QUANTIZATION-AWARE PRECODING FOR MU-MIMO WITH LIMITED-CAPACITY FRONTHAUL

Yasaman Khorsandmanesh, Emil Björnson, and Joakim Jaldén

School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. Email: {yasamank, emilbjo, jalden}@kth.se

ABSTRACT

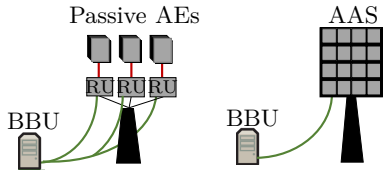
Base stations in 5G and beyond use advanced antenna systems (AASs), where multiple passive antenna elements and radio units are integrated into a single box. A critical bottleneck of such a system is the digital fronthaul between the AAS and baseband unit (BBU), which has limited capacity. In this paper, we study an AAS used for precoded downlink transmission over a multi-user multiple-input multiple-output (MU-MIMO) channel. First, we present the baseline quantization-unaware precoding scheme created when a precoder is computed at the BBU and then quantized to be sent over the fronthaul. We propose a new precoding design that is aware of the fronthaul quantization. We formulate an optimization problem to minimize the mean squared error at the receiver side. We rewrite the problem to utilize mixed-integer programming to solve it. The numerical results manifest that our proposed precoding greatly outperforms quantization-unaware precoding in terms of sum rate.

Index Terms— Quantization-aware precoding, Advanced antenna system, MU-MIMO, limited fronthaul.

1. INTRODUCTION

Multi-user multiple-input multiple-output (MU-MIMO) is a classical technology for spatial multiplexing of user equipments (UEs) on the same time-frequency resource by utilizing a base station (BS) with multiple antennas [1, 2]. MU-MIMO has been supported by several standards, but 5G networks are the first to make widespread use of it [3]. A key reason for the slow adoption is that a traditional BS contains one baseband unit (BBU) and then two boxes per antenna: one passive antenna element (AE) and one radio unit (RU), as depicted in Fig. 1(a). However, 5G BSs integrate all AEs and RUs into a single enclosure, called an *advanced antenna system (AAS)* [4, Chapter 1] shown in Fig. 1(b). This hardware evolution has made massive MU-MIMO practically feasible [5] and enables the BBU to be virtualized in the cloud. A new implementation bottleneck is the digital fronthaul between the AAS and BBU, which needs a capacity proportional to the number of antennas. This interface must carry received uplink signals (to be decoded at the BBU) and precoded downlink signals, which are computed at the BBU. In this paper, we propose a new linear block-level precoding technique that is aware of the necessary fronthaul quantization.

This work was supported by the Knut and Alice Wallenberg Foundation. E. Björnson is also with Department of Electrical Engineering, Linköping University, Linköping, Sweden.



(a) The traditional BS (b) The 5G BS

Fig. 1. The base stations' configuration (a) shows the analog connection and (b) digital fronthaul.

1.1. Related Works

The impact of impairments in analog hardware on MU-MIMO systems has received much attention in prior literature (see e.g., [6–10]). There are related works on quantization distortion caused by low-resolution analog-to-digital converters in the uplink [11, 12] and low-resolution digital-to-analog converters (DAC) in the downlink [13–15]. A key characteristic of these prior works is that the distortion is created in the RU, the analog domain, or the converters. This implies that the transmit signal obtained after precoding is distorted. One way to handle low-resolution DACs is to perform symbol-level precoding [16, 17], where a new precoder is selected for each symbol vector to minimize signal distortion. This approach requires much more fronthaul signaling and computational complexity than symbol-level precoding. Another related line of work is quantized feedback [18, 19], where UEs estimate and feed back their channels to the BS. The precoding is computed based on the quantized channels, which leads to extra unremovable interference even if zero-forcing is used [19], but the key difference is that the quantization appears before the precoding. The effect of limited fronthaul capacity is studied in [20–22] among others, but the focus was not on precoding design. To our knowledge, this is the first paper to analyze the precoding distortion that occurs over the digital interface between the BBU and AAS, where the key difference is that the precoding matrix is quantized before the transmit signal is computed; that is, before the quantized precoding matrix is multiplied with the data symbols at the AAS. An extended version of this paper is available in [23].

1.2. Contributions

In this paper, we consider MU-MIMO precoding quantization over a limited-capacity fronthaul connection. Since the data symbols originate from a finite-resolution codebook and change much more rapidly than the precoding matrix, an efficient implementation will send these quantities separately over the fronthaul so that only the precoder is quantized.

