discussion above, we have considered the power allocation* (P_p, P_d) *and the time allocation* (T_p, T_d) *as fixed. The optimization of these parameters can be carried out similar to [3] and is not further detailed here.*

1.3.5.2 **Joint Compression of Channel and Received Data Signal**

Here we propose a more sophisticated method to convey the sequence of the channel estimates $\widehat{\mathbf{H}}$ in (1.13) and of received data signals $\widehat{\mathbf{Y}}_d$ in (1.14) over the fronthaul link. This method leverages the fact that channel estimates $\widetilde{\mathbf{H}}$ in (1.12) and received signals $\mathbf{Y}_d$ in (1.3b), and thus $\widehat{\mathbf{H}}$ and $\widehat{\mathbf{Y}}_d$, are correlated. As in Section 1.3.5.1, we assume an uncorrelated compression covariance $\mathbf{R}_d = \sigma_d^2 \mathbf{I}_{N_r}$ in (1.14) and we are interested in finding the optimal pair (σ_p^2, σ_d^2).

Proposition 2 *The ergodic achievable sum-rate for joint compression strategy can be bounded as (1.18), where ρ_{eff} is given by (1.19). Moreover, the quantization noise powers (σ_p^2, σ_d^2) must satisfy the fronthaul constraint $C_p + C_d = \bar{C}$, where*

$$C_d = \frac{T_d}{T}\Big(E\left[\log_2\det\left(\mathbf{I}_{N_r}{+}\rho_{\mathit{eff}}\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right)\right]{+}\ N_r\log_2\left(\sigma_{pe}^2{+}\sigma_d^2\right){-}N_r\log_2\sigma_d^2\Big)\,, \tag{1.22}$$

*and C_p is defined in (**??**), with $\widehat{\mathbf{H}}$ being distributed as in (1.13) and σ_{pe}^2 being given in (1.17).*

The ergodic achievable sum-rate (1.18) can now be optimized over the quantization noise powers (σ_p^2, σ_d^2) under the fronthaul constraint $\bar{C} = C_p + C_d$, with C_p in (**??**) and C_d in (1.22), using a two-dimensional search.

Remark 5 *It is useful to compare the fronthaul constraint in (1.20), corresponding to separate compression, with $\bar{C} = C_p{+}C_d$, which applies to joint compression with C_p in (**??**) and C_d in (1.22). To this end, we observe that (1.20) can be expressed in terms of the quantization noise variance σ_p^2 and σ_d^2 using (**??**) and (1.20b), leading to the condition*

$$\bar{C} = C_p + \frac{T_d}{T}\log_2\det\left(\mathbf{I}_{N_r} + \frac{\frac{P_d}{N_t}E[\mathbf{H}\mathbf{H}^\dagger] + \mathbf{I}_{N_r}}{\sigma_d^2}\right). \tag{1.23}$$

*The difference between (1.23) and the condition $\bar{C} = C_p + C_d$, with C_p in (**??**) and C_d in (1.22), is given as*

$$\frac{T_d}{T}\left(\log_2\det\left(\mathbf{I}_{N_r}{+}\rho_{\mathit{eff}}E\left[\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right]\right)E\left[\log_2\det\left(\mathbf{I}_{N_r}{+}\rho_{\mathit{eff}}\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right)\right]\right) \geq 0, \tag{1.24}$$

*where the latter condition follows by Jensen's inequality since we have $E\left[\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right] = \frac{K}{K+1}\bar{\mathbf{H}}\bar{\mathbf{H}}^\dagger + (\sum_{i=1}^{N_M} N_{t,i}\sigma_{\tilde{h}_i}^2 - N_t\sigma_p^2)\mathbf{I}_{N_r}$. Inequality (1.24) shows that joint compression has the potential of improving the efficiency of fronthaul utilization. This will be further explored via numerical results in Section **??**.*

1.3.5.3 **Joint Adaptive Compression of Channel and Received Data Signal**

In this section, we introduce an improved method for joint compression of channel and received data signal. The main idea is that of adapting the covariance matrix $\mathbf{R}_d$

of the compression noise added to the data signal (see (1.14)) to the channel estimate in each channel coherence block. The rationale for this approach is that if, e.g., the channel quality in a coherence block is poor, there is no reason to invest significantly fronthaul capacity for the compression of the corresponding received data signal. We recall that, in the strategy studied in the previous section, the covariance matrix $\mathbf{R}_d$ was instead selected to be equal for all the coherence blocks (and given as $\mathbf{R}_d = \sigma_d^2 \mathbf{I}_{N_r T_d}$).

We start by observing that (**??**) suggests that joint compression can be performed in two steps: (i) first, the channel estimate sequence in compressed with required fronthaul rate $\frac{1}{T} I(\widetilde{\mathbf{H}}; \widehat{\mathbf{H}})$; (ii) then, given that the sequence of channel estimates $\widehat{\mathbf{H}}$ for all coherence blocks is known at both the RRH an the BBU, the RRH uses a different compression strategy for the quantization of $\mathbf{Y}_d$ depending on the value of $\widehat{\mathbf{H}}$[4)]. Based on this observation, we propose here to adapt the choice of matrix $\mathbf{R}_d$ to the current value of $\widehat{\mathbf{H}}$ for each coherence block. To emphasize this fact, we use the notation $\mathbf{R}_d(\widehat{\mathbf{H}})$.

Proposition 3 *For a given adaptive choice $\mathbf{R}_d(\widehat{\mathbf{H}})$ of the compression covariance matrix on the data signal, the ergodic achievable sum-rate for joint adaptive compression strategy is given as*

$$R = \frac{T_d}{T} E\left[\log_2 \det\left(\mathbf{I}_{N_t} + \frac{P_d}{N_t} \widehat{\mathbf{H}}^\dagger \left(\mathbf{R}_d(\widehat{\mathbf{H}}) + \sigma_{pe}^2 \mathbf{I}_{N_r}\right)^{-1} \widehat{\mathbf{H}}\right)\right], \tag{1.25}$$

where $\widehat{\mathbf{H}}$ is distributed as in (1.13) and σ_{pe}^2 is given in (1.17). Moreover, the quantization noise power σ_p^2 and the covariance matrices $\mathbf{R}_d(\widehat{\mathbf{H}})$ must satisfy the fronthaul constraint $C_p + C_d = \bar{C}$, where

$$C_d = \frac{T_d}{T} E\left[\log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{R}_d^{-1}(\widehat{\mathbf{H}}) \left(\frac{P_d}{N_t} \widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger + \sigma_{pe}^2 \mathbf{I}_{N_r}\right)\right)\right] \tag{1.26}$$

*and C_p is defined in (**??**).*

We now observe that the optimization of the compression covariance matrices $\mathbf{R}_d(\widehat{\mathbf{H}})$ of the data signal for a given the variance σ_p^2 can be carried out analytically. The problem of maximizing the ergodic achievable sum-rate (1.25) then reduces to a one-dimensional search over σ_p^2.

Proposition 4 *Define the eigenvalue decomposition*

$$\frac{P_d}{N_t} \widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger + \sigma_{pe}^2 \mathbf{I}_{N_r} = \mathbf{U}(\widehat{\mathbf{H}}) diag\left(t_1(\widehat{\mathbf{H}}), \ldots, t_{N_r}(\widehat{\mathbf{H}})\right) \mathbf{U}^\dagger(\widehat{\mathbf{H}}). \tag{1.27}$$

*The problem of maximizing the ergodic achievable sum-rate (1.25) under the constraint $\bar{C} = C_p + C_d$, with C_p in (**??**) and C_d in (1.26), admits the solution*

[4)] In practice, the values of $\widehat{\mathbf{H}}$ can be quantized in order to reduce the number of codebooks.

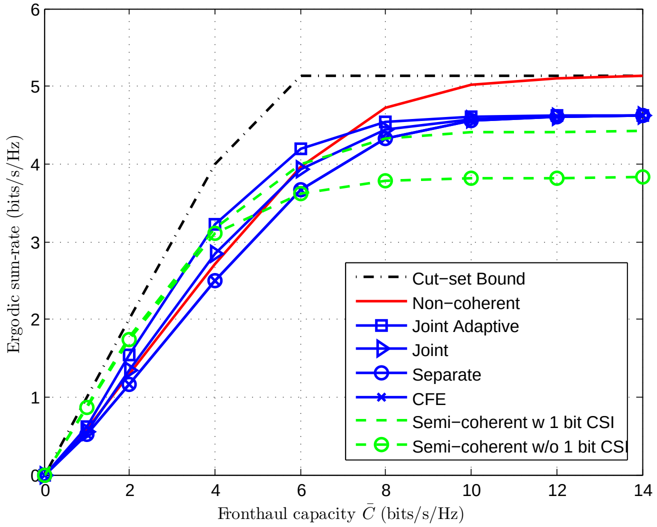

Figure 1.2 Ergodic achievable sum-rate vs. fronthaul capacity ($N_R = N_M = 1$, $N_t = N_r = 1$, $P = 20dB$, T = 10, and $K = 0$).

$\mathbf{R}_d(\widehat{\mathbf{H}}) = \mathbf{U}(\widehat{\mathbf{H}})diag(\lambda_1(\widehat{\mathbf{H}}), \ldots, \lambda_{N_r}(\widehat{\mathbf{H}}))^{-1}\mathbf{U}^{\dagger}(\widehat{\mathbf{H}})$, *where the inverse eigenvalues are given as*

$$\lambda_n^*(\widehat{\mathbf{H}}) = \left[\frac{1}{\mu}\left(\frac{1}{\sigma_{pe}^2} - \frac{1}{t_n(\widehat{\mathbf{H}})}\right) - \frac{1}{\sigma_{pe}^2}\right]^+, \tag{1.28}$$

*for $n = 1, \ldots, N_r$; σ_{pe}^2 is given in (1.17); the Lagrange multiplier μ^* is such that the condition $\bar{C} = C_p + C_d$, with C_p in (**??**) and C_d in (1.26), is satisfied with the equality.*

1.3.6
Numerical Results

In this section, we evaluate the performance of the proposed compression strategies for the uplink of a multi-cell system. Throughout, we assume that every MS is subject to the same power constraint P and that each RRH has the same fronthaul capacity $\bar{C}$, that is $P_i = P$ for $i \in \mathcal{N}_M$ and $\bar{C}_j = \bar{C}$ for $j \in \mathcal{N}_R$. Moreover, we set $\bar{\mathbf{H}}_j = \mathbf{1}_{N_{r,j} \times N_t}$. We optimize over the power allocation (P_p, P_d) and we set $T_p = N_t$ (except for the non-coherent scheme where $T_p = 0$), which was shown to be optimal in [3] for a point-to-point link with no fronthaul limitation.

