Multivariate Fronthaul Quantization for C-RAN Downlink: Channel-Adaptive Joint Quantization in the Cloud

Wonju Lee*, Osvaldo Simeone†, Joonhyuk Kang* and Shlomo Shamai (Shitz)‡
*Dept. of Electrical Engineering, KAIST
Email: wonjulee@kaist.ac.kr
†ECE Dept., NJIT
‡Dept. of Electrical Engineering, Technion

***Abstract*—In the downlink of the Cloud-Radio Access Network (C-RAN) cellular architecture, complex baseband signals are transmitted from a central unit (CU) in the "cloud" over digital fronthaul links to distributed radio units (RUs). The standard design of digital fronthauling is based on quantization that operates separately over each fronthaul link. In this paper, a fronthaul quantization scheme is proposed that, unlike conventional schemes, implements a joint quantization mapping across all fronthaul links that is adapted to the current channel conditions. As compared to the current standard approach, the proposed multivariate quantization (MQ) scheme only requires additional processing at the CU, while no modification is needed at the RUs. The algorithm is extended to enable variable-length compression, and is compared via numerical results to a related approach based on the information-theoretic technique of multivariate compression.**

***Index Terms*—C-RAN, fronthaul, quantization, downlink, 5G.**

I. Introduction

Cloud Radio Access Network (C-RAN) is one of the key technologies for 5G systems due to its significant advantages in terms of lower expenses, flexibility and enhanced spectral efficiency. In a C-RAN, the Radio Unit (RU) of each base station is separated from its baseband unit, with the latter being implemented in the "cloud" at a Central Unit (CU) (see Fig. 1) [1]. The main obstacle to the realization of the promised performance of C-RAN is due to the restrictions on the capacity and latency of the so-called fronthaul links that provide connectivity between RUs and the CU.

The standard design of digital fronthauling, that is, of the transmission of digitized baseband complex samples on the fronthaul links, is based on quantization that applies to each fronthaul link separately in order to match the fronthaul capacity limitations. This is currently implemented by following the Common Public Radio Interface (CPRI) specification [1]. We refer to this conventional approach as *point-to-point* quantization (PtPQ). PtPQ can also be improved upon by allowing for compression, applied again separately on each fronthaul link, as done in, e.g., [2]-[6], which will be termed henceforth point-to-point compression (PtPC).

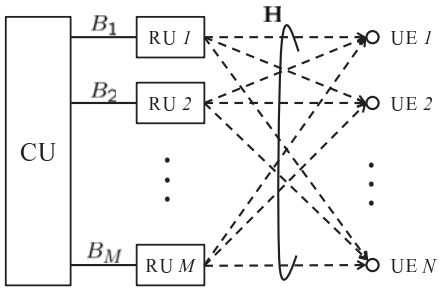

Fig. 1: Downlink C-RAN system with M RUs and N UEs. Each fronthaul link i can carry B_i bits per complex baseband sample.

In this paper, we consider the C-RAN downlink, and propose a fronthaul quantization scheme that, unlike conventional PtPQ, implements a *joint* quantization mapping across all fronthaul links. This mapping is adapted to the current channel state information (CSI) conditions with the aim of reducing the impact of the quantization noise on the received signal space. As compared to the current standard PtPQ, the proposed multivariate quantization (MQ) scheme only requires additional processing at the CU, while no modification is needed at the RUs. As it will be discussed, MQ can be interpreted as having a dual role as compared to precoding, or beamforming: while precoding decides which "spatial directions" should be occupied by the signal, MQ controls which "spatial directions" are mostly affected by the quantization error.

Specifically, we propose an algorithm that designs the quantization codebooks, one for each fronthaul link, based on long-term CSI, while the joint quantization mapping is adapted to current CSI. The proposed algorithm is based on iterative optimization approaches similar to the classical Lloyd-Max and Linde-Buzo-Gray algorithms [7]. Furthermore, we investigate the potential advantages of complementing MQ with variable-length compression by designing an entropy-constrained version of MQ. The proposed MQ approach can be seen as a low-complexity implementation of the Multivariate Compression (MC) strategy introduced in [8] by means of a network information-theoretic analysis. Via numerical results, we compare PtPQ, PtPC, MQ and MC, yielding insights into

their relative performance.