We formulate and solve a novel quantization-aware precoding problem, where the communication performance after quantization is maximized. The main contributions are:

- Inspired by practical AAS implementation, we formulate a new downlink precoding framework where the precoding matrix is quantized when sent over the limited-capacity fronthaul from the BBU to the AAS, while data symbols are inherently quantized.
- We formulate a quantization-aware precoding problem, where the precoder is selected to minimize the mean-squared error (MSE) at the receiver side. We use mixed-integer programming (MIP) [24, 25] to solve it to global optimality for a fixed precoding coefficient.
- We provide numerical results to show the benefits of quantization-aware precoding over the quantization-unaware baseline. We describe how the number of quantization levels and UEs affect performance.

2. SYSTEM MODEL

We consider the downlink data transmission in a single-cell MU-MIMO system, where a BS equipped with an AAS with M antenna-integrated radios serves K single-antenna UEs on the same time-frequency resource. The AAS is connected to a BBU through a digital fronthaul link with limited capacity, as depicted in Fig. 1(b). Hence, any signal that is sent over the fronthaul must be quantized to finite resolution. Each transmitted signal vector is the product between a precoder matrix and a vector with data symbols, where the former is assumed fixed for the duration of the transmission while the latter changes at the symbol rate. The BBU computes a precoder based on channel state information (CSI) and then forwards it to the AAS. As the data symbols represent bit sequences from a codebook, we can send them without quantization errors and map them to modulation symbols at the AAS. However, the precoder matrix normally contains arbitrary complex-valued entries and must be quantized before being sent over the limited-capacity fronthaul. The quantized matrix is then multiplied with the UEs' data symbols at the AAS, and finally the product is transmitted wirelessly.

Before analyzing the proposed quantization-aware precoder in Section 3, in the following subsections, we first introduce the considered channel model. Then, conventional uniform quantizer-mapping and the quantization operator are described.

2.1. Channel Model

As the main focus of this work is on quantization effects, we assume the BBU has perfect CSI and neglect all other potential hardware impairments. The imperfect CSI case will be considered in future work. The downlink system model can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{P}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} = [y_1, \ldots, y_K]^{\mathrm{T}} \in \mathbb{C}^K$ contains the received signals at all UEs and $y_k \in \mathbb{C}$ denotes the signal received at the k-th UE. The downlink channel matrix $\mathbf{H} \in \mathbb{C}^{K \times M}$ has entries $h_{k,m}$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$. It represents a narrowband channel and might be one subcarrier of a multi-carrier system. The vector $\mathbf{n} \in \mathbb{C}^K$ represents i.i.d. additive white complex Gaussian noise with zero mean and variance N_0. The vector $\mathbf{s} = [s_1, \ldots, s_K]^{\mathrm{T}} \in \mathcal{O}^K$ contains the information, where s_k denotes the random data symbol intended for UE k and is normalized to unit power. Here, $\mathcal{O}$ is the finite set of constellation points (e.g., a conventional QAM alphabet). The quantized precoder matrix is denoted by $\mathbf{P} \in \mathcal{P}^{M \times K}$, where the set of quantization alphabet $\mathcal{P}$ coincides with the complex numbers set $\mathbb{C}$ in the case of infinite resolution. We denote $\mathcal{L} = \{l_0, \ldots, l_{L-1}\}$ as the set of real-valued quantization labels. We assume that the same quantization alphabet is used for the real and imaginary parts. Under these assumptions, the entries of the precoded vector $\mathbf{P}$ are $p_{m,k} = l_R + j l_I$ where $l_R, l_I \in \mathcal{L}$. The number of quantization levels is denoted by $L = |\mathcal{L}|$, $N = \log_2(L)$ refers to the number of quantization bits per real dimension and the set of complex-valued precoder outputs for each antenna is $\mathcal{P} = \mathcal{L} \times \mathcal{L}$.

The precoder matrix $\mathbf{P}$ must satisfy the power constraint

$$\|\mathbf{P}\|_{\mathrm{F}}^2 \leq q, \tag{2}$$

where $\|\mathbf{P}\|_{\mathrm{F}}$ denotes the Frobenius norm and q is the maximum average transmit power of the downlink signals. To estimate the transmitted information symbol $\hat{s}_k \in \mathbb{C}$, we assume that the k-th UE scales the received signal y_k by the precoding factor $\beta \in \mathbb{C}$ as $\hat{s}_k = \beta y_k$ with the goal of minimizing the MSE $\mathbb{E}[\|\mathbf{s} - \hat{\mathbf{s}}\|^2]$, where $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_K]^{\mathrm{T}}$. Specifically, by using the same scalar for all UEs, we will equalize their performance [16].

