This is a repository copy of *On The Uplink Throughput of Zero-Forcing in Cell-Free Massive MIMO with Coarse Quantization*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/147065/

Version: Accepted Version

Article:

Maryopi, Dick, Bashar, Manijeh and Burr, Alister Graham orcid.org/0000-0001-6435-3962 (2019) On The Uplink Throughput of Zero-Forcing in Cell-Free Massive MIMO with Coarse Quantization. IEEE Transactions on Vehicular Technology. pp. 1-5. ISSN 0018-9545

https://doi.org/10.1109/TVT.2019.2920070

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2920070, IEEE Transactions on Vehicular Technology

On The Uplink Throughput of Zero-Forcing in Cell-Free Massive MIMO with Coarse Quantization

Dick Maryopi, Manijeh Bashar and Alister Burr

***Abstract*—The recently proposed Cell-Free massive MIMO architecture is studied for the uplink. In contrast to most previous works, joint detection is performed using global CSI. Therefore, we study strategies for transferring CSI to the CPU taking into account the fronthaul capacity which limits CSI quantization. Two strategies for pilot-based CSI acquisition are considered: *estimate-and-quantize* and *quantize-and-estimate*. These are analysed using the Bussgang decomposition. For a given quantization constraint for the data and CSI the achievable rate per user with Zero-Forcing is determined. Numerical results show that quantize-and-estimate (the simpler strategy) is similar to or better than estimate-and-quantize at low resolution, especially for 1-bit.**

***Index Terms*—Cell-Free Massive MIMO, Fronthaul, Quantization, Bussgang, Channel Estimation.**

I. INTRODUCTION

THE next generation of wireless networks (including 5G) will be required to provide a high capacity per user and per unit area due to the increasing number of users and the variety of applications expected in the near future. Cell-free massive Multiple Input Multiple Output (MIMO) has been gaining more attention recently as it has the potential to meet this demand [1]. It can be regarded as a form of network MIMO which makes use of a large number of distributed antennas, referred to Access Points (AP), spread over a large coverage area. The term "cell-free" was motivated by the notion of blurring the role of cells so that all users can be served by all APs over the same resources using network MIMO techniques to avoid mutual interference. Because a large number of APs serve a smaller number of users, it still benefits from *channel hardening* as in co-located massive MIMO [2].

Nevertheless, the joint transmission/detection in the most current cell-free system is based only on local Channel State Information (CSI). We identify this a limitation, since relying on local CSI at the APs restricts the feasible choice of processing to conjugate beamforming or Maximum Ratio Combining (MRC). Other forms of processing such as Zero Forcing (ZF) could be performed at Central Processing Unit (CPU), but would require additional CSI transfer via the fronthaul. However, CF-massive MIMO already faces the problem of high fronthaul load requirement. To address these issues, we study in this paper joint detection with global CSI at the CPU and the strategy of acquiring the required CSI. To deal with the growth of fronthaul load we assume a coarse quantization constraint, which is also of interest for the low-cost implementation of APs. We show that using appropriate CSI acquisition strategies, much improved detection techniques can be applied at the CPU resulting in a significant rate improvement in uplink. In [3], a performance improvement of ZF over MRC has also been shown in downlink where a sort of global CSI is used at the CPU for precoding. Nevertheless, they didn't address specifically the CSI acquisition schemes and the limited fronthaul capacity.

After a brief description of our system model we investigate two strategies of CSI acquisition. The first is called *estimate-and-quantize* (EQ) where channel estimation is carried out at the AP. The channel estimate is quantized and then the quantized form is sent to the CPU. This is similar to the sharing of quantized CSI between the base stations in the coordinated multipoint (CoMP) scheme [4]. As alternative we consider *quantize-and-estimate* (QE), where the APs quantize the received pilot and send it to the CPU. From these quantized received pilots the CPU performs the channel estimation. Further, we compare their performance and their corresponding throughput for ZF detection. Surprisingly enough, the QE strategy, which is simpler for the implementation at the AP, has good performance and a significant performance improvement over EQ for 1-bit fronthaul resolution. Overall, the superiority of utilizing global CSI, even with coarse quantization, is shown to be significant compared to utilizing only local CSI with infinite resolution.

Notation: Roman letters, lower-case boldface letters and upper-case boldface letters are used respectively to denote scalars, column vectors and matrices. The set of all complex and real $M \times N$ matrices are represented by $\mathbb{R}^{M \times N}$ and $\mathbb{C}^{M \times N}$ respectively. The real part and imaginary part of complex numbers are expressed respectively by $\mathbb{R}\mathrm{e}\{\cdot\}$ and $\mathbb{I}\mathrm{m}\{x\}$. By $\langle \cdot, \cdot \rangle$ we denote the inner product with $\|\cdot\|$ as its corresponding vector norm or Frobenius norm. The expectation of random variables is represented by $\mathbb{E}\{\cdot\}$. We denote the circularly complex Gaussian distribution with zero mean and unit variance by $\mathcal{CN}(0, 1)$. We use $\mathbf{I}_K$ for the $K \times K$ identity matrix and $\mathbf{1}_K$ for all-one vector of dimension K. For a vector $\mathbf{a}$, $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with the diagonal elements taken from vector $\mathbf{a}$.