We start by considering case of a single MS and a single RRH, namely $N_R = 1$ and $N_M = 1$ and consider the performance of the ECF schemes, of CFE and of non-coherent and semi-coherent processing. For the latter, we focus on the semi-coherent scheme with one-bit CSI and without one-bit CSI. Fig. 1.2 and Fig. 1.3

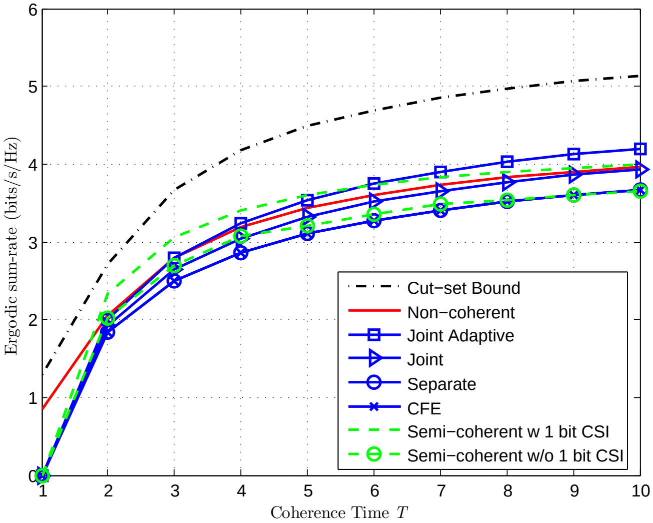

Figure 1.3 Ergodic achievable sum-rate vs. coherence time ($N_R = N_M = 1$, $N_t = N_r = 1$, $\bar{C}$ = 6 bits/s/Hz, $P = 20dB$, and $K = 0$).

show the ergodic achievable sum-rate for all the mentioned schemes as function of the fronthaul capacity $\bar{C}$ and coherence time T[5], respectively. For reference, in both figures, we also show the upper bound obtained by standard cut-set arguments, namely $\min(\bar{C}, R_{nc})$, where R_{nc} is the non-coherent capacity of the MS-RRH channel [6]. In Fig. 1.2, we set $N_t = N_r = 1$, power $P = 20dB$, coherence time $T = 10$ and consider Rayleigh fading channel, i.e., $K = 0$. At low fronthaul capacity $\bar{C}$ (here, $\bar{C} < 4$), it is seen that the semi-coherent strategy is to be preferred due to its ability to devote the limited fronthaul resources to convey only information about the data block to the BBU. Note that the semi-coherent scheme with one-bit CSI outperforms the case with no CSI unless the fronthaul capacity $\bar{C}$ is smaller or very close to $1/T$ (i.e., the overhead for the one-bit CSI on the fronthaul). Conversely, for sufficiently large fronthaul capacities (here, $\bar{C} > 7$), the non-coherent approach turns out to be advantageous. This is because, when the compression noise is negligible, the achievable rate is upper bounded by the non-coherent capacity[6] (see, e.g., [6]). Instead, for intermediate fronthaul values, ECF and CFE schemes are the preferred choice. Concerning the comparison between ECF and CFE, Fig. 1.2 demonstrates

[5] Consider a multicarrier system. The coherence bandwidth can be approximated as $1/(50\sigma_\tau)$, where σ_τ is the delay spread [22]. Therefore, by imposing $1/(50\sigma_\tau) = T\Delta f$, where Δf is the subcarrier spacing, one can find that a delay spread equal to $\sigma_\tau = 1/(50T\Delta f)$ causes a coherent block equal to T channel uses. For instance, with $\Delta f = 15kHz$, as for LTE systems, we get that $T = 1$ corresponds to $\sigma_\tau = 13\mu s$.

[6] In a non-coherent information-theoretic set-up, the optimization of the transmit signals allows, as a special case, the selection of a pilot-based transmission in which all codewords contain the same training sequence.

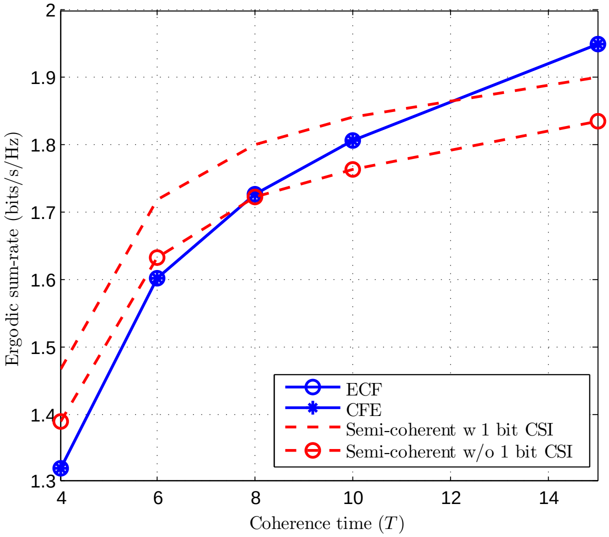

Figure 1.4 Ergodic achievable sum-rate vs. coherence time ($N_R = N_M = 1$, $N_t = N_r = 2$, $\bar{C} = 5$ bits/s/Hz, $P = 5dB$, and $K = 0$).

that the ECF strategy is advantageous. In particular, for the scenario at hand, CFE performs as ECF with separate compression as discussed in Section 1.3.5.1. However, progressively more complex ECF schemes have better performance, with the joint adaptive strategy outperforming the joint approach and the separate strategy. Finally, we note that the gains obtained by more complex ECF compression strategies are especially pronounced in the region of interest of moderate fronthaul capacity, in which the fronthaul capacity is at a premium and should be used efficiently.

The effect of an increase of the coherence time on the ergodic achievable sum-rate is instead investigated with $N_t = N_r = 1$, fronthaul capacity $\bar{C} = 6$, power $P = 20dB$, and Rayleigh fading in Fig. 1.3. The figure illustrates that the non-coherent strategy is clearly advantageous over the other schemes for $T = 1$ given that it operates without transmitting any pilot signal. Moreover, ECF with Joint adaptive compression is especially advantageous for large coherence time due to the increased relevance of an efficient compression of the data signal when $T_d \gg T_p$.

We set $N_t = N_r = 2$ and compare the performance of ECF, CFE and of non-coherent and semi-coherent processing with one-bit CSI and without one-bit CSI. The effect of an increase of the coherence time on the ergodic achievable sum-rate is investigated in Fig. 1.4 with fronthaul capacity $C = 5$ bits/s/Hz, and power $P = 5dB$. Note that the rate of the non-coherent scheme is significantly smaller than those of the semi-coherent and coherent schemes for the considered range of values of T and is hence not shown to enhance legibility. The figure illustrates that the semi-coherent strategy is clearly advantageous over the other schemes in the regime of a small coherence period due to its reduced fronthaul overhead. Instead, for larger

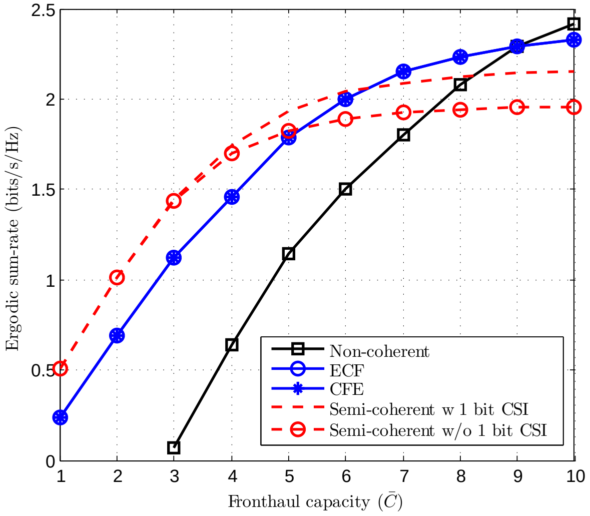

Figure 1.5 Ergodic achievable sum-rate vs. fronthaul capacity ($N_R = N_M = 1$, $N_t = N_r = 2$, $P = 10dB$, T = 4, and $K = 0$).

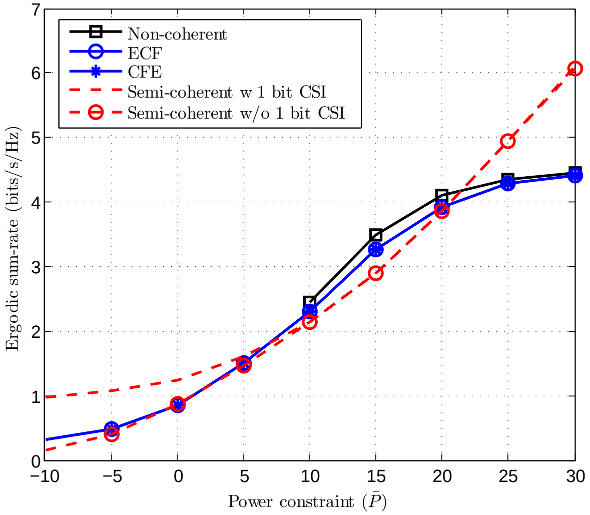

Figure 1.6 Ergodic achievable sum-rate vs. power constraint ($N_R = N_M = 1$, $N_t = N_r = 2$, $\bar{C} = 10$ bits/s/Hz, T = 4, and $K = 0$).

coherence times, the coherent ECF and CFE schemes are the preferred choice due to a decreasing share of the fronthaul rate required for CSI. In particular, for the scenario at hand, CFE and ECF have comparable performance (see also Remark 1).