2.2. Quantized Precoding

When it comes to quantized precoding, we will consider two cases. The naive baseline approach is to first design a precoder $\mathbf{P}_{\text{ideal}}$ based on CSI and then quantize it, which we call *quantization-unaware precoding*. We notice that there are prior works that consider a similar procedure [13, 16], but quantize the product $\mathbf{P}_{\text{ideal}}\mathbf{s}$ and not just the precoder $\mathbf{P}_{\text{ideal}}$. An alternative is to find the quantized precoder that minimizes the MSE, based on both CSI and knowledge of the quantization alphabet, which we call *quantization-aware precoding*. We will now present the quantization-unaware precoding approach and then the proposed quantization-aware precoding scheme will be described in Section 3.

In quantization-unaware precoding, the precoder matrix $\mathbf{P}$ is designed without taking the limited-capacity fronthaul effect into account. The precoder is designed based on CSI as if there would be no quantization and then a uniform quantization scheme is applied. The precoder matrix $\mathbf{P} = \alpha\hat{\mathbf{P}} = \alpha\mathcal{Q}(\mathbf{W})$ describes the limited-capacity fronthaul effect, where $\mathbf{W}$ is the preferred precoder matrix without quantization and $\mathcal{Q}(\cdot) : \mathbb{C}^{M \times K} \to \mathcal{P}^{M \times K}$ denotes the quantizer-mapping function. Since uniform quantization is normally used in practice, we model $\mathcal{Q}(\cdot)$ as a symmetric uniform quantizer with step size Δ. In general, the average power in the quantized signal is not preserved. As the condition $\|\mathbf{W}\|_{\mathrm{F}}^2 = q$ does not imply $\|\mathbf{P}\|_{\mathrm{F}}^2 = q$ and for assurance that the power constraint (2) is satisfied with equality, the output of the quantizer is scaled by a constant $\alpha = \sqrt{q/\|\hat{\mathbf{P}}\|_{\mathrm{F}}^2}$ at the AAS.

Each entry of the quantization labels $\mathcal{L}$ is defined as

$$l_z = \Delta\Big(z - \frac{L-1}{2}\Big), \quad z = 0, \ldots, L-1. \tag{3}$$

Furthermore, we let $\mathcal{T} = \{\tau_0, \ldots, \tau_L\}$, where $-\infty = \tau_0 <$

$\tau_1 < \ldots < \tau_{(L-1)} < \tau_L = \infty$, specify the set of $L+1$ quantization thresholds. For uniform quantizers, the quantization thresholds are

$$\tau_z = \Delta\Big(z - \frac{L}{2}\Big), \quad z = 1, \ldots, L. \tag{4}$$

The quantizer function $\mathcal{Q}(\cdot)$ can be uniquely described by the set of quantization labels $\mathcal{L} = \{l_z | z = 0, \ldots, L-1\}$ and the set of quantization thresholds $\mathcal{T}$. The quantizer maps $w_{m,k} \in \mathbb{C}$ into the quantized output $p_{m,k} = l_r + jl_i$ if $\Re\{w_{m,k}\} \in [\tau_r, \tau_{r+1})$ and $\Im\{w_{m,k}\} \in [\tau_i, \tau_{i+1})$. The step size Δ of the quantizers should be chosen to minimize the distortion between the quantized and unquantized vector. The optimal step size Δ depends on the minimum MSE distribution of the input, which in our case depends on the precoding scheme. Since the distribution of the precoder elements is generally unknown, we set the step size to minimize the distortion under the maximum-entropy assumption that the per-antenna input to the quantizers is $\mathcal{CN}(0, q/M)$ distributed. The corresponding optimal step size was found in [26].

3. QUANTIZATION-AWARE PRECODING

The quantization-unaware precoding scheme is clearly not the optimum quantized precoder since it neglects the quantization effect, which leads to extra interference and reduced beamforming gain. For example, canceling all interference using zero-forcing is optimal when the signal-to-noise ratio (SNR) is high and there is no quantization [27], while it will not be the case in our setup. In this section, we propose a scheme that finds an optimum quantization-aware precoder that minimizes the MSE between the received signal and the transmitted symbol vector $\mathbf{s}$ under the power constraint (2).