The rest of the paper is organized as follows. Sec. II presents the system model and Sec. III introduces with a simple example the key ideas behind MQ. Sec. IV describes the proposed algorithm design, and Sec. V discusses the entropy-constrained MQ design. Sec. VI offers numerical results and Sec. VII some final remarks.

II. System Model and Design Criterion

We consider the downlink of a C-RAN in which M RUs cover an area with N active user equipments (UEs), as illustrated in Fig 1. Baseband processing for the M RUs is carried out at a CU. The CU transfers the baseband signals to each RU through a fronthaul link with capacity B_i as measured in bits per complex baseband sample, for $i = 1, \ldots, M$.

Let us define as $\mathbf{s} = [s_1, \ldots, s_N]^T$ the $N \times 1$ vector of complex information-bearing symbols at a given channel use, where s_k is the symbol intended for UE k that satisfies the normalization $E[|s_k|^2] = 1$. Each symbol s_k may be typically assumed to be distributed as a zero-mean complex Gaussian variable, so as to obtain a modulation-independent solution or to account for OFDM transmission in the time domain by the law of large numbers as in e.g., [3].

Assuming full CSI at the CU, the information-bearing vector $\mathbf{s}$ is linearly precoded by means of an $M \times N$ precoding matrix $\mathbf{W_H} = [\mathbf{w}_{\mathbf{H},1}, \ldots, \mathbf{w}_{\mathbf{H},N}]$, where $\mathbf{w}_{\mathbf{H},k}$ is the beamforming, or precoding, vector for the signal s_k intended for UE k. The subscript $\mathbf{H}$ indicates the dependence of the precoding vectors on the channel matrix $\mathbf{H}$, to be introduced below. Hence, the precoded signal $\mathbf{x} = [x_1, \ldots, x_M]^T$ is given as

$$\mathbf{x} = \sum_{k=1}^{N} \mathbf{w}_{\mathbf{H},k} s_k = \mathbf{W_H s}. \tag{1}$$

To satisfy the capacity limitations of the fronthaul links, the precoded signal x_i for the RU i is quantized to B_i bits producing the signal $\hat{x}_i$. The signal $\hat{x}_i$ is selected from a set $\hat{\mathcal{X}}_i$ of cardinality 2^{B_i}, which will be referred to as a codebook. We define the quantized $M \times 1$ vector as $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_M]^T$ and assume the per-RU power constraint

$$E[|\hat{x}_i|^2] \leq 1 \tag{2}$$

for all $i = 1, \ldots, M$. Each RU i is assumed to be informed about the codebook $\hat{\mathcal{X}}_i$. No additional information, such as CSI, is instead assumed at the RU. Furthermore, codebooks are assumed to be updated only at the time scale of the variations of the long-term statistical properties of the channels $\mathbf{H}$. As a result, RUs need to be informed about new codebooks only when the statistics of the channels, such as the path-loss and shadowing, change significantly.

Since each RU i transmits $\hat{x}_i$, the received signal at the UE k can be written as

$$\begin{aligned} y_k &= \sqrt{P}\mathbf{h}_k^T \hat{\mathbf{x}} + z_k \\ &= \sqrt{P}\mathbf{h}_k^T \mathbf{w}_{\mathbf{H},k} s_k + \sqrt{P}\mathbf{h}_k^T (\hat{\mathbf{x}} - \mathbf{w}_{\mathbf{H},k} s_k) + z_k, \end{aligned} \tag{3}$$

where P is a dimensionless parameter that accounts for the transmitted power of the RUs; $\mathbf{h}_k$ is the $M \times 1$ channel vector which is assumed to have a given distribution e.g., Rayleigh with possibly correlated entries; and $z_k \sim \mathcal{CN}(0,1)$ is the additive Gaussian noise. From (3), the effective signal-to-noise ratio (SNR) at the UE k can be obtained as

$$\text{SNR}_k^{\text{eff}} = \frac{PE\left[\left|\mathbf{h}_k^H \mathbf{w}_{\mathbf{H},k}\right|^2\right]}{1 + PE\left[\left|\mathbf{h}_k^H (\mathbf{w}_{\mathbf{H},k} s_k - \hat{\mathbf{x}})\right|^2\right]}, \tag{4}$$

where the expectation is taken with respect to the distribution of the channel $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$ and of $\mathbf{s}$, and the second term in the denominator measures the power of the interference term in (3) due to quantization.