In Fig. 1.5, we set the power as $P = 10dB$ and the coherence time as $T =$

4 and plot the ergodic achievable sum-rate versus the fronthaul capacity. At low fronthaul capacity C (here, $C < 6$), it is seen that the semi-coherent strategy is to be preferred due to its ability to reduce the fronthaul overhead. Note that the semi-coherent scheme with one-bit CSI outperforms the one with no CSI unless the fronthaul capacity C is smaller or very close to $1/T$ (i.e., the overhead for the one-bit CSI on the fronthaul). Conversely, for sufficiently large fronthaul capacities (here, $C > 9$), the non-coherent approach turns out to be advantageous. This is because, when the compression noise is negligible, the achievable rate is upper bounded by the non-coherent capacity[7)] (see, e.g., [6]). Instead, for intermediate fronthaul values, ECF and CFE schemes are to be preferred.

In Fig. 1.6, the ergodic sum-rate is plotted versus the power constraint P with backhaul capacity $C = 10$ bits/s/Hz, and coherence time $T = 4$. The rate with non-coherent decoding is plotted from power $10dB$ due to the validity of the formula in [4, Theorem 9] only in the high-SNR regime. It is seen that the semi-coherent strategy, which only quantizes the data signal, is superior to the other schemes for low power constraint. This is because, in the high-SNR regime, the quantization noise downrates the performance.

We now turn to consider a multiple RRHs and multiple MSs scenario with $N_R = N_M = 2$, $N_t = N_r = 4$ and focus on the comparison among the different proposed ECF schemes and CFE[8)]. The performance comparison among the proposed ECF schemes discussed above is confirmed by the results reported in Fig. 1.7, 1.8 and 1.9. Fig. 1.7 shows the ergodic achievable sum-rate of the three compression methods versus the transmit power P with fronthaul capacity $\bar{C} = 6$, coherence time $T = 10$, channel gain $\alpha_{ji} = 1$ for all $j \in \mathcal{N}_R, i \in \mathcal{N}_M$, and Rayleigh fading channel ($K = 0$). It is seen that the performance gains of more complex compression strategies is more evident in the high SNR regime, in which the compression noise imposes a significant bottleneck to the system performance.

In Fig. 1.8, the ergodic achievable sum-rate is plotted versus the inter-cell channel gain α_{ji} assumed to be the same for all $i \neq j$, while $\alpha_{jj} = 1$ for $j \in \mathcal{N}_R$, with fronthaul capacity $\bar{C} = 6$, power $P = 20dB$, coherence time $T = 10$ and Rayleigh fading. As it is well known (see, e.g., [23]), at low inter-cell gain, the inter-cell interference is deleterious; instead, when the inter-cell gain is large enough, the central decoder can take advantage of the additional signal paths and the sum-rate increases.

Finally, we show the impact of the Rician factor K in Fig. 1.9 with fronthaul capacity $\bar{C} = 6$, power $P = 20dB$ and channel gain $\alpha_{ji} = 1$ for all $j \in \mathcal{N}_R, i \in \mathcal{N}_M$. We observe that the performance of the joint adaptive compression method approaches that of the joint compression method as the Rician factor K increases. This is because the joint adaptive compression scheme is based on an optimization of the compression strategy that adapts the quantization error on the data signal to the channel

[7)] In a non-coherent information-theoretic set-up, the optimization of the transmit signals allows, as a special case, the selection of a pilot-based transmission in which all codewords contain the same training sequence.

[8)] With multiple RRHs and MSs, evaluating the non-coherent capacities, and thus also the cut-set bound is an open problem. Moreover, the evaluation of the performance of semi-coherent strategies is left for future work.

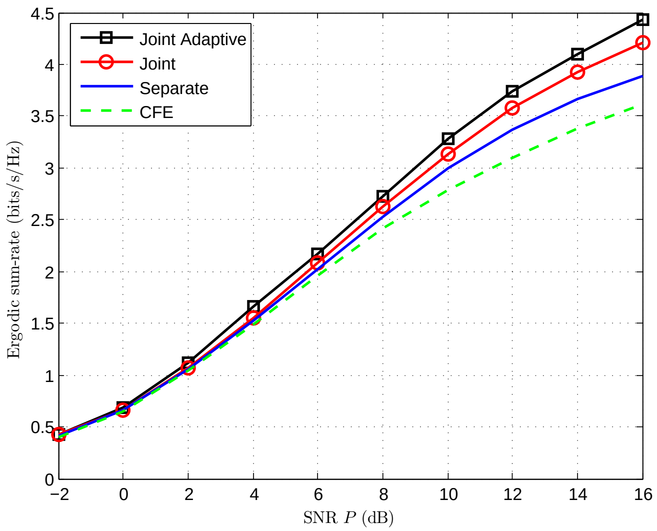

Figure 1.7 Ergodic achievable sum-rate vs. SNR (dB) ($N_R = N_M = 2$, $N_t = N_r = 4$, $\bar{C} = 6$ bits/s/Hz, $T = 10$, $\alpha_{ji} = 1$, and $K = 0$).

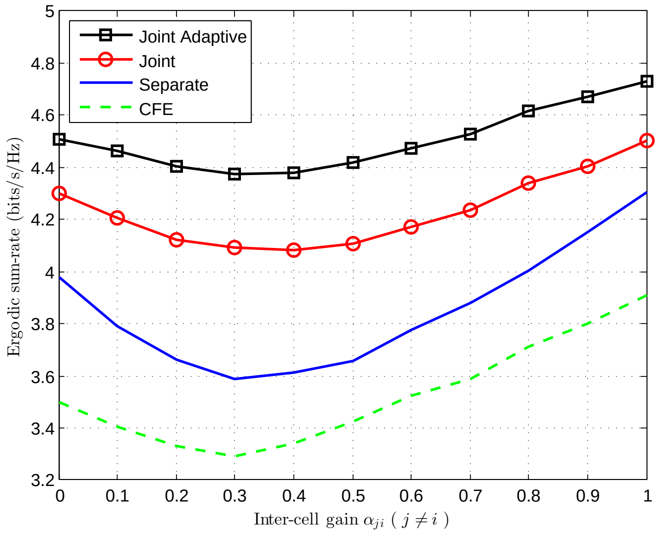

Figure 1.8 Ergodic achievable sum-rate vs. inter-cell gain α_{ji} ($N_R = N_M = 2$, $N_t = N_r = 4$, $\bar{C} = 6$ bits/s/Hz, $P = 20dB$, $T = 10$ and $K = 0$).

estimates for each coherence block. Therefore, in the presence of reduced channel variations due to a larger Rician factor K, the performance gain of the adaptive joint approach are reduced.

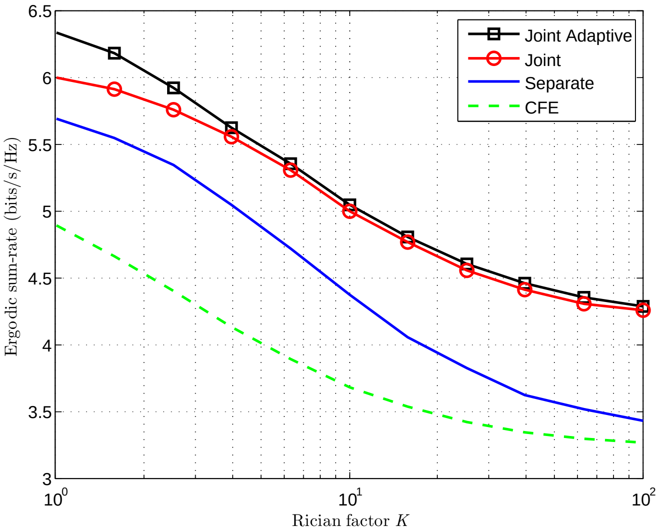

Figure 1.9 Ergodic achievable sum-rate vs. Rician factor K ($N_R = N_M = 2$, $N_t = N_r = 4$, $\bar{C} = 6$ bits/s/Hz, $P = 20dB$, $T = 20$ and $\alpha_{ji} = 1$).

1.4 Downlink: Where to Perform Channel Encoding and Precoding?

In this section, we turn to the downlink and address the question of whether it may be more advantangeous to implement channel encoding and precoding at the RRHs rather than at the BBU as in the conventional implementation. Specifically, we compare the following two approaches: *i Conventional approach* the BBU performs channel coding and precoding and then the BBU compresses and forwards the resulting baseband signals on the fronthaul links to the RRHs; *ii Channel encoding and precoding at the RRHs*: The BBU does not perform precoding but rather forwards separately the information messages of a subset of MSs along with the compressed precoding matrices to the each RRH, which then performs precoding. Under a simplified quasi-static channel, rather than ergodic, channel model, the conventional approach was studied in [25, 26, 27, 28], while the alternative functional split *ii* was investigated in [**?**]. This section is adapted from the reference [**?**]. We start by detailing the system model in Sec. 1.4.1. In Section 1.4.2, we study the conventional approach, while the alternative functional split mentioned above is is studied in 1.4.3. In Section 1.4.4, numerical results are presented to provide insight into the comparison under study.

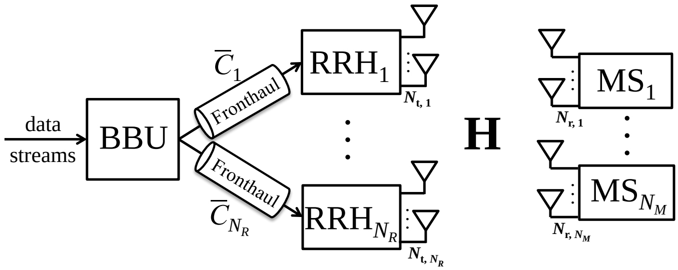

Figure 1.10 Downlink of a C-RAN system in which a single cluster of RRHs is connected to a BBU via finite-capacity fronthaul links. The downlink channel matrix $\mathbf{H}$ varies in an ergodic fashion along the channel coherence blocks and its instantaneous realization is unknown to the BBU and the RRHs.