II. SYSTEM MODEL

We consider the uplink transmission of a cell-free system [1], where we have K single-antenna users (UEs) and M single-antenna APs connected to a CPU by M error-free fronthaul links. The main processing for these M APs are virtualized at

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The authors are with the Department of Electronic Engineering, University of York, Heslington, York, UK. email: dm1110@york.ac.uk. The paper was supported by Indonesia Endowment Fund for Education (LPDP) and in part by the European Horizon 2020 Programme under GA H2020-MSCA-ITN-2016-722788

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2920070, IEEE Transactions on Vehicular Technology

the CPU, where the communication between them occurs in baseband form. We assume that the fronthaul link connecting the m-th AP with the CPU can in practice transmit reliably at a maximum rate of R_m.

A. Channel Model

The channel between the k-th user and the m-th AP is specified (as in [1]) by

$$g_{mk} = h_{mk}\beta_{mk}^{1/2}, \tag{1}$$

where the coefficient h_{mk} models the small-scale fading between the k-th user and the m-th AP with the assumption that it is i.i.d. $\sim \mathcal{CN}(0,1)$. The large-scale fading is denoted by β_{mk} which is likely to be different for each user k and each AP m due to the distributed configuration. The channel from all K users to all M APs can then be expressed as the element-wise product of small-scale fading matrix $\mathbf{H} \in \mathbb{C}^{M\times K}$ and large-scale fading matrix $\mathbf{D} \in \mathbb{R}^{M\times K}$ given by

$$\mathbf{G} = \mathbf{H} \odot \mathbf{D}^{1/2}, \text{ where } [\mathbf{H}]_{mk} = h_{mk}, [\mathbf{D}]_{mk} = \beta_{mk}. \tag{2}$$

B. Quantization Scheme

To simplify our analysis, we consider fronthaul links with $R_m = R$ bits, $\forall m \in \{1,\dots,M\}$, corresponding to the quantization level $L = 2^R$. Therefore, we apply an L-level scalar quantizer Q at each AP as an interface to the fronthaul with

$$Q(x) = \sum_{l=0}^{L-1} q_l T_l(x), \tag{3}$$

where $T_l(\mathbf{x})$ is equal 1 for $x_l < x \le x_{l+1}$ and 0 otherwise. We consider Q as a uniform quantizer with a fixed step size $\Delta = x_{l+1} - x_l$ and a reconstruction value $q_l = (l - \frac{L-1}{2})\Delta$. For a complex-valued signal $x \in \mathbb{C}$ we quantize the real and imaginary part separately. In this case, whenever we have $x_l < \mathbb{Re}\{x\} \le x_{l+1}$ and $x_{l'} < \mathbb{Im}\{x\} \le x_{l'+1}$ for $(l,l') \in \{0,\dots,L-1\}$, we obtain

$$\begin{aligned} x_q = Q(x) &= Q(\mathbb{Re}\{x\}) + iQ(\mathbb{Im}\{x\}) \qquad (4)\\ &= q_l^R + iq_{l'}^I, \qquad (5)\end{aligned}$$

where q_l^R and $q_{l'}^I$ are respectively the reconstruction values of the real and imaginary part with the pair $(q_l^R, q_{l'}^I) \in \{q_0^R,\dots,q_{L-1}^R\} \times \{q_0^I,\dots,q_{L-1}^I\}$. Moreover, the quantization operation should apply elementwise for a vector valued input. We assume that the large scale fading β_{mk} is relatively constant over a long period and known at the APs. Thus, we can scale the input-output signal of the quantizer according to β_{mk} and approximate the normalised input as normally distributed.

The function Q is the scalar quantization process, which is particularly nonlinear for small L. To analyse it, we use the Bussgang decomposition [5]. Accordingly, for a nonlinear function $Q(x)$ we can write it as

$$x_q = Q(x) = \alpha_q x + d. \tag{6}$$

The distortion term d is uncorrelated to the input signal x. The linear factor α_q depends on the characteristic of the quantizer Q and the distribution $f(x)$ of the input signal x given by [5, 6]

$$\begin{aligned}\alpha_q &= \frac{1}{P_x}\int_x xQ^*(x)f(x)dx\\ &= \frac{1}{P_x}\sum_{l=0}^{L-1} q_l \int_{x_l}^{x_{l+1}} xf(x)dx, \qquad (7)\end{aligned}$$

where $P_x = \mathbb{E}\{|x|^2\}$ is the power of x. As shown in [7], for normally distributed input and uniform quantizer it can be expressed in closed form as a function of Δ and L

$$\alpha_q = \frac{\Delta}{\sqrt{2\pi}}\left(1 + 2\sum_{l=1}^{L/2-1}\exp(-l^2\Delta^2)\right). \tag{8}$$