We observe that quantization only affects the interference power in (4) and hence optimal quantizers should minimize the powers $E[|\mathbf{h}_k^H (\mathbf{w}_{\mathbf{H},k} s_k - \hat{\mathbf{x}})|^2]$ for all $k = 1, \ldots, N$. To tackle this multiobjective problem, following the standard scalarization approach [9], here we propose to design quantizers that minimize the scalarized weighted mean squared error

$$\sum_{k=1}^{N} \alpha_k E\left[\left|\mathbf{h}_k^H (\mathbf{w}_{\mathbf{H},k} s_k - \hat{\mathbf{x}})\right|^2\right], \tag{5}$$

where the weights $\alpha_k \geq 0$ can be selected to enforce some fairness criterion (see [10]). We also note that (5) differs from standard quantization error metrics, such as the error vector magnitude (EVM) or mean squared error [2]-[6], since (5) directly captures the overall system performance while the mentioned metrics apply only on a per-fronthaul link basis.

III. Introducing Multivariate Quantization

In this section, we present intuitive arguments to illustrate the basic principles and potential benefits of MQ. This is done by contrasting MQ with standard PtPQ that operates separately on each fronthaul link. To this end, we focus on the case of a single UE, i.e., $N = 1$ and two RU, i.e., $M = 2$, and assume for simplicity of visualization a real-valued system model. Moreover, to further streamline the discussion, we adopt the matched beamformer $\mathbf{w_H} = \mathbf{h}$, where $\mathbf{h}$ is the 2×1 (real) channel vector for the given UE, and we have dropped the UE subscript to simplify the notation. The transmitted signal (1) can hence be written as $\mathbf{x} = \mathbf{h}s$, and some realization of $\mathbf{x}$ are shown as dots along the line with slopes -55 degrees line in Fig. 2 under the assumption that s is a Gaussian random variable and the channel vector is $\mathbf{h} = [-\sqrt{1/3}, \sqrt{2/3}]^T$. The figure also shows as squares on the horizontal and vertical axes the quantization levels that define the quantization codebooks $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$ for the two RUs. Note that the codebooks are the same for both conventional PtPQ (Fig. 2(a)) and MQ (Fig. 2(b)), as further discussed below. As mentioned, each RU is informed only about its own codebook.

Because of quantization, the signal $\hat{\mathbf{x}}$ sent by two RUs must correspond to one of the quantization points (crosses) on the plane. Therefore, the quantization error $(\mathbf{h}s - \hat{\mathbf{x}})$ between the desired signal, which lies on the -55° line, and the selected

point (cross) should be considered as a disturbance to the reception of the UE. The key observation is that the impact of the quantization error $(\mathbf{h}s - \hat{\mathbf{x}})$ on the reception of the UE depends, by (3)-(4), solely on the power $|\mathbf{h}^T(\mathbf{h}s - \hat{\mathbf{x}})|^2$ of its projection on the $-55°$ line. That is, the only component of the quantization error that affects the UE is its projection onto the signal subspace.

Fig. 2(a) illustrates the quantization mapping resulting from standard uniform PtPQ of the signals to be transmitted by the two RUs. As seen in the figure, with PtPQ, the shape of the quantization regions is constrained to be rectangular. Instead, with MQ, quantization mapping is performed jointly for the two RUs, as illustrated in Fig. 2(b), as a function of current CSI. MQ hence enables a finer control of the impact of the quantization error on the received signals so that the projection onto the subspace occupied by the signal is minimized.

IV. Multivariate Quantization Design

In this section, we introduce the proposed MQ design. Throughout this paper, we assume that the downlink precoder $\mathbf{W_H}$ is fixed and not subject to optimization. Joint optimization of quantization and precoding is elaborated on in [11]. We aim to minimize the distortion criterion (5).