1.4.1 System Model

We consider the downlink of a C-RAN in which a cluster of N_R RRHs provides wireless service to N_M MSs as illustrated in Fig. 1.10. Most of the baseband processing for all the RRHs in the cluster is carried out at a BBU that is connected to each i-th RRH via a fronthaul link of finite capacity, as further discussed below. Each i-th RRH has $N_{t,i}$ transmit antennas and each j-th MS has $N_{r,j}$ receive antennas. We denote the set of all RRHs 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 RRH 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

$$\mathbf{Y}_j = \mathbf{H}_j \mathbf{X} + \mathbf{Z}_j, \tag{1.29}$$

where $\mathbf{Z}_j$ is the $N_{r,j} \times T$ noise matrix, which consist of i.i.d. $\mathcal{CN}(0, 1)$ entries; $\mathbf{H}_j = [\mathbf{H}_{j1}, \ldots, \mathbf{H}_{jN_R}]$ 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 RRH to the j-th MS; and $\mathbf{X}$ is the collection of the signals transmitted by all the RRHs, i.e., $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_{N_R}^T]^T$. 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 both the scenarios in which the BBU has either perfect instantaneous information about the channel matrix $\mathbf{H}$ or 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 $E[||\mathbf{X}_i||^2]/T \leq \bar{P}_i$.

Remark 6 *A specific channel model of interest is the standard Kronecker model, whereby the channel matrix $\mathbf{H}_{ji}$ is written as*

$$\mathbf{H}_{ji} = \mathbf{\Sigma}_{R,ji}^{1/2}\widetilde{\mathbf{H}}_{ji}\mathbf{\Sigma}_{T,ji}^{1/2}, \tag{1.30}$$

where the $N_{t,i} \times N_{t,i}$ matrix $\mathbf{\Sigma}_{T,ji}$ and the $N_{r,j} \times N_{r,j}$ matrix $\mathbf{\Sigma}_{R,ji}$ are the transmit-side and receiver-side spatial correlation matrices, respectively, 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 [24]. With this model, stochastic CSI entails that the BBU is hence only aware of the correlation matrices $\mathbf{\Sigma}_{T,ji}$ and $\mathbf{\Sigma}_{R,ji}$. Moreover, in case that the RRHs are placed in a higher location than the MSs, one can assume that the receive-side fading is uncorrelated, i.e., $\mathbf{\Sigma}_{R,ji} = \mathbf{I}_{N_{r,j}}$, while the transmit-side covariance matrix $\mathbf{\Sigma}_{T,ji}$ is determined by the one-ring scattering model (see [24] and references therein). In particular, if the RRHs are equipped with $\lambda/2$-spaced uniform linear arrays, we have $\mathbf{\Sigma}_{T,ji} = \mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})$ for the j-th MS and the i-th RRH located at a relative angle of arrival θ_{ji} and having angular spread Δ_{ji}, where the element (m,n) of matrix $\mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})$ is given by

$$[\mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})]_{m,n} = \frac{\alpha_{ji}}{2\Delta_{ji}} \int_{\theta_{ji}-\Delta_{ji}}^{\theta_{ji}+\Delta_{ji}} \exp^{-j\pi(m-n)\sin(\phi)} \, d\phi, \tag{1.31}$$

with the path loss coefficient α_{ji} between the j-th MS and the i-th RRH being given as

$$\alpha_{ji} = \frac{1}{1 + \left(\frac{d_{ji}}{d_0}\right)^{\eta}}, \tag{1.32}$$

where d_{ji} is the distance between the j-th MS and the i-th RRH, d_0 is a reference distance, and η is the path loss exponent.

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. In other words, the capacity of the i-th fronthaul link is $\bar{C}_i$ bits per channel use of the downlink. The fronthaul capacity constraint limits the fronthaul rate that is allocated in the coding block, and hence across all the fading states, to be no larger than $\bar{C}_i$. The fronthaul constraint will be further discussed in Section 1.4.2 and 1.4.3.

1.4.2 Conventional Approach

In this section, we first describe the CAP strategy in Section 1.4.2.1. Then, we briefly review known strategies for the joint optimization of fronthaul compression and pre-

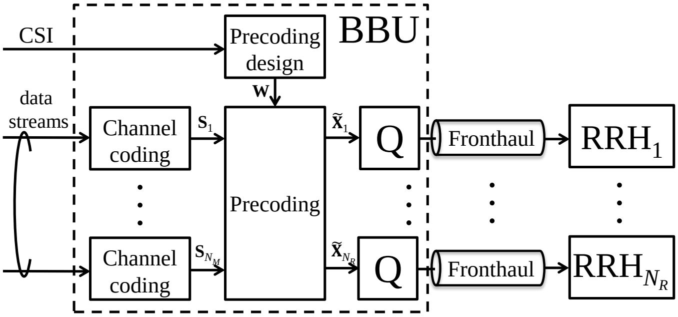

Figure 1.11 Block diagram of the Compression-After-Precoding (CAP) scheme ("Q" represents fronthaul compression).

coding with perfect instantaneous channel knowledge at the BBU in Section 1.4.2.2. Finally, we propose an optimization algorithm under the assumption of stochastic CSI at the BBU in Section 1.4.2.3.

1.4.2.1 Problem Formulation

With the CAP scheme as illustrated in Fig. 1.11, the BBU performs channel coding and precoding, and then compresses the resulting baseband signals so that they can be forwarded on the fronthaul links to the corresponding RRHs. This strategy corresponds to the standard approach envisioned for C-RANs [25, 26, 27, 28]. 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.

The precoded data signal computed by the BBU for any given coherence time can be written as $\widetilde{\mathbf{X}} = \mathbf{WS}$, 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 RRH 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 RRH 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 RRH, 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.

The BBU quantizes each sequence of baseband signal $\widetilde{\mathbf{X}}_i$ for transmission on the

i-th fronthaul link to the i-th RRH. We write the compressed signals $\mathbf{X}_i$ for i-th RRH as

$$\mathbf{X}_i = \widetilde{\mathbf{X}}_i + \mathbf{Q}_{x,i}, \tag{1.33}$$

where the quantization noise matrix $\mathbf{Q}_{x,i}$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma^2_{x,i})$ entries. The quantization noises $\mathbf{Q}_{x,i}$ are independent across the RRH index i, which can be realized via separate quantizers for the signals of different RRHs. Note that the possibility to leverage quantization noise correlation across the RRHs via joint quantization is explored in [28] for static channels. Based on (1.33), the design of the fronthaul compression reduces to the optimization of the quantization noise variances $\sigma^2_{x,1}, \ldots, \sigma^2_{x,N_R}$. The power transmitted by i-th RRH is then computed as

$$P_i\left(\mathbf{W}, \sigma^2_{x,i}\right) = \frac{1}{T}E[||\mathbf{X}_i||^2] = \text{tr}\left(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r + \sigma^2_{x,i}\mathbf{I}\right), \tag{1.34}$$

where we have emphasized the dependence of the power $P_i(\mathbf{W}, \sigma^2_{x,i})$ on the precoding matrix $\mathbf{W}$ and quantization noise variances $\sigma^2_{x,i}$. Moreover, using standard rate-distortion arguments, the rate required on the fronthaul between the BBU and i-th RRH in a given coherence interval can be quantified by $I(\widetilde{\mathbf{X}}_i; \mathbf{X}_i)/T$ (see, e.g., [8, Ch. 3]). Therefore, the rate allocated on the i-th fronthaul link is equal to

$$C_i\left(\mathbf{W}, \sigma^2_{x,i}\right) = \log\det\left(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r + \sigma^2_{x,i}\mathbf{I}\right) - N_{t,i}\log\left(\sigma^2_{x,i}\right), \tag{1.35}$$

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

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 $E[R_j^{CAP}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}^2_x)]$, with $R_j^{CAP}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}^2_x) = I_{\mathbf{H}}(\mathbf{S}_j; \mathbf{Y}_j)/T$, 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

$$\begin{aligned} R_j^{CAP}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}^2_x) = & \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\mathbf{W}\mathbf{W}^\dagger + \boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right) \\ & - \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\sum_{k\in\mathcal{N}_M\backslash j}\mathbf{W}_k^c\mathbf{W}_k^{c\dagger} + \boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right), \end{aligned} \tag{1.36}$$

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

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

1.4.2.2 **Perfect Instantaneous CSI**

In the case of perfect channel knowledge at the BBU, the design of the precoding matrix $\mathbf{W}$ and the compression noise variances $\boldsymbol{\sigma}_x^2$, is adapted to the channel realization $\mathbf{H}$ for each coherence block. To emphasize this fact, we use the notation $\mathbf{W}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$. The problem of optimizing the ergodic weighted achievable sum-rate with given weights $\mu_j \geq 0$ for $j \in \mathcal{N}_M$ is then formulated as follows:

$$\underset{\mathbf{W}(\mathbf{H}),\boldsymbol{\sigma}_x^2(\mathbf{H})}{\text{maximize}} \quad \sum_{j\in\mathcal{N}_M} \mu_j E\left[R_j^{CAP}\left(\mathbf{H},\mathbf{W}(\mathbf{H}),\boldsymbol{\sigma}_x^2(\mathbf{H})\right)\right] \tag{1.37a}$$

$$\text{s.t.} \quad C_i\left(\mathbf{W},\sigma_{x,i}^2(\mathbf{H})\right) \leq \bar{C}_i, \tag{1.37b}$$

$$P_i\left(\mathbf{W}(\mathbf{H}),\sigma_{x,i}^2(\mathbf{H})\right) \leq \bar{P}_i, \tag{1.37c}$$

where (1.37b)-(1.37c) apply for all $i \in \mathcal{N}_R$ and all channel realizations $\mathbf{H}$. Due to the separability of the fronthaul and power constraints across the channel realizations $\mathbf{H}$, the problem (1.37) can be solved for each $\mathbf{H}$ independently. Note that the achievable rate in (1.37a) and the fronthaul constraint in (1.37b) are non-convex. However, the functions $R_j^{CAP}(\mathbf{H},\mathbf{W}(\mathbf{H}),\boldsymbol{\sigma}_x^2(\mathbf{H}))$ and $C_i(\mathbf{W}(\mathbf{H}),\sigma_{x,i}^2(\mathbf{H}))$ can be then seen to be difference of convex (DC) functions of the covariance matrices $\widetilde{\mathbf{V}}_j(\mathbf{H}) = \widetilde{\mathbf{W}}_j^c(\mathbf{H})\widetilde{\mathbf{W}}_j^{c\dagger}(\mathbf{H})$ for all $j \in \mathcal{N}_M$ and the variance $\boldsymbol{\sigma}_x^2(\mathbf{H})$. The resulting relaxed problem can be tackled via the Majorization-Minimization (MM) algorithm as detailed in [28], from which a feasible solution of problem (1.37) can be obtained. We refer to [28] for details.