Further, we define the power ratio of the input x and the output x_q in terms of Δ and L as given in [7] by

$$\begin{aligned}\lambda_q = \frac{\mathbb{E}\{|x_q|^2\}}{\mathbb{E}\{|x|^2\}} &= \frac{1}{P_x}\int_x |Q(x)|^2 f(x)dx \qquad (9)\\ &= \frac{1}{P_x}\sum_{l=0}^{L-1} q_l^2 \int_{x_i}^{x_{l+1}} f(x)dx \qquad (10)\\ &= \Delta^2\left(\frac{1}{4} + 4\sum_{l=1}^{L/2-1} l(1-\Phi(l\Delta))\right), \qquad (11)\end{aligned}$$

where Φ is the Gaussian cumulative distribution function. We choose here the step size Δ that maximizes the Signal to Distortion Noise Ratio (SDNR) at the output of the quantizer defined as

$$\text{SDNR} = \mathbb{E}\{|\alpha_q x|^2\}/\mathbb{E}\{|d|^2\}. \tag{12}$$

From (6) and (11) the power of the distortion is given by

$$\mathbb{E}\{|d|^2\} = \mathbb{E}\{|x_q - \alpha x|^2\} = (\lambda_q - \alpha_q^2)\mathbb{E}\{|x|^2\}. \tag{13}$$

Using equations (8) and (11) we characterize the Bussgang decomposition such that it is directly related to the parameter Δ and L. This will be useful for the analysis and numerical evaluation of the quantization process.

III. CSI ACQUISITION STRATEGIES

The CSI is acquired based on the estimation of known pilots transmitted by the users. In this case, the k-th user transmits $\sqrt{\tau_p}\boldsymbol{\varphi}_k$ as its pilot, where a specific random sequence $\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau_p\times 1}$ is taken from an orthonormal basis with $|\langle\boldsymbol{\varphi}_k, \boldsymbol{\varphi}_k'\rangle| = \delta_{kk'}$ and $\|\boldsymbol{\varphi}_k\|^2 = 1$. The sequence length τ_p is assumed to be less than or equal to the coherence interval τ_c. The m-th AP observes the received pilot $\mathbf{y}_m$ from all K users as

$$\mathbf{y}_{p,m} = \sqrt{\tau_p\rho_p}\sum_{k=1}^{K} g_{mk}\boldsymbol{\varphi}_k + \mathbf{w}_p, \tag{14}$$

where ρ_p is the transmit SNR of the pilot and $\mathbf{w}_p \sim \mathcal{CN}(0, \mathbf{I}_K)$ is an additive noise vector with zero mean and identity covariance. To ensure that all pilots are orthogonal for all K users, one should only allow $K \le \tau_p$ users who transmit their pilots simultaneously. In this case, the transmitted pilots satisfy

$$\Theta^H\Theta = \tau_p\rho_p\mathbf{I}_K, \text{ where } \Theta = \sqrt{\tau_p\rho_p}[\boldsymbol{\varphi}_1,\dots,\boldsymbol{\varphi}_K]. \tag{15}$$

In the ideal case of perfect fronthaul [1] the channel g_{mk} can be estimated at the AP and sent to the CPU which then has the global CSI. In this case, the received pilot $\mathbf{y}_{p,m}$ at the m-th AP is projected onto $\boldsymbol{\varphi}_k^H$ giving:

$$\begin{aligned} r_{p,mk} &= \boldsymbol{\varphi}_k^H\mathbf{y}_{p,m}\\ &= \sqrt{\tau_p\rho_p}g_{mk} + \sqrt{\tau_p\rho_p}\sum_{k'\neq k}^{K} g_{mk'}\boldsymbol{\varphi}_k^H\varphi_k' + \boldsymbol{\varphi}_k^H\mathbf{w}_p. \qquad (16)\end{aligned}$$

To obtain the estimate of g_{mk} we use the Linear Minimum Mean Squared Error (LMMSE) estimator given by

$$\hat{g}_{mk} = c_{mk}r_{p,mk}. \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2920070, IEEE Transactions on Vehicular Technology

We choose c_{mk} that minimizes the Mean Squared Error (MSE)

$$\epsilon_{mk} = \mathbb{E}\{|g_{mk} - \hat{g}_{mk}|^2\}. \tag{18}$$

The unique minimum is obtained by taking the derivative of ϵ_{mk} and setting it equal to zero giving

$$\begin{aligned} c_{mk} &= \frac{\mathbb{R}\mathrm{e}\{\mathbb{E}\{r^*_{p,mk} g_{mk}\}\}}{\mathbb{E}\{|r_{p,mk}|^2\}} \\ &= \frac{\sqrt{\tau_p \rho_p}\beta_{mk}}{\tau_p \rho_p \sum_{k'=1}^{K} \beta_{mk'} |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + 1}, \end{aligned} \tag{19}$$

where the last equation follows from (16). With the optimal coefficient c_{mk} the minimum mean squared error is then given by