A quantizer consists of two elements, namely a quantization codebook and a mapping [7]. For MQ, the quantization codebook $\hat{\mathcal{X}} = \hat{\mathcal{X}}_1 \times \cdots \times \hat{\mathcal{X}}_M$ is given by the Cartesian product of the sets $\hat{\mathcal{X}}_i = \{\hat{x}_i^{(1)}, \ldots, \hat{x}_i^{(2^{B_i})}\}$ of the quantization levels $\hat{x}_i^{(j)}$ for each RU i (i.e., the crosses in Fig. 2(b)). The mapping, instead, is a function $f_{\hat{\mathcal{X}},\mathbf{H}}(\mathbf{x})$ that takes as input the baseband signal $\mathbf{x}$ in (1), the CSI $\mathbf{H}$, and the codebook $\hat{\mathcal{X}}$, and outputs the corresponding quantization levels $[\hat{x}_1^{(j_1)}, \ldots, \hat{x}_M^{(j_M)}]^T$ or, equivalently, their indices $[j_1, \ldots, j_M]$. The mapping defines the quantization regions illustrated in the example of Fig. 2.

For a fixed codebook $\hat{\mathcal{X}}$, the optimal mapping, from (5), is given by the function

$$f_{\hat{\mathcal{X}},\mathbf{H}}(\mathbf{x}) = \arg\min_{j_1,\ldots,j_M} \sum_{k=1}^{N} \alpha_k \left| \mathbf{h}_k^H \left(\mathbf{w}_{\mathbf{H},k} s_k - \hat{\mathbf{x}}^{(j_1,\ldots,j_M)} \right) \right|^2, \quad (6)$$

where $\hat{\mathbf{x}}^{(j_1,\ldots,j_M)} = [\hat{x}_1^{(j_1)}, \ldots, \hat{x}_M^{(j_M)}]$ is taken from the codebook $\hat{\mathcal{X}}$ and we have made explicit the dependence of the function on both the codebook $\hat{\mathcal{X}}$ and on the channel $\mathbf{H}$. Moreover, the design of the quantization codebook $\hat{\mathcal{X}}$ can be formulated as the problem

$$\hat{\mathcal{X}} = \arg\min_{\{\hat{x}_i^{(j)}\}_{i=1}^{M} {}_{j=1}^{2^{B_i}}} \sum_{k=1}^{N} \alpha_k E\left[\left| \mathbf{h}_k^H \left(\mathbf{w}_{\mathbf{H},k} s_k - \hat{\mathbf{x}}^{(f_{\hat{\mathcal{X}},\mathbf{H}}(\mathbf{x}))} \right) \right|^2 \right]$$
$$\text{s.t. } E\left[|\hat{x}_i|^2 \right] \leq 1 \text{ for } i = 1, \ldots, M, \quad (7)$$

where the expectations are taken with respect to $\mathbf{s}$ and $\mathbf{H}$. We observe that, as per (7), the codebook $\hat{\mathcal{X}}$ is not a function of the instantaneous CSI but only of the distribution of $\mathbf{H}$. This guarantees that the RUs need not be informed about a new codebook any time the channel $\mathbf{H}$ changes but only at the time scale of the variations of long-term CSI. The RUs need also not be informed about the mapping (6).