1.4.2.3 **Stochastic CSI**

With only stochastic CSI at the BBU, in contrast to the case with instantaneous CSI, 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:

$$\underset{\mathbf{W},\boldsymbol{\sigma}_x^2}{\text{maximize}} \quad \sum_{j\in\mathcal{N}_M} \mu_j E\left[R_j^{CAP}\left(\mathbf{H},\mathbf{W},\boldsymbol{\sigma}_x^2\right)\right] \tag{1.38a}$$

$$\text{s.t.} \quad C_i\left(\mathbf{W},\sigma_{x,i}^2\right) \leq \bar{C}_i, \tag{1.38b}$$

$$P_i\left(\mathbf{W},\sigma_{x,i}^2\right) \leq \bar{P}_i, \tag{1.38c}$$

where (1.38b)-(1.38c) apply to all $i \in \mathcal{N}_R$. In order to tackle this problem, we adopt the Stochastic Successive Upper-bound Minimization (SSUM) method [29], whereby, at each step, a stochastic lower bound of the objective function is maximized around the current iterate[9]. To this end, similar to [28], 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 (1.38b) is also a DC function, with respect to the covariance $\mathbf{V} = [\mathbf{V}_1 \ldots \mathbf{V}_{N_M}]$ and

[9] We mention here that an alternative method to attack the problem would be the strategy introduced in [30]. We leave the study of this approach to future work.

Table 1.1 CAP Design of Fronthaul Compression and Precoding

```
**Initialization (outer loop)**: Initialize the covariance matrices $\mathbf{V}^{(0)}$ and the quantization noise variances $\boldsymbol{\sigma}_x^{2\ (0)}$, and set $n = 0$.
**repeat**
    $n \leftarrow n + 1$
    Generate a channel matrix realization $\mathbf{H}^{(n)}$ using the available stochastic CSI.
    **Initialization (inner loop)**: Initialize $\mathbf{V}^{(n,0)} = \mathbf{V}^{(n-1)}$ and $\boldsymbol{\sigma}_x^{2\ (n,0)} = \boldsymbol{\sigma}_x^{2\ (n-1)}$, and set $r = 0$.
    **repeat**
        $r \leftarrow r + 1$
        $\max_{\mathbf{V},\boldsymbol{\sigma}_x^2} \quad \frac{1}{n}\sum_{l=1}^{n}\sum_{j\in\mathcal{N}_M} \mu_j \widetilde{R}_j^{CAP}\left(\mathbf{H}^{(l)}, \mathbf{V}, \boldsymbol{\sigma}_x^2 | \mathbf{V}^{(l-1)}, \boldsymbol{\sigma}_x^{2\ (l-1)}\right)$
        $\text{s.t.} \quad \widetilde{C}_i\left(\mathbf{V}, \sigma_{x,i}^2 | \mathbf{V}^{(n,r-1)}, \sigma_{x,i}^{2\ (n,r-1)}\right) \leq \bar{C}_i,$    (1.39)
        $P_i\left(\mathbf{V}, \sigma_{x,i}^2\right) \leq \bar{P}_i, \quad \text{for all } i \in \mathcal{N}_R.$
        Update $\mathbf{V}^{(n,r)} \leftarrow \mathbf{V}$ and $\boldsymbol{\sigma}_x^{2\ (n,r)} \leftarrow \boldsymbol{\sigma}_x^2$.
    **until** a convergence criterion is satisfied.
    Update $\mathbf{V}^{(n)} \leftarrow \mathbf{V}^{(n,r)}$ and $\boldsymbol{\sigma}_x^{2\ (n)} \leftarrow \boldsymbol{\sigma}_x^{2\ (n,r)}$.
**until** a convergence criterion is satisfied.
**Solution**: Calculate the precoding matrix $\mathbf{W}$ from the covariance matrices $\mathbf{V}^{(n)}$ via rank reduction as $\mathbf{W}_j = \gamma_j \nu_{\max}^{(M_j)}(\mathbf{V}_j^{(n)})$ for all $j \in \mathcal{N}_M$, where $\gamma_j$ is obtained by imposing $P_i\left(\mathbf{W}, \sigma_{x,i}^2\right) = \bar{P}_i$ using (1.34).
```

the quantization noise variances $\boldsymbol{\sigma}_x^2$. As discussed above, the resulting problem is a rank-relaxation of the original problem (1.38). Due to the DC structure, locally tight (stochastic) convex lower bounds can be calculated for objective function (1.38a) and the constraint (1.38b) (see, e.g., [31]).

The proposed algorithm based on SSUM [29] 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 BBU. For example, with the model (1.30), the channel matrices are generated based on the knowledge of the spatial correlation matrices. Following the SSUM scheme, the outer loop aims at maximizing a stochastic lower bound on the objective function, given as

$$\frac{1}{n}\sum_{l=1}^{n} \widetilde{R}_j^{CAP}\left(\mathbf{H}^{(l)}, \mathbf{V}, \boldsymbol{\sigma}_x^2 | \mathbf{V}^{(l-1)}, \boldsymbol{\sigma}_x^{2\ (l-1)}\right), \tag{1.40}$$

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 $R_j^{CAP}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}_x^2)$ 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, e.g., [29])

$$\widetilde{R}_j^{CAP}\left(\mathbf{H}^{(l)}, \mathbf{V}, \boldsymbol{\sigma}_x^2 | \mathbf{V}^{(l-1)}, \boldsymbol{\sigma}_x^{2\ (l-1)}\right) \triangleq \log\det\left(\mathbf{I} + \mathbf{H}_j^{(l)}\left(\sum_{k=1}^{N_M}\mathbf{V}_k + \boldsymbol{\Omega}_x\right)\mathbf{H}_j^{(l)\ \dagger}\right)$$
$$-f\left(\mathbf{I} + \mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j^{(l-1)}\mathbf{H}_j^{(l)\ \dagger}, \mathbf{I} + \mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j\mathbf{H}_j^{(l)\ \dagger}\right), \quad (1.41)$$

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

$$f(\mathbf{A}, \mathbf{B}) \triangleq \log\det(\mathbf{A}) + \frac{1}{\ln 2}\text{tr}\left(\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\right). \quad (1.42)$$

Since the maximization of (1.40) is subject to the non-convex DC constraint (1.38b), 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 (1.38b) [32]. 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 (1.38b) at the r-th inner iteration can be locally approximated as

$$\widetilde{C}_i\left(\mathbf{V}, \sigma_{x,i}^2 | \mathbf{V}^{(n,r-1)}, \sigma_{x,i}^{2\ (n,r-1)}\right) \triangleq \quad (1.43)$$
$$f\left(\sum_{k=1}^{N_M}\mathbf{D}_i^{rT}\mathbf{V}_k^{(n,r-1)}\mathbf{D}_i^r + \sigma_{x,i}^{2\ (n,r-1)}\mathbf{I}, \sum_{k=1}^{N_M}\mathbf{D}_i^{rT}\mathbf{V}_k\mathbf{D}_i^r + \sigma_{x,i}^2\mathbf{I}\right) - N_{t,i}\log(\sigma_{x,i}^2).$$

The resulting combination of SSUM and MM algorithms for the solution of problem (1.38) is summarized in Table Algorithm 1.1. The algorithm is completed by calculating, from the obtained solution $\mathbf{V}^*$ of the relaxed problem, the precoding matrix $\mathbf{W}$ by using the standard rank-reduction approach [33], which is given as $\mathbf{W}_j^* = \gamma_j\nu_{\text{max}}^{(M_j)}(\mathbf{V}_j^*)$ with the normalization factor γ_j, selected so as to satisfy the power constraint with equality, namely $P_i\left(\mathbf{W}, \sigma_{x,i}^2\right) = \bar{P}_i$.

Two remarks are in place on the properties of the proposed algorithm. First, since the approximated functions (1.41) and (1.43) are local lower bounds, the algorithm provides a feasible solution of the relaxed problem at each inner and outer iteration (see, e.g., [29]). The second remark is that, from [29, 31], 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.

1.4.3 Compression-Before-Precoding

With the Compression-Before-Precoding (CBP) scheme, the BBU calculates the precoding matrices, but does not perform precoding. Instead, as illustrated in Fig. 1.12,

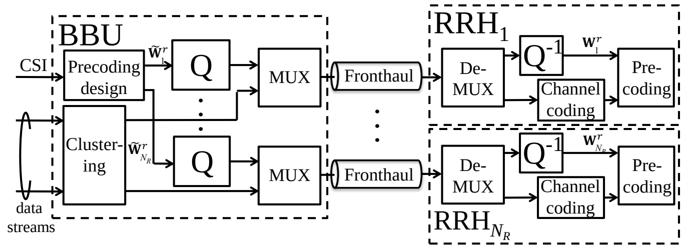

Figure 1.12 Block diagram of the Compression-Before-Precoding (CBP) scheme (“Q" and “Q^{-1}" represents fronthaul compression and decompression, respectively).

it uses the fronthaul links to communicate the information messages of a given subset of MSs to each RRH, along with the corresponding compressed precoding matrices. Each RRH 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 RRHs. In the following, we first describe the CBP strategy in Section 1.4.3.1; then we review the design problem under instantaneous CSI in Section 1.4.3.2; and, finally, we introduce an algorithm for the joint optimization of fronthaul compression and precoding with stochastic CSI at the BBU.

1.4.3.1 **Precoding and Fronthaul Compression for CBP**

As shown in Fig. 1.12, in the CBP method, the precoding matrix $\widetilde{\mathbf{W}}$ and the information streams are separately transmitted from the BBU to the RRHs, and the received information bits are encoded and precoded at each RRH 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 RRH, since the RRHs should receive all the corresponding information streams.