$$\epsilon_{mk} = \mathbb{E}\{|g_{mk}|^2\} - \frac{(\mathbb{R}\mathrm{e}\{\mathbb{E}\{r^*_{p,mk} g_{mk}\}\})^2}{\mathbb{E}\{|r_{p,mk}|^2\}} = \beta_{mk} - \gamma_{mk}, \tag{20}$$

where we use γ_{mk} to denote the mean squared of the channel estimate given by

$$\begin{aligned} \gamma_{mk} \triangleq \mathbb{E}\{|\hat{g}_{mk}|^2\} &= c^2_{mk} \mathbb{E}\{|r_{p,mk}|^2\} \\ &= c_{mk}\, \mathbb{R}\mathrm{e}\{\mathbb{E}\{r^*_{p,mk} g_{mk}\}\} \\ &= \frac{\tau_p \rho_p \beta^2_{mk}}{\tau_p \rho_p \sum_{k'=1}^{K} \beta_{mk'} |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + 1}. \end{aligned} \tag{21}$$

We suppose that the CSI is transferred to the CPU in the same time frame as the uplink data. To possibly maximize the rate, the same proportion of power is allocated to pilot and to data as in the training-based scheme of general MIMO system [8]. Let ρ and ρ_u denote the total transmit SNR and the transmit Signal to Noise Ratio (SNR) for the uplink data respectively, the power allocation for pilot of length τ_p and for data of length τ_u follows

$$\rho_u \tau_u = \frac{\rho \tau_c}{2} \text{ and } \rho_p \tau_p = \frac{\rho \tau_c}{2}, \text{ where } \tau_c = \tau_p + \tau_u. \tag{22}$$

A. Estimate-and-Quantize

In this scheme we estimate the channel coefficient g_{mk} first as given in (17). So that it may be sent via limited fronthaul to the CPU, the estimated channel $\hat{g}_{mk}$ is quantized at each AP. Because we send the quantized version $\hat{g}^{eq}_{mk}$ to the CPU, the amount of CSI overhead resulted by this scheme is proportional to the number of users K. For symbol frame of length τ_c the portion of CSI overhead is then K/τ_c. After transferring via the fronthaul the CPU receives $\hat{g}^{eq}_{mk}$, which can be decomposed by Bussgang as

$$\hat{g}^{eq}_{mk} = Q(\hat{g}_{mk}) = \alpha_{eq}\hat{g}_{mk} + d_{eq}. \tag{23}$$

The mean squared error after quantization is given by

$$\epsilon^{eq}_{mk} = \mathbb{E}\{|g_{mk} - \hat{g}^{eq}_{mk}|^2\} \tag{24}$$

$$= \mathbb{E}\{|g_{mk}|^2\} + \mathbb{E}\{|\hat{g}^{eq}_{mk}|^2\} - 2\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{g^*_{mk}\hat{g}^{eq}_{mk}\}\}. \tag{25}$$

We can apply (23) to express

$$\begin{aligned} \mathbb{E}\{g^*_{mk}\hat{g}^{eq}_{mk}\} &= \alpha_{eq}\mathbb{E}\{g^*_{mk}\hat{g}_{mk}\} + \mathbb{E}\{g^*_{mk} d_{eq}\} \\ &= \alpha_{eq}\mathbb{E}\{g^*_{mk}\hat{g}_{mk}\}, \end{aligned} \tag{26}$$

where the second term vanishes because $g*_{mk}$ is uncorrelated with d_{eq}. This follows because $\mathbb{E}\{\hat{g}_{mk} d_{eq}\} = 0$ and our use of a linear MMSE estimator means that the estimation error prior to quantization is also uncorrelated with $\hat{g}_m k$ and hence also with d_eq. We then obtain

$$\begin{aligned} \epsilon^{eq}_{mk} &= \mathbb{E}\{|g_{mk}|^2\} + \lambda_{eq}\mathbb{E}\{|\hat{g}_{mk}|^2\} - 2\alpha_{eq}\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{g^*_{mk}\hat{g}_{mk}\}\}. \\ &= \mathbb{E}\{|g_{mk}|^2\} + \lambda_{eq}\gamma_{mk} - 2\alpha_{eq}\gamma_{mk} \\ &= \beta_{mk} - (2\alpha_{eq} - \lambda_{eq})\gamma_{mk}. \end{aligned} \tag{27}$$

Note that in the practical implementation of this scheme the channel estimation does not have to be performed at low resolution: the channel can be estimated at the AP at high precision, in the same way as CSI quantization in the CoMP scenario, and the estimate subsequently quantized at a lower resolution, in order to reduce the fronthaul load.