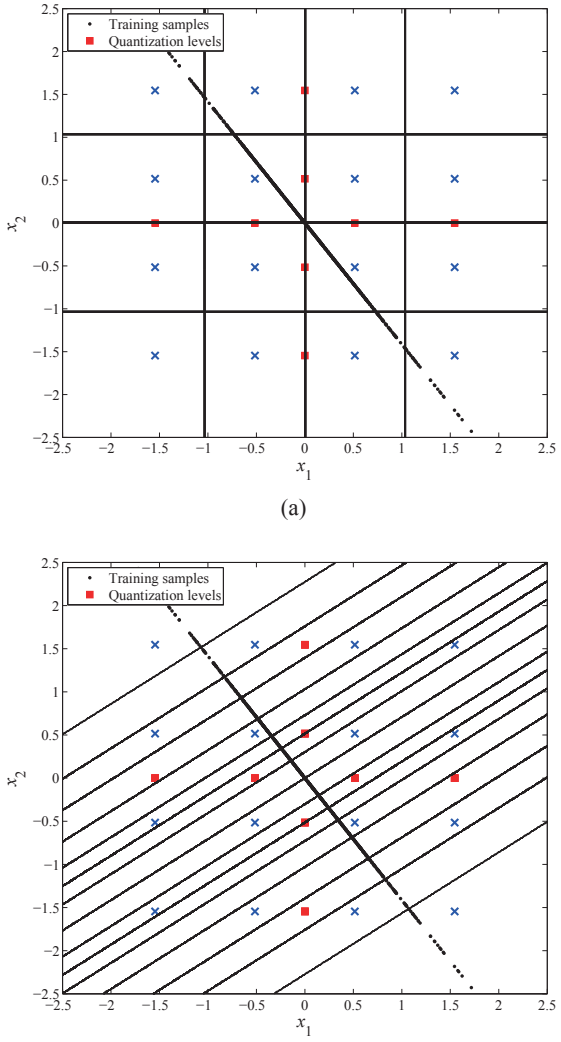

Fig. 2: Illustration of: (a) conventional Point-to-Point Quantization (PtPQ); (b) Multivariate Quantization (MQ).

Codebook Optimization: In order to address the optimization (7) over the codebook $\hat{\mathcal{X}}$, we follow the standard approach of the Lloyd-Max algorithm, and its extension to vector quantization due to Linde-Buzo-Gray [7], by iterating between the application of the mapping (6) for a fixed codebook and the optimization of the codebook (7) for the obtained mapping given the current codebook iterate. The algorithm is based on randomly generated training samples for $\mathbf{s}$ and for $\mathbf{H}$, namely $\mathcal{S} = \{\mathbf{s}(1), \ldots, \mathbf{s}(N_s)\}$ and $\mathcal{H} = \{\mathbf{H}(1), \ldots, \mathbf{H}(N_h)\}$, respectively, and is detailed in Algorithm 1. We emphasize that the algorithm is run offline based only on long-term CSI. Moreover, once the codebook $\hat{\mathcal{X}}$ is designed, the mapping (6) is applied for the given instantaneous CSI $\mathbf{H}$.

Referring to Algorithm 1, the proposed MQ design scheme updates the codebook $\hat{\mathcal{X}}$ by solving the quadratic convex

Algorithm 1 Multivariate Quantization (MQ) Design

Input: Generate N_h independent training channels $\mathcal{H} = \{\mathbf{H}(1), \ldots, \mathbf{H}(N_h)\}$ from the known channel distribution and N_s independent training values $\mathcal{S} = \{\mathbf{s}(1), \ldots, \mathbf{s}(N_s)\}$ from the known distribution of $\mathbf{s}$. Set a threshold $\epsilon \geq 0$.

Initialization: Initialize the codebook as $\hat{\mathcal{X}}^{[1]}$ and set $t = 1$ and $D^{[0]} = \infty$.

repeat

for $m = 1$ to N_h **do**

• Find the minimum-distortion partition of $\mathcal{S}$ when the channel is $\mathbf{H}(m)$ as $\mathcal{S}^{(j_1,\ldots,j_M,m)} = \{\mathbf{s} \in \mathcal{S} : f_{\hat{\mathcal{X}}^{[t]},\mathbf{H}(m)}(\mathbf{W}_{\mathbf{H}(m)}\mathbf{s}) = [j_1, \ldots, j_M]^T\}$ with $j_i = 1, \ldots, 2^{B_i}$, $i = 1, \ldots, M$ and $m = 1, \ldots, N_h$.

end for

• Compute the average distortion $D^{[t]} = 1/(N_h N_s) \sum_{m=1}^{N_h} \sum_{n=1}^{N_s} \sum_{k=1}^{N} |\mathbf{h}_k(m)^H (\mathbf{w}_{\mathbf{H}(m),k} s_k(n) - \hat{\mathbf{x}}^{(f_{\hat{\mathcal{X}}^{[t]},\mathbf{H}(m)}(\mathbf{W}_{\mathbf{H}(m)}\mathbf{s}(n)))})|^2$.