Given the above, with the CBP strategy, we allow for a preliminary clustering step at the BBU whereby each RRH is assigned by a subset of the MSs. We denote the set of MSs assigned by i-th RRH as $\mathcal{M}_i \subseteq \mathcal{N}_M$ for all $i \in \mathcal{N}_R$. This implies that i-th RRH only needs the information streams intended for the MSs in the set $\mathcal{M}_i$. We also denote the set of RRHs that serve the j-th MS, as $\mathcal{B}_j = \{i|j \in \mathcal{M}_i\} \subseteq \mathcal{N}_R$ 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 RRH in the sets $\mathcal{M}_i$ and $\mathcal{B}_j$, respectively. We define the number of all transmit antennas for the RRHs, which serve the j-th MS, as $N_{t,\mathcal{B}_j}$. We assume here that the sets of MSs assigned by i-th RRH are given and not subject to optimization (see Section 1.4.4 for further details).

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

$$\widetilde{\mathbf{W}} = \left[\mathbf{E}_1^c \widetilde{\mathbf{W}}_1^c, \ldots, \mathbf{E}_{N_M}^c \widetilde{\mathbf{W}}_{N_M}^c\right], \tag{1.44}$$

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 RRHs 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 RRHs 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.

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 RRH for all coherence times in the coding block is compressed by the BBU and forwarded over the fronthaul link to the i-th RRH. The compressed precoding matrix $\mathbf{W}_i^r$ for i-th RRH is given by

$$\mathbf{W}_i^r = \widetilde{\mathbf{W}}_i^r + \mathbf{Q}_{w,i}, \tag{1.45}$$

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 RRHs is represented as

$$\mathbf{W} = \widetilde{\mathbf{W}} + \mathbf{Q}_w, \tag{1.46}$$

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

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

$$\begin{aligned} R_j^{CBP}\Big(\mathbf{H}, \widetilde{\mathbf{W}}, \boldsymbol{\sigma}^2_w\Big) = \frac{1}{T} I_{\mathbf{H}}\left(\mathbf{S}_j; \mathbf{Y}_j\right) = \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\dagger + \mathbf{\Omega}_w\right)\mathbf{H}_j^\dagger\right) \\ - \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\sum_{k\in\mathcal{N}_M\setminus j} \widetilde{\mathbf{W}}_k^c\widetilde{\mathbf{W}}_k^{c\dagger} + \mathbf{\Omega}_w\right)\mathbf{H}_j^\dagger\right). \end{aligned} \tag{1.47}$$

1.4.3.2 Perfect Instantaneous CSI

With perfect CSI at the BBU, as discussed in Section 1.4.2.2, one can adopt the precoding matrix $\widetilde{\mathbf{W}}(\mathbf{H})$, the user rates $\{R_j(\mathbf{H})\}$ and the quantization noise variances $\boldsymbol{\sigma}^2_w(\mathbf{H})$ to the current channel realization at each coherence block. The rate required to transmit precoding information on the i-th fronthaul in a given channel realizations $\mathbf{H}$ is given by $C_i(\mathbf{H}, \widetilde{\mathbf{W}}_i^r, \sigma^2_{w,i})/T$, with

$$\begin{aligned} \frac{1}{T}C_i\left(\mathbf{H}, \widetilde{\mathbf{W}}_i^r, \sigma^2_{w,i}\right) &= \frac{1}{T} I_{\mathbf{H}}(\widetilde{\mathbf{W}}_i^r; \mathbf{W}_i^r) \\ &= \frac{1}{T}\Big\{\log\det\Big(\mathbf{D}_i^{rT}\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\dagger\mathbf{D}_i^r + \sigma^2_{w,i}\mathbf{I}\Big) - N_{t,i}\log\big(\sigma^2_{w,i}\big)\Big\}, \end{aligned} \tag{1.48}$$

where the rate $C_i(\widetilde{\mathbf{W}}_i^r, \sigma^2_{w,i})$ required on i-fronthaul link is defined in (1.35). Note that the normalization by T is needed since only a single precoding matrix is needed for each channel coherence interval. Then, under the fronthaul capacity constraint, the remaining fronthaul capacity that can be used to convey precoding information

Table 1.2 CBP Design of Fronthaul Compression and Precoding

```
**Initialization**: Initialize the covariance matrices $\widetilde{\mathbf{V}}^{(0)}$ and the user rate $\{R_j^{(0)}\}$ and set $n = 0$.
**repeat**
    $n \leftarrow n + 1$
    Generate a channel matrix realization $\mathbf{H}^{(n)}$ using the available stochastic CSI.
        $\max_{\widetilde{\mathbf{V}}, \{R_j\}} \sum_{j \in \mathcal{N}_M} \mu_j R_j$    (1.50)
        $R_j \leq \frac{1}{n} \sum_{l=1}^{n} \widetilde{R}_j^{CBP} \left( \mathbf{H}^{(l)}, \widetilde{\mathbf{V}} | \widetilde{\mathbf{V}}^{(l-1)} \right),$
        $\sum_{j \in \mathcal{M}_i} R_j \leq \bar{C}_i,$
        $P_i \left( \widetilde{\mathbf{V}}, 0 \right) \leq \bar{P}_i,$    for all $i \in \mathcal{N}_R$ and $j \in \mathcal{N}_M$.
    Update $\widetilde{\mathbf{V}}^{(n)} \leftarrow \widetilde{\mathbf{V}}$ and $\{R_j^{(n)}\} \leftarrow \{R_j\}$.
**until** a convergence criterion is satisfied.
**Solution**: Calculate the precoding matrix $\widetilde{\mathbf{W}}$ from the covariance matrices $\widetilde{\mathbf{V}}^{(n)}$ via rank reduction as $\widetilde{\mathbf{W}}_j = \gamma_j \nu_{\text{max}}^{(M_j)}(\widetilde{\mathbf{V}}_j^{(n)})$ for all $j \in \mathcal{N}_M$, where $\gamma_j$ is obtained by imposing $P_i \left( \widetilde{\mathbf{W}}, \sigma_{w,i}^2 \right) = \bar{P}_i$ using (1.34).
```

corresponding to the i-th RRH is $\bar{C}_i - \sum_{j \in \mathcal{M}_i} R_j$. As a result, the optimization problem of interest can be formulated as

$$\underset{\widetilde{\mathbf{W}}(\mathbf{H}), \boldsymbol{\sigma}_{w,i}^2(\mathbf{H}), \{R_j(\mathbf{H})\}}{\text{maximize}} \quad \sum_{j \in \mathcal{N}_M} \mu_j R_j(\mathbf{H}) \tag{1.49a}$$

$$s.t. \quad R_j(\mathbf{H}) \leq R_j^{CBP} \left(\mathbf{H}, \widetilde{\mathbf{W}}(\mathbf{H}), \boldsymbol{\sigma}_w^2(\mathbf{H}) \right), \tag{1.49b}$$

$$\frac{1}{T} C_i \left(\mathbf{H}, \widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}) \right) \leq \bar{C}_i - \sum_{j \in \mathcal{M}_i} R_j(\mathbf{H}), \tag{1.49c}$$

$$P_i \left(\widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}) \right) \leq \bar{P}_i, \tag{1.49d}$$

where the constraints apply to all channel realization, (1.49b) applies to all $j \in \mathcal{N}_M$, (1.49c) - (1.49d) apply to all $i \in \mathcal{N}_R$ and the transmit power $P_i(\widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}))$ at i-th RRH is defined in (1.34). Similar to Section 1.4.2.2, the problem (1.49) can be studied for each $\mathbf{H}$ independently. In addition, each subproblem can be tackled by using MM algorithm as explained in [28].

1.4.3.3 **Stochastic CSI**

With stochastic CSI at the BBU, 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 RRH becomes negligible. As a result, we can neglect the effect of the quantization noise and 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

$$\underset{\widetilde{\mathbf{W}},\{R_j\}}{\text{maximize}} \quad \sum_{j\in\mathcal{N}_M} \mu_j R_j \tag{1.51a}$$

$$s.t. \quad R_j \leq E\left[R_j^{CBP}\left(\mathbf{H},\widetilde{\mathbf{W}},\mathbf{0}\right)\right], \tag{1.51b}$$

$$\sum_{j\in\mathcal{M}_i} R_j \leq \bar{C}_i, \tag{1.51c}$$

$$P_i\left(\widetilde{\mathbf{W}}_i^r, 0\right) \leq \bar{P}_i, \tag{1.51d}$$

where (1.51b) applies to all $j \in \mathcal{N}_M$, (1.51c)-(1.51d) 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 RRH is defined in (1.34). In problem (1.51), the constraint (1.51b) is not only non-convex but also stochastic. Similar to Section 1.4.2.3, 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. Referring to Section 1.4.2.3, for details, given the solutions $\widetilde{\mathbf{V}}^{(l-1)}$ at the previous iterations, $l \leq n$, the algorithm approximates the function $E[R_j^{CBP}(\mathbf{H},\widetilde{\mathbf{W}})]$ in (1.51b) with the stochastic upper bound as

$$\frac{1}{n}\sum_{l=1}^{n} \widetilde{R}_j^{CBP}\left(\mathbf{H}^{(l)}, \widetilde{\mathbf{V}}|\widetilde{\mathbf{V}}^{(l-1)}\right) \tag{1.52}$$

with

$$\widetilde{R}_j^{CBP}\left(\widetilde{\mathbf{V}}|\mathbf{H}^{(l)}, \widetilde{\mathbf{V}}^{(l-1)}\right) \triangleq \log\det\left(\mathbf{I} + \mathbf{H}_j^{(l)}\left(\sum_{k=1}^{N_M}\widetilde{\mathbf{V}}_k\right)\mathbf{H}_j^{\dagger\ (l)}\right) \tag{1.53}$$

$$-f\left(\mathbf{I} + \mathbf{H}_j^{(l)}\left(\sum_{k=1,k\neq j}^{N_M}\widetilde{\mathbf{V}}_k^{(l-1)}\right)\mathbf{H}_j^{\dagger\ (l)}, \mathbf{I} + \mathbf{H}_j^{(l)}\left(\sum_{k=1,k\neq j}^{N_M}\widetilde{\mathbf{V}}_k\right)\mathbf{H}_j^{\dagger\ (l)}\right),$$

where the linearization function $f(\mathbf{A},\mathbf{B})$ is defined in (1.42). The algorithm which is summarized in Table Algorithm 1.2, has the same properties discussed for the algorithm in Table Algorithm 1.1, namely it provides a feasible solution of the relaxed problem at each iteration and it converge to a stationary point of the same problem.