B. Quantize-and-Estimate

Unlike the previous scheme, here we quantize the pilot first and send it to the CPU to estimate g_{mk}. In this case, at the CPU we have the quantized received pilots which once again may be decomposed using the Bussgang decomposition as

$$\mathbf{y}^q_{p,m} = Q(\mathbf{y}_{p,m}) = \alpha_{qe}\mathbf{y}_{p,m} + \mathbf{d}_{qe}. \tag{28}$$

The noisy quantized observation at the CPU is given as

$$\begin{aligned} r^q_{p,mk} &= \boldsymbol{\varphi}_k^H \mathbf{y}^q_{p,m} = \alpha_{qe}\boldsymbol{\varphi}_k^H \mathbf{y}_{p,m} + \boldsymbol{\varphi}_k^H \mathbf{d}_{qe} \\ &= \alpha_{qe} r_{p,mk} + \boldsymbol{\varphi}_k^H \mathbf{d}_{qe}. \end{aligned} \tag{29}$$

We then apply the LMMSE estimator to obtain the quantize-and-estimate channel coefficient $\hat{g}^{qe}_{mk}$ given by

$$\hat{g}^{qe}_{mk} = c^{qe}_{mk} r^q_{p,mk}, \tag{30}$$

where we choose c^{qe}_{mk} that minimizes the MSE $\mathbb{E}\{|g_{mk} - \hat{g}^{qe}_{mk}|^2\}$. As derived in Appendix A the coefficient c^{qe}_{mk} is given by

$$c^{qe}_{mk} = c_{mk}\frac{\alpha_{qe} a_{mk}}{\alpha^2_{qe} a_{mk} + (\lambda_{qe} - \alpha^2_{qe}) b_m}, \text{ where} \tag{31}$$

$$a_{mk} \triangleq \tau_p \rho_p \sum_{k'=1}^{K} \beta_{mk'} |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + 1, \text{ and } b_m \triangleq \rho_p \sum_{k=1}^{K} \beta_{mk} + 1.$$

We then also obtain the MSE for the QE scheme expressed as (see Appendix A)

$$\begin{aligned} \epsilon^{qe}_{mk} &= \mathbb{E}\{|g_{mk} - \hat{g}^{qe}_{mk}|^2\} \\ &= \beta_{mk} - \left(\frac{\alpha^2_{qe} a_{mk}}{\alpha^2_{qe} a_{mk} + (\lambda_{qe} - \alpha^2_{qe}) b_m}\right)\gamma_{mk}. \end{aligned} \tag{32}$$

Because we quantize the received pilot in this scheme, the amount of the resulted CSI overhead is proportional to the length of pilot τ_p which doesn't scale directly with the number of users. However, in the case of orthogonal pilots the EQ and QE scheme have the same amount of CSI overhead. In terms of complexity at AP, this scheme apparently has lower complexity than EQ scheme bacause no estimator and only a single quantizer are needed.

IV. THE ACHIEVABLE RATES WITH COARSE QUANTIZATION

The uplink data received at all M APs may be described by

$$\mathbf{y}_d = \sqrt{\rho_u}\mathbf{G}\mathbf{x}_d + \mathbf{w}_d, \tag{33}$$

where $\mathbf{x}_d \in \mathbb{C}^K$ is the transmitted data from all K users, $\mathbf{G}$ is the channel matrix defined in (2) and $\mathbf{w}_d \sim \mathcal{CN}(0, \mathbf{I}_M)$ is an additive noise vector. After quantization and transmission via the fronthaul the CPU obtains the data signal $\mathbf{r}_d$, which can also be decomposed as

$$\begin{aligned} \mathbf{r}_d &= Q(\mathbf{y}_d) = \alpha_{qd}\mathbf{y}_d + \mathbf{d}_{qd} \\ &= \sqrt{\rho_u}\alpha_{qd}\mathbf{G}\mathbf{x}_d + \alpha_{qd}\mathbf{w}_d + \mathbf{d}_{qd} \\ &= \sqrt{\rho_u}\alpha_{qd}\hat{\mathbf{G}}\mathbf{x}_d + \sqrt{\rho_u}\alpha_{qd}\tilde{\mathbf{G}}\mathbf{x}_d + \alpha_{qd}\mathbf{w}_d + \mathbf{d}_{qd}, \end{aligned} \tag{34}$$

where $\tilde{\mathbf{G}}$ is the channel estimation error including the quantization error. Further, we treat $\hat{\mathbf{G}}$ as the true channel and treat

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2920070, IEEE Transactions on Vehicular Technology

the second term and so forth as an effective noise $\mathbf{z}$ such that

$$\mathbf{r}_d = \sqrt{\rho_u}\alpha_{qd}\hat{\mathbf{G}}\mathbf{x}_d + \mathbf{z} \tag{35}$$

To detect the transmitted data we can use a ZF detection matrix $\bar{\mathbf{A}}^H = (\hat{\mathbf{G}}^H\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H$, with $\bar{\mathbf{A}}^H\hat{\mathbf{G}} = \mathbf{I}_K$. We then obtain the estimated data as

$$\hat{\mathbf{x}}_d = \bar{\mathbf{A}}^H\mathbf{r}_d = \sqrt{\rho_u}\alpha_{qd}\mathbf{x}_d + \bar{\mathbf{A}}^H\mathbf{z}, \tag{36}$$

such that the SINR for the k-th user is given by

$$\text{SINR}_k^{ZF} = \frac{\rho_u\alpha_{qd}^2}{\left[\mathbb{E}\{\bar{\mathbf{A}}^H\mathbf{z}\mathbf{z}^H\bar{\mathbf{A}}\}\right]_{k,k}}. \tag{37}$$