3.1. Optimal Quantized Precoding

We formulate the precoder optimization problem as

$$\begin{aligned} \underset{\mathbf{P} \in \mathcal{P}^{M\times K}, \beta \in \mathbb{C}}{\text{minimize}} \quad & \mathbb{E}[\|\mathbf{s} - \beta\mathbf{y}\|^2] \\ \text{subject to} \quad & \|\mathbf{P}\|_{\mathrm{F}}^2 \le q. \end{aligned} \tag{5}$$

For a given quantized precoder matrix $\mathbf{P}$, we can compute the optimal value of β by taking the Wirtinger derivative and equate it to zero, which yields

$$\beta^{\mathrm{Opt}} = \frac{\mathrm{tr}(\mathbf{P}^H\mathbf{H}^H)}{\mathrm{tr}(\mathbf{P}^H\mathbf{H}^H\mathbf{H}\mathbf{P}) + KN_0}. \tag{6}$$

If we substitute (6) into (5), our problem will be highly complex. An iterative method where we switch between optimizing $\mathbf{P}$ and β is possible, but we have noticed that β will not change much between iterations. Hence, we propose to pick a "good" β and then solve (5) once for that β. In particular, we will choose $\beta = \beta^{\mathrm{WF}}$, where WF refers to Wiener filter. The WF precoder minimizes the MSE with an *infinite-capacity fronthaul* [27] and is given by $\mathbf{P}^{\mathrm{WF}} = \bar{\alpha}\bar{\mathbf{W}}$, where $\bar{\mathbf{W}} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \frac{KN_0}{q}\mathbf{I}_K)^{-1}$ and $\bar{\alpha} = \sqrt{q/\mathrm{tr}(\bar{\mathbf{W}}\bar{\mathbf{W}}^H)}$. By substituting $\mathbf{P}^{\mathrm{WF}}$ into (6), we simplify the expression to [27]

$$\beta^{\mathrm{WF}} = \frac{1}{\sqrt{q}}\Big[\mathrm{tr}\Big(\mathbf{H}^H\Big(\mathbf{H}\mathbf{H}^H + \frac{KN_0}{q}\mathbf{I}_K\Big)^{-2}\mathbf{H}\Big)\Big]^{1/2}. \tag{7}$$

3.2. Optimization Solution

To find the optimal precoder for an arbitrary fixed β, first we simplify the objective function of problem (5) as

$$\begin{aligned} &\mathbb{E}[\|\mathbf{s} - \beta\mathbf{y}\|^2] = \mathbb{E}[\|\mathbf{s} - \beta\mathbf{H}\mathbf{P}\mathbf{s} - \beta\mathbf{n}\|^2] \\ &= \mathrm{tr}\Big((\mathbf{I}_K - \beta\mathbf{H}\mathbf{P})\mathbb{E}[\mathbf{s}\mathbf{s}^H](\mathbf{I}_K - \beta\mathbf{H}\mathbf{P})^H + \beta\beta^*\mathbb{E}[\mathbf{n}\mathbf{n}^H]\Big) \\ &= \mathrm{tr}\Big(\mathbf{I}_K - \beta\mathbf{H}\mathbf{P} - \beta^*\mathbf{P}^H\mathbf{H}^H + |\beta|^2\mathbf{P}^H\mathbf{H}^H\mathbf{H}\mathbf{P}\Big) + |\beta|^2KN_0 \\ &= \mathrm{tr}\Big(|\beta|^2\mathbf{P}^H\mathbf{H}^H\mathbf{H}\mathbf{P} - \beta\mathbf{H}\mathbf{P} - \beta^*\mathbf{P}^H\mathbf{H}^H\Big) + K(|\beta|^2N_0 + 1). \end{aligned} \tag{8}$$

To minimize (8) with respect to $\mathbf{P}$, we can drop the constant term $K(|\beta|^2N_0 + 1)$ and rewrite the relaxed version of (5) as

$$\begin{aligned} \underset{\mathbf{P} \in \mathcal{P}^{M\times K}}{\text{minimize}} \quad & \mathrm{tr}\Big(\mathbf{P}^H\mathbf{H}^H\mathbf{H}\mathbf{P} - \frac{1}{\beta^*}\mathbf{H}\mathbf{P} - (\frac{1}{\beta^*}\mathbf{H}\mathbf{P})^H\Big) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \le q. \end{aligned} \tag{9}$$