• Obtain the updated codebook as $\hat{\mathcal{X}}^{[t+1]}$ by solving the quadratic convex problem (8).

• Set $t = t + 1$.

until $(D^{[t-1]} - D^{[t]})/D^{[t]} \leq \epsilon$.

Output: Codebook $\hat{\mathcal{X}} = \hat{\mathcal{X}}^{[t]}$.

problem

$$\hat{\mathcal{X}}^{[t+1]} = \arg \min_{\{\hat{x}_i^{(j_i)}\}_{i=1}^{M} {}_{j_i=1}^{2^{B_i}}} \sum_{m=1}^{N_h} \sum_{j_1,\cdots,j_M} \sum_{n:\ \mathbf{s}(n) \in \mathcal{S}^{(j_1,\ldots,j_M,m)}} \times \sum_{k=1}^{N} \alpha_k \left| \mathbf{h}_k(m)^H (\mathbf{w}_{\mathbf{H}(m),k} s_k(n) - \hat{\mathbf{x}}^{(j_1,\ldots,j_M)}) \right|^2$$
$$\text{s.t.} \sum_{j_i=1}^{2^{B_i}} p(\hat{x}_i^{(j_i)}) |\hat{x}_i^{(j_i)}|^2 \leq 1 \text{ for } i = 1, \ldots, M, \quad (8)$$

with $\sum_{i=1}^{M} 2^{B_i}$ unknown variables, where $p(\hat{x}_i^{(j_i)}) = \sum_{m=1}^{N_h} \sum_{j_k \neq j_i} \left|\mathcal{S}^{(j_1,\ldots,j_M,m)}\right| / N_h N_s$ is the fraction of the $N_h N_s$ training samples that are quantized to $\hat{x}_i^{(j_i)}$ for each RU i, and the set $\mathcal{S}^{(j_1,\ldots,j_M,m)}$ contains all training samples in $\mathcal{S}$ that are mapped to the quantization index $[j_1, \ldots, j_M]$ when the CSI is $\mathbf{H}(m)$ by mapping (6).

V. Entropy-Constrained Multivariate Quantization

To reduce the bit rate produced by MQ, in this section, we consider the practically relevant case in which the joint quantizer of MQ is followed by a separate entropy encoder, or compressor, for each fronthaul link that produces variable-length descriptions for each sample. The variable-rate compressor assigns more bits to the most used quantization levels and less bits to the least used levels, so that the average quantization output is B bit/symbol. The resulting quantization-compression system has the advantage of potentially reducing the fronthaul overhead for a given quantization resolution, although this gain comes at the price of requiring the implementation of a buffer per fronthaul link in order to smooth out the variance of the variable-length descriptions [7]. Alternatively, optimal entropy encoders can also be implemented using block processing.

We pursue here the optimization of the discussed quantization-compression system with both PtPQ and MQ by adopting the standard framework of entropy-constrained optimization [7]. To this end, we modify the codebook design problem (5) as the minimization of

$$\sum_{k=1}^{N} \alpha_k E\left[\left| \mathbf{h}_k^H \left(\mathbf{x} - \hat{\mathbf{x}}^{(f_{\hat{\mathcal{X}}}(\mathbf{x}))} \right) \right|^2 \right] + \sum_{i=1}^{M} \lambda_i H\left(\hat{x}_i^{(f_{\hat{\mathcal{X}}_i}(x_i))} \right), \quad (9)$$

where $f_{\hat{\mathcal{X}}}(\cdot)$ is a mapper to be optimized based on (9); $H(\cdot)$ measures the entropy of its input; and the parameters $\lambda_i \geq 0$ define the desired tradeoff between distortion and fronthaul rate. The quantization codebook $\hat{\mathcal{X}} = \hat{\mathcal{X}}_1 \times \cdots \times \hat{\mathcal{X}}_M$ consists of the Cartesian product of the sets $\hat{\mathcal{X}}_i = \{\hat{x}_i^{(1)}, \ldots, \hat{x}_i^{(2^{B_i'})}\}$ of the quantization levels $\hat{x}_i^{(j)}$ for each RU i, where $B_i' \geq B_i$ is a parameter that defines the resolution of the quantizer for RU i. The optimization algorithm follows a similar iterative approach as in the previous section, and is omitted here for lack of space. It can be found in [11].