However, due to the nature of the matrix $\mathbf{G}$ in the case of distributed massive MIMO, which tends to have independent large scale fading coefficients, the closed form expression of Signal to Interference Noise Ratio (SINR) for ZF is intractable. To obtain the SINR expression for our quantized CF massive MIMO we follow the approximation derived in [9]. We apply $\hat{\mathbf{G}}$ in our zero forcing detector to detect the data from (35) and apply a filter $\mathbf{\Lambda}^{-1/2}$ to $\mathbf{r}_d$ to whiten $\mathbf{z}$, such that we have a ZF detector matrix

$$\mathbf{A}^H = (\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1/2}, \text{ where} \tag{38}$$

$$\mathbf{\Lambda} = \mathbb{E}\{\mathbf{z}\mathbf{z}^H\} \text{ and } \mathbf{A}^H\mathbf{\Lambda}^{-1/2}\hat{\mathbf{G}} = \mathbf{I}_K. \tag{39}$$

After detection we obtain

$$\begin{aligned}\hat{\mathbf{x}}_d &= \mathbf{A}^H\mathbf{\Lambda}^{-1/2}\mathbf{r}_d \\ &= \sqrt{\rho_u}\alpha_{qd}\mathbf{x}_d + (\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\mathbf{z}.\end{aligned} \tag{40}$$

The instantaneous SINR (i.e. the SINR for a specific realization of $\mathbf{z}$) for the k-th user can then be expressed as

$$\text{SINR}_k^{ZF} = \frac{\rho_u\alpha_{qd}^2}{\left[(\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\mathbf{z}\mathbf{z}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}}(\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\right]_{k,k}} \tag{41}$$

Following [9] we may approximate $\mathbf{z}\mathbf{z}^H$ in (41) by its expectation $\mathbf{\Lambda}$ such that it remains

$$\text{SINR}_k^{ZF} \overset{(39)}{\approx} \frac{\rho_u\alpha_{qd}^2}{\left[(\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\right]_{k,k}} \tag{42}$$

In this way, we can express the SINR_k^{ZF} as [9]

$$\text{SINR}_k^{ZF} \approx \rho_u\alpha_{qd}^2\left(\frac{M-K+1}{M}\right)\hat{g}_k^H\mathbf{\Lambda}^{-1}\hat{g}_k. \tag{43}$$

Due to the independent realization of the additive noise and estimation error at each AP we may assume that the effective noise $\mathbf{z}$ is uncorrelated over M APs. Thus, the matrix $\mathbf{\Lambda}$ is a diagonal matrix given by

$$\mathbf{\Lambda} = \text{diag}\{\Lambda_1, ..., \Lambda_M\} \text{ and } \Lambda_m = \sigma_{d_{qd}}^2 + \alpha_{qd}^2\sigma_n^2 + \rho_u\alpha_{qd}^2\sum_{k=1}^{K}\epsilon_{mk}^q,$$

where $\sigma_{d_{qd}}^2$ is the distortion variance resulted from quantizing data, σ_n^2 is the noise variance and $\epsilon_{mk}^q \in \{\epsilon_{mk}^{eq}, \epsilon_{mk}^{qe}\}$ is the estimation error from (27) or (32) depending on the scheme. The achievable rate per user in the uplink is then given by

$$R_{u,k}^{ZF} = \log_2\left(1 + \text{SINR}_k^{ZF}\right). \tag{44}$$

V. NUMERICAL RESULTS

In the following, we provide some numerical results for the considered schemes above. We do simulations with system parameters similar to [1] where there are $M = 200$ APs and $K = 20$ users distributed uniformly in an area of 1×1 km^2. We assume that this simulation area is wrapped around to avoid the boundary effects. For the channel g_{mk} given in (1) we model the large scale fading $\beta_{mk} = \text{PL}_{mk} \cdot 10^{(\sigma_{sh}z_{mk})/10}$, where the factor $10^{(\sigma_{sh}z_{mk})/10}$ is the uncorrelated shadowing with the standard deviation $\sigma_{sh} = 8$ dB and $z_{mk} \sim \mathcal{N}(0, 1)$. The path loss coefficient follows the three-slope model according to

$$\text{PL}_{mk} = \begin{cases} -\mathcal{L} - 35\log_{10}(d_{mk}), d_{mk} > d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) - 20\log_{10}(d_{mk}), d_0 < d_{mk} \leq d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) - 20\log_{10}(d_0), d_{mk} \leq d_0, \end{cases}$$

where d_{mk} is the distance between the m-th AP and the k-th user, $d_0 = 0.01$km, $d_1 = 0.05$km, and

$$\begin{aligned}\mathcal{L} \triangleq\ & 46.3 + 33.9\log_{10}(f) - 13.83\log_{10}(h_{AP}) \\ & - (1.1\log_{10}(f) - 0.7)h_u + (1.56\log_{10}(f) - 0.8).\end{aligned} \tag{45}$$