To turn (9) into a vector optimization problem, we set $\mathbf{a} = \mathrm{vec}(\mathbf{P}), \mathbf{h} = \mathrm{vec}(\frac{1}{\beta^*}\mathbf{H}^T)$, so we have

$$\begin{aligned} \underset{\mathbf{a} \in \mathcal{P}^{MK\times 1}}{\text{minimize}} \quad & \mathbf{a}^H\left(\mathbf{I}_K \otimes \mathbf{H}^H\mathbf{H}\right)\mathbf{a} - \mathbf{h}^T\mathbf{a} - \left(\mathbf{h}^T\mathbf{a}\right)^H \\ \text{subject to} \quad & \mathbf{a}^H\mathbf{a} \le q, \end{aligned} \tag{10}$$

where $\otimes$ denotes the Kronecker product. For solving the problem in the quantized domain and reusing standard MIP results, we should transfer (10) into an equivalent real-valued problem utilizing the following definitions:

$$\begin{aligned} &\mathbf{a}_{\mathbb{R}} = \begin{bmatrix}\Re\{\mathbf{a}\}\\ \Im\{\mathbf{a}\}\end{bmatrix}, \mathbf{c}_{\mathbb{R}} = \begin{bmatrix}\Re\{\mathbf{h}\}\\ \Im\{\mathbf{h}\}\end{bmatrix}, \text{and} \\ &\mathbf{V}_{\mathbb{R}} = \begin{bmatrix}\Re\{\mathbf{I}_K \otimes \mathbf{H}^H\mathbf{H}\} & -\Im\{\mathbf{I}_K \otimes \mathbf{H}^H\mathbf{H}\}\\ \Im\{\mathbf{I}_K \otimes \mathbf{H}^H\mathbf{H}\} & \Re\{\mathbf{I}_K \otimes \mathbf{H}^H\mathbf{H}\}\end{bmatrix}. \end{aligned} \tag{11}$$

These definitions enable us to rewrite (10) as

$$\begin{aligned} \underset{\mathbf{a}_{\mathbb{R}} \in \tilde{\mathcal{P}}^{2MK\times 1}}{\text{minimize}} \quad & \mathbf{a}_{\mathbb{R}}^T\mathbf{V}_{\mathbb{R}}\mathbf{a}_{\mathbb{R}} - 2\mathbf{c}_{\mathbb{R}}^T\mathbf{a}_{\mathbb{R}} \\ \text{subject to} \quad & \mathbf{a}_{\mathbb{R}}^T\mathbf{a}_{\mathbb{R}} \le q, \end{aligned} \tag{12}$$

where $\tilde{\mathcal{P}} = \mathcal{L}$ assure that we are not using more quantization steps than allowed. Both the objective function and constraint of (12) are convex function of $\mathbf{a}_{\mathbb{R}}$. Due to the $\mathbf{a}_{\mathbb{R}} \in \tilde{\mathcal{P}}^{2MK\times 1}$ criteria, the search domain of the problem is discrete. Numerical algorithms for solving such integer optimization problems to global optimality are well known, e.g., see [24]. Hence, we can use standard integer convex solvers to find the optimal solutions efficiently by defining following equivalent problem

$$\begin{aligned} \underset{\mathbf{x} \in \mathbb{Z}^{2MK\times 1}}{\text{minimize}} \quad & \mathbf{a}_{\mathbb{R}}^T\mathbf{V}_{\mathbb{R}}\mathbf{a}_{\mathbb{R}} - 2\mathbf{c}_{\mathbb{R}}^T\mathbf{a}_{\mathbb{R}} \\ \text{subject to} \quad & \mathbf{a}_{\mathbb{R}}^T\mathbf{a}_{\mathbb{R}} \le q, \\ & \mathbf{a}_{\mathbb{R}} = \Delta\Big(\mathbf{x} - (\frac{L-1}{2})\mathbf{1}_{2MK\times 1}\Big), \\ & \mathbf{0}_{2MK\times 1} \le \mathbf{x} \le (L-1)\mathbf{1}_{2MK\times 1}. \end{aligned} \tag{13}$$