VI. Numerical Results

Throughout this section, we assume that every RU is subject to the same power constraint P and has equal fronthaul capacity $B_1 = \cdots = B_M = B$ bit/symbol. Furthermore, we consider a single user ($N = 1$) and fix the precoding to the matched beamformer as in Sec. III. For all design schemes, we assume Gaussian training samples for $\mathbf{s}$, and we set $\epsilon = 0.001$ in Algorithms 1 and $B_i' = B_i + 1$ for $i = 1, \ldots, M$ for entropy-coded quantization. We compare the performance of PtPQ, MQ – with and without entropy coding – and of PtPC and MC. We refer to [11] for a detailed description of the design of PtPQ, which follows the standard approach [7]. As for PtPC and MC, we use the information-theoretic expressions presented in [8], which imply optimal block-based, rather than symbol-by-symbol, as in PtPQ and MQ, quantization and compression. We observe that the performance is generally enhanced as one moves from less complex schemes based only on symbol-by-symbol quantization, such as PtPQ or MQ, to more complex schemes involving entropy coding, with the best performance being achieved by the information-theoretically optimal block processing, such as PtPC and MC.

We first investigate the performance as a function of the fronthaul capacity B in Fig. 3. We consider $M = 4$ and $P = 10$ dB. Fig. 3 demonstrates the remarkable gains that can be achieved using MQ as compared to PtPQ, particularly for low values of the fronthaul capacity. For instance, for $B = 2$ bit/symbol, MQ achieves a 197% gain over PtPQ. Furthermore, block processing, as in PtPC and MC, is seen to be able to significantly improve the performance of symbol-by-symbol quantization for low values of B. As an example, for $B = 1$ bits/symbol, the performance of MC in terms of

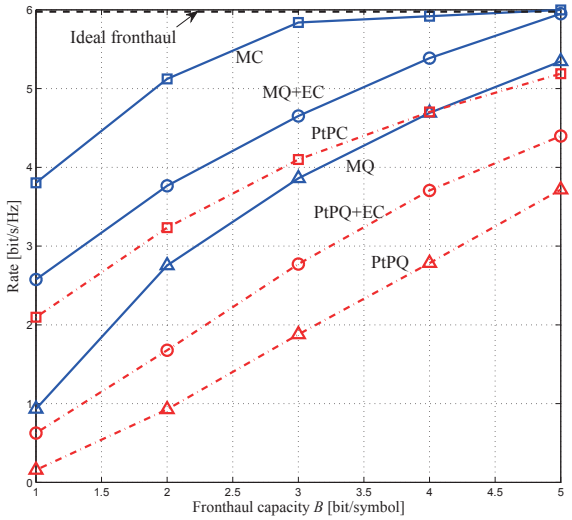

Fig. 3: Rate R versus the fronthaul capacity B for PtPQ and MQ, with and without entropy encoding, and for PtPC and MC ($M = 4$, $N = 1$, and $P = 10$ dB).

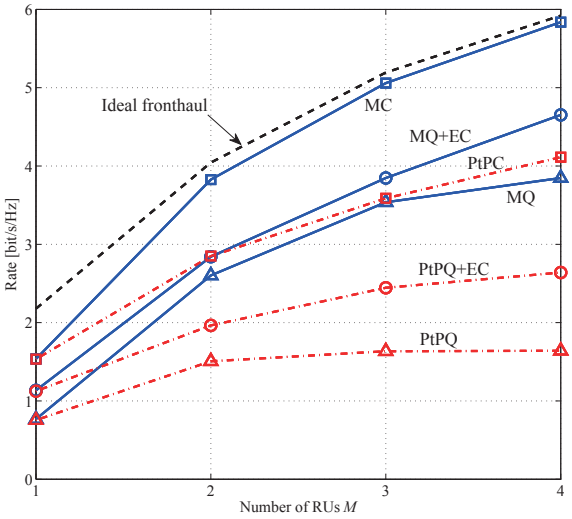

Fig. 4: Rate R versus the number M of RUs for PtPQ and MQ, with and without entropy encoding, and for PtPC and MC ($N = 1$, $P = 10$ dB, and $B = 3$ bit/symbol).