We choose the carrier frequency $f = 1.9$ GHz, the AP antenna height $h_{AP} = 15$m and the user antenna height $h_u = 1.65$m. In our simulation the normalized transmit SNRs ρ_u and ρ_p are defined as the transmit power divided by the noise power which is $B \times k_b \times T_0 \times$ noise figure. We suppose that the bandwidth $B = 20$ MHz, the Boltzmann constant $k_b = 1.381 \times 10^{-23}$, the noise temperature $T_0 = 290$ Kelvin and the noise figure $= 9$ dB. To make a fair comparison, our simulation considers the orthogonal case with $\tau_p = K$ where EQ and QE scheme spend the same length of CSI overhead. We allocate 10% of symbols for acquiring CSI where $\tau_p = 20$ symbols are spent for the pilot from overall $\tau_c = 200$.

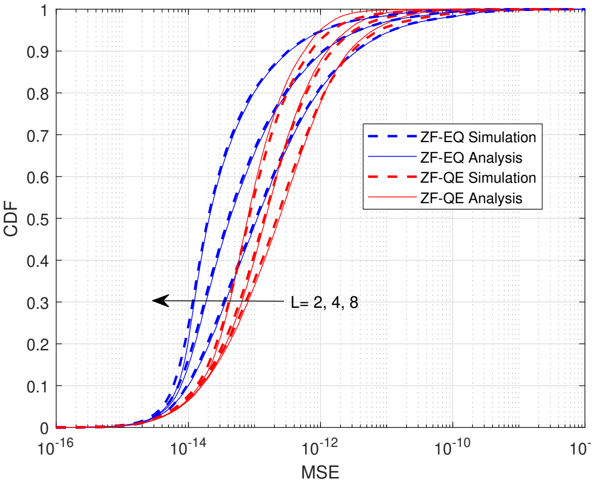

Fig. 1. The cumulative distribution of the channel estimation MSE ϵ_{mk}^q for the schemes estimate-and-quantize (EQ) in (27) and quantize-and-estimate (QE) in (32) with $K = 20$, $M = 200$ and Transmit Power $= 0$ dBW

In Fig 1, we first validate with simulations our analytical MSE approximations which are obtained in (27) and (32) using Bussgang decomposition. Note that in our simulation setup the large scale fading has very small value up to -17 order of magnitude. This boils down to very small channel gain and to very small typical value of MSE. It is shown in Fig. 1 that our analyses for both strategies are quite close to simulations especially for small L and high transmit power. In at least 80% of cases (those with the lower MSE) the QE scheme gives a poorer MSE than EQ and for $L > 2$ this proportion increases. However, it is the larger channel estimate errors that have stronger influence on the rate.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2920070, IEEE Transactions on Vehicular Technology

Using the corresponding channel estimation errors we then evaluate the average achievable rates per user given in (44). In this case, we compare their performance in terms of their per-user net throughput defined as

$$S_{u,k}^{ZF} \triangleq B\frac{1-\tau_p/\tau_c}{2}R_{u,k}^{ZF}, \tag{46}$$

where the CSI overhead is taken into account by the term $1-\tau_p/\tau_c$. As shown in Fig. 2 the QE scheme achieves higher throughput than the EQ scheme for small L over the whole range of transmit power. The performance gap is decreasing as we increase the quantization level. For small L, the achievable rates computed by our approximation (43) has only relatively small deviation from the rate computed by (37). It can also clearly be observed that ZF with low quantization level $L = 4$ can already outperform MRC even with infinite quantization precision. This demonstrates the great improvement resulting from having global CSI available at the CPU. With $L = 32 = 5$-bits we are about 5 dB away from ZF with ideal fronthaul to reach 60 Mbits/s/Hz average throughput per user. Meanwhile, the trade off between the increasing throughput and the resulting latency due to CSI overhead is left for future works.

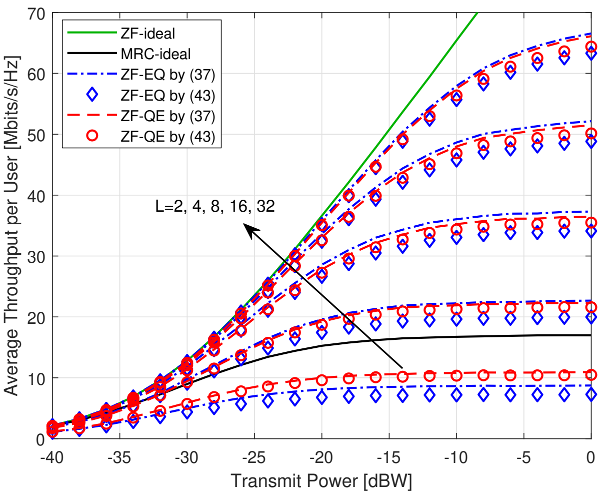

Fig. 2. The average per user throughput for different number of quantization level L, transmit power and CSI acquisition schemes for $K = 20$ and $M = 200$.