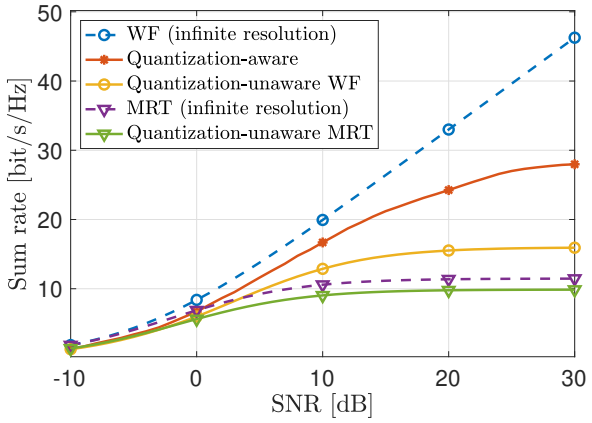


Fig. 2. Average achievable sum rate versus the SNR for different precoding schemes.

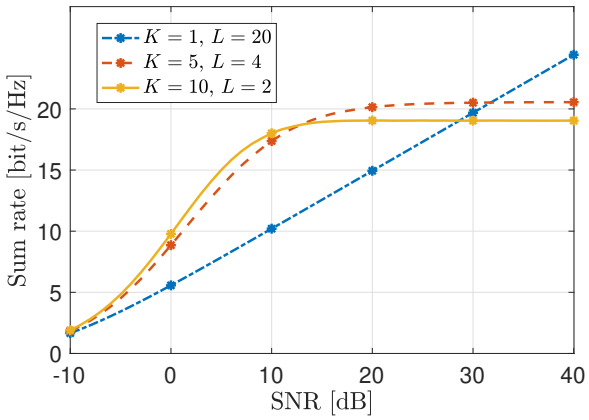


Fig. 3. Average achievable sum rate versus the SNR for a fixed number of BS antennas.

In the next section, we will use the Gurobi solver with CVX [25] to solve this problem. Although the complexity of the problem increases exponentially with MK, the numerical results show that it is still solvable for practically-sized MU-MIMO systems. The complexity of problem (13) is also exponential in the number of quantization levels L, but we consider a fixed and relatively small number of quantization bits.

4. NUMERICAL RESULTS

In this section, we compare the sum rates achieved by quantization-aware precoding and quantization-unaware precoding under different conditions. The entries of the channel matrix $\mathbf{H}$ are generated as independent circularly-symmetric complex Gaussian random variables with variance γ and the common SNR of all UEs is defined as $\text{SNR} = \frac{q\gamma}{N_0}$. The number of BS antennas is $M = 16$ and the number of UEs is $K = 4$. The sum rate is calculated using Monte Carlo simulations for the case of Gaussian signaling and perfect CSI at the receiver. We will compare different precoding schemes as a function of the SNR and the number of quantization levels L. We will compare quantization-based precoding with the infinite-resolution case. The classic WF precoding and maximum ratio transmission (MRT) schemes are considered.

Fig. 2 depicts the average sum rate as a function of the SNR for different precoding schemes: quantization-aware, quantization-unaware, and infinite resolution. The number of quantization levels is $L = 8$. The infinite-resolution WF precoding outperforms all the quantized precoding schemes and the gap increases linearly (in dB scale) at high SNR. However, the gap between the quantization-aware and infinite-resolution WF precoding is remarkably smaller than the gap between quantization-unaware WF and the infinite-resolution WF precoding. The quantization-aware and -unware MRT precoding have the same performance as WF at low SNR, but the lack of interference cancellation results in a large gap at higher SNRs. The gap between the MRT curves corresponds to the loss in beamforming gain due to quantization. We notice that for quantization-aware and quantization-unaware precoding, the sum rate converges to specific limits at high SNR, since all interference cannot be canceled due to the quantization effect; that is, the system is interference-limited at high SNR.

Similar to [19], the degrees-of-freedom of the considered MU-MIMO system is 1, due to the finite-resolution quantization. This implies that it is optimal to serve one UE at a time when the SNR is large. To demonstrate this, Fig. 3 presents the average achievable sum rate as a function of SNR for $M = 16$ BS antennas and different values of K and L, such that $K \cdot L = 20$. At low and medium SNR, the sum rate is maximized by serving many UEs. At high SNR, serving a lower number of UEs with a high-resolution quantizer outperforms the opposite case. This is because the system is heavily interference-limited, which can be partially resolved by increasing L. For each SNR value, there is an optimal number of UEs to serve and it is imperative to schedule the right number of UEs. Despite the fact that single-user transmission prevails at high SNR, MU-MIMO remains the preferable case in practice since the crossing point is at 20 bit/s/Hz, which would require enormous constellations for a single UE.