1.4.4 **Numerical Results**

In this section, we compare the performance of the CAP and CBP schemes in the setup under study of block-ergodic channels. To this end, we consider a system in which

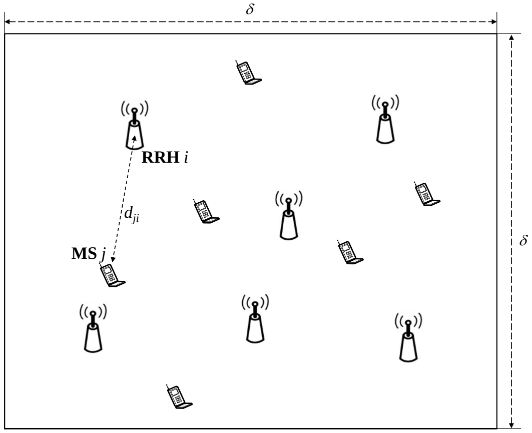

Figure 1.13 Set-up under consideration for the numerical results in Section 1.4.4, where the RRHs are randomly located in a square with side δ and all MSs and RRHS are randomly uniformly placed.

the RRHs and the MSs are randomly located in a square area with side $\delta = 500m$ as in Fig. 1.13. In the path loss formula (1.32), we set the reference distance to $d_0 = 50m$ and the path loss exponent to $\eta = 3$. We adopt the spatial correlation model in (1.31) with the angular spread $\Delta_{ji} = \arctan(r_s/d_{ji})$, with the scattering radius $r_s = 10m$ and with d_{ji} being the Euclidean distance between the i-th RRH and the j-th MS. Throughout, we assume that the every RRH 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-RRH assignment is carried out by choosing, for each RRH, 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. Note also that a given MS is generally assigned to multiple RRHs.

The effect of the fronthaul capacity limitation on the ergodic achievable sum-rate is investigated in Fig. 1.14, where the number of RRHs and MSs is $N_R = N_M = 4$, the number of transmit antennas is $N_{t,i} = 2$ for all $i \in \mathcal{N}_R$, the number of receive antennas is $N_{r,j} = 1$ for all $j \in \mathcal{N}_M$, the power is $\bar{P} = 10dB$, and the coherence time is $T = 20$. We first observe that, with instantaneous CSI, the CAP strategy is uniformly better than CBP as long as the fronthaul capacity is sufficiently large (here $\bar{C} > 2$). This is due to the enhanced interference mitigation capabilities of CAP resulting from its ability to coordinate all the RRHs via joint baseband processing without requiring the transmission of all messages on all fronthaul links. Note, in fact, that, with CBP, only N_c MSs are served by each RRH, and that making N_c larger entails a significant increase in the fronthaul capacity requirements. We will later see that this advantage of CAP is offset by the higher fronthaul efficiency of CBP in transmitting precoding information for large coherence periods T (see Fig.

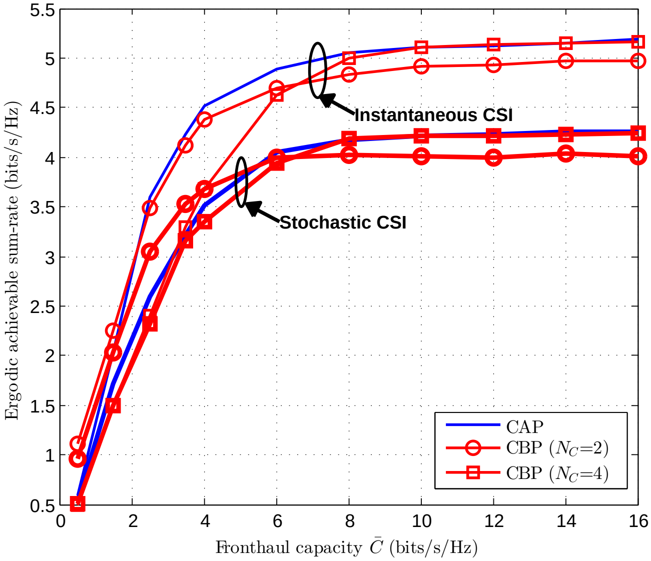

Figure 1.14 Ergodic achievable sum-rate vs. the fronthaul capacity $\bar{C}$ ($N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{P} = 10$ dB, $T = 20$, and $\mu = 1$).

1.16). Instead, with stochastic CSI, in the low fronthaul capacity regime, here about $\bar{C} < 6$, the CBP strategy is generally advantageous due to the additional advantage that is accrued by amortizing the precoding overhead over the entire coding block. Another observation is that, for small $\bar{C}$, the CBP schemes with progressively smaller N_c have better performance thanks to the reduced fronthaul overhead. Moreover, for large $\bar{C}$, the performance of the CBP scheme with $N_c = N_M$, whereby each RRH serves all MSs, approaches that of the CAP scheme.

The effect of the power constraint $\bar{P}$ is investigated in Fig. 1.15, where the number of RRHs and MSs is $N_R = N_M = 4$, the number of transmit antennas is $N_{t,i} = 2$, the number of receive antennas is $N_{r,j} = 1$, the fronthaul capacity is $\bar{C} = 6$ bits/s/Hz, and the coherence time is $T = 15$. As a general rule, increasing $\bar{P}$ enhances the relative impact of the quantization noise on the performance. This can be seen from, e.g., (1.35), from which it follows that the quantization noise variance increases with the power $\bar{P}$ for a fixed value of the fronthaul capacity $\bar{C}$. The CAP approach is seen to be advantageous in the low power regime, in which the RRH coordination gains are not offset by the effect of the quantization noise. In contrast, the CBP method is to be preferred in the larger power regime due to the limited impact of the quantization noise on its performance since only precoding information is quantized.

Fig. 1.16 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. As anticipated, with instantaneous CSI, CBP is seen to benefit from a larger coherence time T, since the fronthaul overhead required to transmit precoding information gets

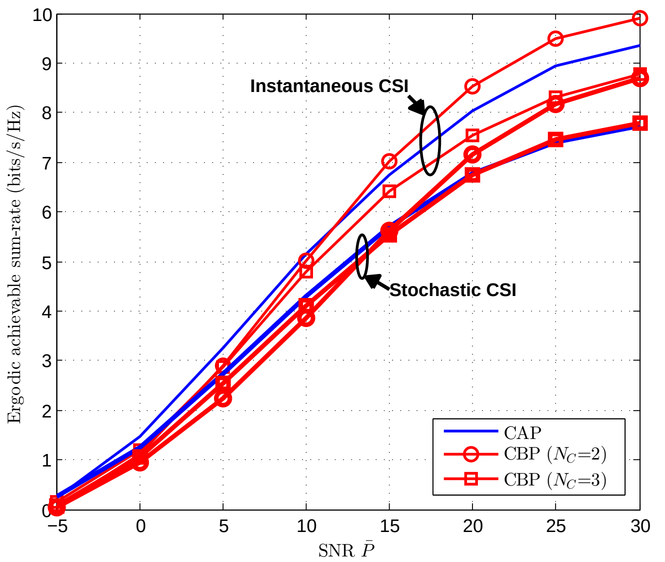

Figure 1.15 Ergodic achievable sum-rate vs. the power constraint $\bar{P}$ ($N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 6$ bits/s/Hz, $T = 15$, and $\mu = 1$).

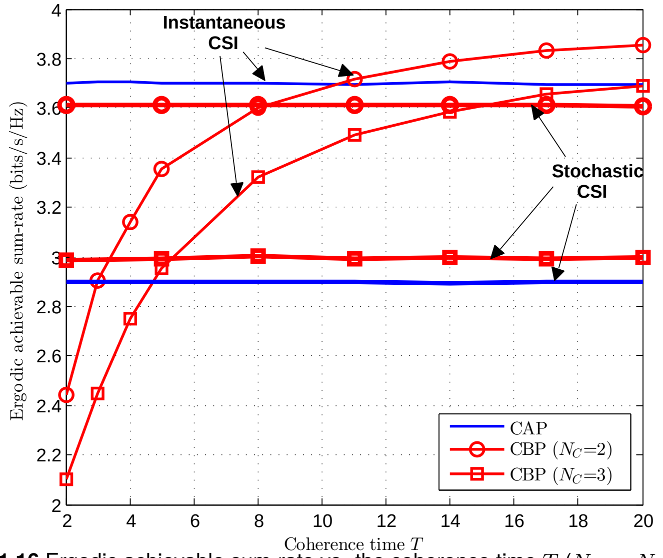

Figure 1.16 Ergodic achievable sum-rate vs. the coherence time T ($N_R = N_M = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 2$ bits/s/Hz, $\bar{P} = 20dB$, and $\mu = 1$).

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 scheme is not affected by the coherence time. As a result, CBP can outperform CAP for sufficiently

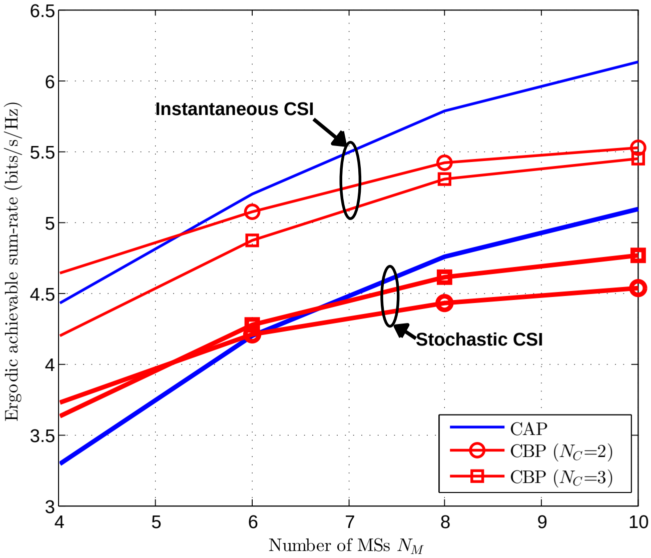

Figure 1.17 Ergodic achievable sum-rate vs. the number of MSs N_M ($N_R = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 4$ bits/s/Hz, $\bar{P} = 10$ dB, $T = 10$, and $\mu = 1$).