VI. CONCLUSION

This paper shows the benefit of having global CSI at the CPU for the uplink of cell-free massive MIMO. We have established the MSE expression of CSI-acquisition strategies and compared their performance. We have presented their corresponding average throughput for ZF detection. In this case, the low-complexity scheme ZF-QE outperforms ZF-EQ at low resolution especially for 1-bit.

APPENDIX A
DERIVATION OF EQUATION (31) AND (32)

For the estimator in (31) we have c_{mk}^{qe} that minimizes the MSE given by

$$c_{mk}^{qe} = \frac{\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,mk}^{q^*}g_{mk}\}\}}{\mathbb{E}\{|r_{p,mk}^{q}|^2\}}. \tag{47}$$

From (29) we can express the numerator of c_{mk}^{qe} as

$$\begin{aligned}\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,mk}^{q^*}g_{mk}\}\} &= \alpha_{qe}\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,mk}^{*}g_{mk}\}\} + \mathbb{R}\mathrm{e}\{\mathbb{E}\{\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}g_{mk}\}\}\\ &= \alpha_{qe}\sqrt{\tau_p\rho_p}\beta_{mk},\end{aligned} \tag{48}$$

where the second term vanishes due to uncorrelation. Likewise we can express the denominator as

$$\mathbb{E}\{|r_{p,mk}^{q}|^2\} = \alpha_{qe}^2\mathbb{E}\{|r_{p,mk}|^2\} + \mathbb{E}\{|\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}|^2\}, \tag{49}$$

where the first term is given by

$$\alpha_{qe}^2\mathbb{E}\{|r_{p,mk}|^2\} = \alpha_{qe}^2\left(\tau_p\rho_p\sum_{k'=1}^{K}\beta_{mk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2 + 1\right) \tag{50}$$

and the second term is given by

$$\begin{aligned}\mathbb{E}\{|\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}|^2\} = \|\boldsymbol{\varphi}_k^H\|^2\mathbb{E}\{|\mathbf{d}_{qe}|^2\} &\overset{(13)}{=} (\lambda_{qe}-\alpha_{qe}^2)\mathbb{E}\{|\mathbf{y}_{p,m}|^2\}\\ &= (\lambda_{qe}-\alpha_{qe}^2)\left(\rho_p\sum_{k=1}^{K}\beta_{mk}+1\right).\end{aligned} \tag{51}$$

Let a_{mk} and b_m denote the following expressions

$$a_{mk}\triangleq\tau_p\rho_p\sum_{k'=1}^{K}\beta_{mk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1,\ \text{and}\ b_m\triangleq\rho_p\sum_{k=1}^{K}\beta_{mk}+1,$$

then we obtain

$$\begin{aligned}c_{mk}^{qe} &= \frac{\alpha_{qe}\sqrt{\tau_p\rho_p}\beta_{mk}}{\alpha_{qe}a_{mk}}\frac{\alpha_{qe}a_{mk}}{\alpha_{qe}^2a_{mk}+(\lambda_{qe}-\alpha_{qe}^2)b_m}\\ &= c_{mk}\frac{\alpha_{qe}a_{mk}}{\alpha_{qe}^2a_{mk}+(\lambda_{qe}-\alpha_{qe}^2)b_m}.\end{aligned} \tag{52}$$

Further, we have the MSE given by

$$\epsilon_{mk}^{qe} = \mathbb{E}\{|g_{mk}|^2\} - \frac{(\mathbb{E}\{r_{p,mk}^{q^*}g_{mk}\})^2}{\mathbb{E}\{|r_{p,mk}^{q}|^2\}}, \tag{53}$$

where the second term can also be expressed as

$$\begin{aligned}\gamma_{mk}^{qe} &= \frac{\alpha_{qe}^2\tau\rho_p\beta_{mk}^2}{\alpha_{qe}^2a_{mk}}\frac{\alpha_{qe}^2a_{mk}}{\alpha_{qe}^2a_{mk}+(\lambda_{qe}-\alpha_{qe}^2)b_m}\\ &= \gamma_{mk}\frac{\alpha_{qe}^2a_{mk}}{\alpha_{qe}^2a_{mk}+(\lambda_{qe}-\alpha_{qe}^2)b_m}.\end{aligned} \tag{54}$$

REFERENCES

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, March 2017.

[2] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and Power Optimization in Cell-Free Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, July 2017.

[4] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 102–111, February 2011.

[5] J. Bussgang, "Crosscorrelation functions of amplitude-distorted gaussian signals," *RLE Technical Reports*, vol. 216, 1952.

[6] P. Zillmann, "Relationship Between Two Distortion Measures for Memoryless Nonlinear Systems," *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 917–920, Nov. 2010.

[7] A. Burr, M. Bashar, and D. Maryopi, "Cooperative Access Networks: Optimum Fronthaul Quantization in Distributed Massive MIMO and Cloud RAN," *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018.

[8] B. Hassibi and B. M. Hochwald, "How Much Training is Needed in Multiple-Antenna Wireless Links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, April 2003.

[9] R. Senanayake, A. Lozano, P. Smith, and J. Evans, "Analytical handle for ZF reception in distributed massive MIMO," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 16–20.