5. CONCLUSIONS AND FUTURE WORK

5G sites consist of an AAS connected to a BBU via a digital fronthaul with limited capacity. In the downlink, the finite-constellation data symbols can be sent to the AAS without quantization, but the precoder matrix must be quantized to finite precision. We have introduced the concept of novel quantization-aware precoding, where the BBU uses the quantizer structure to select the best finite-precision MU-MIMO precoder that requires no further quantization. In particular, we formulate the MSE-minimizing precoder and solve it by MIP. We have shown numerically that the proposed quantization-aware precoding outperforms the baseline quantization-unaware precoding, where the optimal precoding for the infinite-resolution case is selected and then quantized. The improved interference mitigation gives a large sum rate gain at medium and large SNRs, despite the fact that the degrees-of-freedom is limited to one.

We will develop lower-complexity quantization-aware precoders in future work to make the concept useful also in massive MU-MIMO scenarios.

6. REFERENCES

[1] S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, "The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems," *IEEE Trans. Veh. Technol.*, vol. 39, no. 1, pp. 56–67, Feb. 1990.

[2] D. Gesbert, M. Kountouris, R. W. Heath, C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 36–46, Sept. 2007.

[3] S. Parkvall, E. Dahlman, A. Furuskär, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Stand. Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[4] H. Asplund, D. Astely, P. von Butovitsch, T. Chapman, M. Frenne, F. Ghasemzadeh, M. Hagström, B. Hogan, G. Jöngren, J. Karlsson, et al., *Advanced Antenna Systems for 5G Network Deployments: Bridging the Gap Between Theory and Practice*, Academic Press, 2020.

[5] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, Nov. 2019.

[6] M. Wenk, *MIMO-OFDM testbed: Challenges, implementations, and measurement results*, Series in microelectronics. Hartung-Gorre, 2010.

[7] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[8] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[9] C. Mollén, U. Gustavsson, T. Eriksson, and E. G. Larsson, "Impact of spatial filtering on distortion from low-noise amplifiers in massive MIMO base stations," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6050–6067, Dec. 2018.

[10] S. R. Aghdam, S. Jacobsson, U. Gustavsson, G. Durisi, C. Studer, and T. Eriksson, "Distortion-aware linear precoding for massive MIMO downlink systems with nonlinear power amplifiers," *unpublished paper*, [Online]. Available: https://arxiv.org/pdf/2012.13337.pdf, 2020.

[11] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 1, pp. 87–100, Jan. 2016.

[12] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, June 2016.

[13] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[14] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019.

[15] O. Castañeda, S. Jacobsson, G. Durisi, T. Goldstein, and C. Studer, "Finite-alphabet wiener filter precoding for mmwave massive MU-MIMO systems," in *53rd Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, 2019, pp. 178–183.

[16] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *50th Asilomar Conference on Signals, Systems and Computers (ACSSC)*, Pacific Grove, CA, USA, 2016, pp. 763–767.

[17] C. G. Tsinos, A. Kalantari, S. Chatzinotas, and B. Ottersten, "Symbol-level precoding with low resolution DACs for large-scale array MU-MIMO systems," in *19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Kalamata, Greece, 2018, pp. 1–5.

[18] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[19] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Oct. 2006.

[20] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Adv. Singal Process.*, vol. 2009, pp. 1–10, June 2009.

[21] S.-H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[22] P. Parida, H. S. Dhillon, and A. F. Molisch, "Downlink performance analysis of cell-free massive MIMO with finite fronthaul capacity," in *88th Vehicular Technology Conference (VTC-Fall)*, Chicago, IL, USA, 2018, pp. 1–6.

[23] Y. Khorsandmanesh, E. Björnson, and J. Jaldén, "Optimized precoding for MU-MIMO with fronthaul quantization," *to be submitted*, 2022.

[24] L. A. Wolsey, "Mixed integer programming," *Wiley Encyclopedia of Computer Science and Engineering*, pp. 1–10, Mar. 2007.

[25] G. O. Gurobi, "Reference manual, Gurobi optimization," 2020.

[26] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 957–977, Mar. 2001.

[27] M. Joham, W. Utschick, and J. A. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.