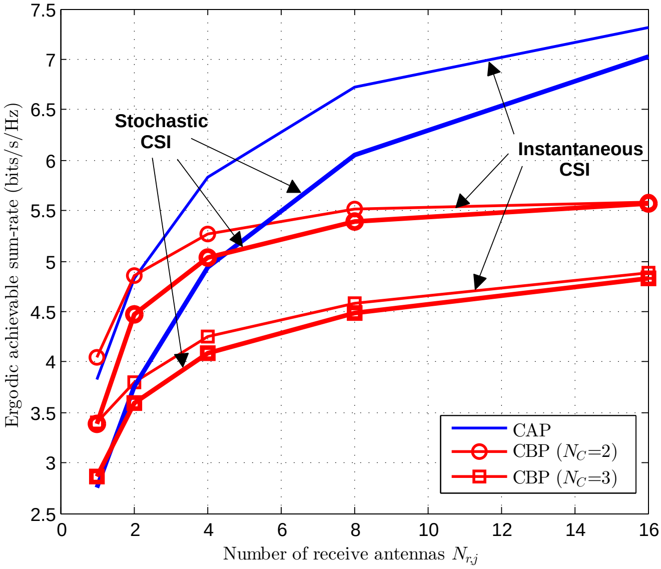

Figure 1.18 Ergodic achievable sum-rate vs. the number of recevie antennas N_r ($N_R = N_M = 4$, $N_{t,i} = 2$, $\bar{C} = 3$ bits/s/Hz, $\bar{P} = 10$ dB, $T = 10$, and $\mu = 1$).

large T in the presence of instantaneous CSI. Instead, with stochastic CSI, given the large SNR, as discussed around Fig. 1.15, CBP is to be preferred.

In Fig. 1.17, 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. 1.18, 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 RRHs 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.

1.4.5 Conclusion

In this chapter, 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 both instantaneous and stochastic CSI. The study compares the Compress-After-Precoding (CAP) and the Compress-Before-Precoding (CBP) approaches, which differ in their fronthaul compression requirements and interference mitigation capabilities. Efficient algorithms have been proposed for the maximization of the ergodic achievable sum-rate based on the stochastic successive upper-bound minimization technique. Extensive numerical results have quantified the regimes, in terms of fronthaul capacity, transmit power, channel coherence time and density of C-RANs, in which CAP and CBP are to be preferred.

In this paper, we have studied the design of the fronthaul compression strategies for the uplink of network MIMO systems by accounting for both CSI and data transfer from the RRHs to the BBU. Motivated by the information-theoretic optimization of separate estimation and compression, we have adopted an Estimate-Compress-Forward (ECF) approach, whereby the RRHs first estimate the CSI and then forward the compressed CSI to the BBU. The alternative Compress-Forward-Estimate (CFE) approach, already studied in previous work, is also considered for reference along with non-coherent transmission. Various schemes of increasing complexity are proposed that aim at optimizing the ergodic achievable sum-rate subject to fronthaul constraints. Specifically, separate and joint data signal and CSI compression strategies are devised. Moreover, in the presence of multiple RRHs, we have combined the proposed fronthaul strategies with distributed source coding to leverage the received signal correlation across RRHs. From numerical results, we have observed that the ECF approach outperforms the CFE approach, and that more complex joint compression strategies have significant advantages in the regime of intermediate fronthaul capacity, in which the fronthaul capacity should be used efficiently, and for sufficiently large SNR and channel coherence times. Finally, we have proposed a semi-coherent strategy that does not convey any CSI or pilot information over the fronthaul links. It was seen by numerical results that this scheme is large enough, while the latter is advantageous in the regime of low fronthaul capacity.

As a general conclusion, 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 large coherence periods and with stochastic, rather than instantaneous CSI. To elaborate, CBP requires data streams and precoding information to be sent on the fronthaul links. Hence, the fronthaul overhead of CBP increases with the network density, due to the larger number of data streams, and decreases with the coherence period and in the presence of stochastic CSI, owing to the reduced overhead for precoding. In contrast, the fronthaul overhead of CAP, which is due to the quantization of the baseband signals, does not depend on the network density, thus enabling to reap the interference management benefits of joint baseband processing at a larger scale. However, for small fronthaul capacities, large coherence periods and insufficiently dense networks, particularly in the presence of stochastic CSI, the interference management benefits of CAP may be outweighted by the lower fronthaul overhead of CBP.

References

1 Caire, G., Taricco, G., and Biglieri, E. (1999) Optimum power control over fading channels. *IEEE Trans. Info. Th.*, **45** (5), 1468–1489.

2 Hoydis, J., Kobayashi, M., and Debbah, M. (2011) Optimal channel training in uplink network MIMO systems. *IEEE Trans. Sig. Proc.*, **59** (6), 2824–2833.

3 Hassibi, B. and Hochwald, B.M. (2003) How much training is needed in multiple-antenna wireless links? *IEEE Trans. Info. Th.*, **49** (4), 951–963.

4 Zheng, L. and Tse, D.N.C. (2002) Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. *IEEE Trans. Info. Th.*, **48** (2), 359–383.

5 Kobayashi, M., Jindal, N., and Caire, G. (2011) Training and feedback optimization for multiuser MIMO downlink. *IEEE Trans. Comm.*, **59** (8), 2228–2240.

6 Marzetta, T.L. and Hochwald, B.M. (1999) Capacity of a mobile multiple-antenna communication link in rayleigh flat fading. *IEEE Trans. Info. Th.*, **45** (1), 139–157.

7 T. M. Cover and J. A. Thomas (2006) *Element of Information Theory*, John Wiley & Sons.

8 Gamal, A.E. and Kim, Y.H. (2011) *Network Information Theory*, Cambridge University Press.

9 Bjornson, E. and Ottersten, B.E. (2010) A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance. *IEEE Trans. Sig. Proc.*, **58** (3), 1807–1820.

10 Abou-Faycal, I.C., Trott, M.D., and Shamai, S. (2001) The capacity of discrete-time memoryless Rayleigh-fading channels. *IEEE Trans. Info. Th.*, **47** (4), 1290–1301.

11 Medard, M. (2000) The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. *IEEE Trans. Info. Th.*, **46** (3), 933–946.

12 Weingarten, H., Steinberg, Y., and Shamai, S. (2004) Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels. *IEEE Trans. Info. Th.*, **50** (8), 1665–1686.

13 S. Boyd and L. Vandenberghe (2004) *Convex Optimization*, Cambridge University Press.

14 Sanderovich, A., Somekh, O., Poor, H.V., and Shamai, S. (2009) Uplink macro diversity of limited backhaul cellular network. *IEEE Trans. Info. Th.*, **55** (8), 3457–3478.

15 Zhou, L. and Yu, W. Uplink multicell processing with limited backhaul via successive interference cancellation. *arXiv:1208.3024*.

16 Lim, S.H., Kim, Y.H., Gamal, A.E., and Chung, S.Y. (2011) Noisy network coding. *IEEE Trans. Info. Th.*, **57** (5), 3132–3152.

17 Shin, H. and Lee, J.H. (2003) Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyholes. *IEEE Trans. Info. Th.*, **49** (10), 2636–2647.

18 del Coso, A. and Simoens, S. (2009) Distributed compression for MIMO coordinated networks with a backhaul constraint. *IEEE Trans. Wireless Comm.*, **8** (9), 4698–4709.

19 Kang, J., Simeone, O., Kang, J., and Shamai, S. Joint signal and channel state information compression for the backhaul of uplink network MIMO systems. *arXiv:1306.0865*.

20 Zhang, X., Chen, J., Wicker, S.B., and Berger, T. (2007) Successive coding in multiuser information theory. *IEEE Trans. Info. Th.*, **53** (6), 2246–2254.

21 Park, S.H., Simeone, O., Sahin, O., and Shamai, S. (2013) Robust and efficient distributed compression for cloud radio access networks. *IEEE Trans. on Veh. Technol.*, **62** (2), 692–703.

22 Sklar, B. (1997) Rayleigh fading channels in mobile digital communication systems. I. characterization. *IEEE Comm. Mag.*, **35** (7), 90–100.

23 Simeone, O., Levy, N., Sanderovich, A., Somekh, O., Zaidel, B.M., Poor, H.V., and Shamai, S. (2011) *Cooperative wireless cellular systems: An information-theoretic view*, Foundations and Trends in Commun. Inf. Theory.

24 Adhikary, A., Nam, J., Ahn, J.Y., and Caire, G. (2014) Joint spatial division and multiplexing: The large-scale array regime. *IEEE Trans. Info. Th.*, **59** (10), 6441–6463.

25 Simeone, O., Somekh, O., Poor, H.V., and Shamai, S. (2009) Downlink multicell processing with limited-backhaul capacity. *EURASIP Jour. Adv. Sig. Proc.*

26 Marsch, P. and Fettweis, G. (2009) On downlink network MIMO under a constrained backhaul and imperfect channel knowledge. *Proc. IEEE Glob. Comm. Conf.*

27 Patil, P. and Yu, W. (2014) Hybrid compression and message-sharing strategy for the downlink cloud radio-access network. *Proc. of IEEE Info. Th. and Application Workshop*.

28 Park, S.H., Simeone, O., Sahin, O., and Shamai, S. (2013) Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks. *IEEE Trans. Sig. Proc.*, **61** (22), 5646–5658.

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

30 Yang, Y., Scutari, G., and Palomar, D.P. (2013) Parallel stochastic decomposition algorithms for multi-agent systems. *Proc. IEEE Workshop on Sign. Proc. Adv. in Wireless Comm.*, pp. 180–184.

31 Hunter, D.R. and Lange, K. (2004) A tutorial on MM algorithms. *The American Statistician*, **58** (1), 30–37.

32 Beck, A. and Teboulle, M. (2010) Gradient-based algorithms with applications to signal recovery problems. *in Convex Optimization in Signal Processing and Communications*.

33 Vandenberghe, L. and Boyd, S. (1996) Semidefinite relaxation of quadratic optimization problems. *SIAM Rev.*, **38** (1), 49–95.