rate is approximately four times larger as compared to MQ. However, this performance gain decreases with B, e.g., with $B = 4$ bit/symbol, MC provides only a gain of 1.2 bit/s/Hz with respect to MQ. Furthermore, MQ is seen to potentially even improve over the information-theoretically optimal PtPC if B is large enough. This demonstrates that a symbol-by-symbol approach can be an effective close-to-optimal solution as long as the fronthaul capacity is not too small.

Fig. 4 compares the mentioned rates when varying the number M of RUs. We set $P = 10$ dB and $B = 3$ bit/symbol. The gains of MQ over PtPQ are seen to become more relevant as M gets larger due to the corresponding increase in the number of design degrees of freedom. Furthermore, the benefits of entropy coding or block processing are more significant for point-to-point techniques than for multivariate schemes. For instance, for $M = 4$, in the case of point-to-point techniques, entropy coding improves the rate of standard quantization by 60% and block processing improves the rate by 160%. Instead, for multivariate approaches, the corresponding gains are 21% and 52%. This can be ascribed to the capabilities of MQ and MC to reduce the impact of the quantization error as compared to PtPQ and PtPC, hence making the use of more sophisticated compression techniques less relevant. Note that, in Fig. 4, this is particularly evident as the number of RUs increases given the enhanced effectiveness of MQ in this regime.

VII. Conclusions

This paper has introduced a novel fronthaul quantization method for the downlink of C-RAN. The approach is based on a CSI-aware optimization of the quantization mapping at the cloud that operates jointly across all fronthaul links. It has been demonstrated that the proposed multivariate quantization (MQ) scheme yields significant performance gain over per-fronthaul link point-to-point quantization (PtPQ) as carried out in the CPRI standard. We refer to [11] for additional discussions, including the joint optimization of precoding and MQ and further numerical experiments.

Acknowledgment

The work of O. Simeone was partially supported by U.S. NSF under grant CCF-1525629. The work of S. Shamai was supported by the Israel Science Foundation and the S. AND N. grand research fund.

References

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - A technology overview," *IEEE Commun. Surveys, Tutorials*, vol. 17, no. 1, pp. 405-426, First quarter 2015.

[2] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3216-3225, Sep. 2012.

[3] K. F. Nieman and B. L. Evans. "Time-domain compression of complex-baseband LTE signals for cloud radio access networks," in *Proc. IEEE Global Confer. Signal, Inform. Process. (GlobalSIP)*, pp. 1198-1201, Austin, USA, Dec. 2013.

[4] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of LTE/LTE-advanced cloud-RAN architectures," in *Proc. IEEE Int. Symp. World of Wireless, Mobile, Multimedia Networks (WoWMoM)* pp. 1-9, Madrid, Spain, Jun. 2013.

[5] A. Vosoughi, M. Wu, and J. R. Cavallaro, "Baseband signal compression in wireless base stations," in *Proc. IEEE Global Commun. Confer. (GLOBECOM)*, pp. 4505-4511, Anaheim, USA, Dec. 2012.

[6] H. Si, B. L. Ng, M. S. Rahman, and J. Zhang, "A novel and efficient vector quantization based CPRI compression algorithm," *arXiv:1510.04940*.

[7] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer Acad. Press, 1992.

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

[9] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2210-2239, Nov. 1998.

[10] K. Miettinen, *Nonlinear multiobjective optimization*, Kluwer Acad. Press, 2012.

[11] W. Lee, O. Simeone, J. Kang, and S. Shamai (Shitz), "Multivariate fronthaul quantization for downlink C-RAN," *arXiv:1510.08301*.