

# Quantized Precoding for Massive MU-MIMO

Sven Jacobsson, *Student Member, IEEE*, Giuseppe Durisi, *Senior Member, IEEE*, Mikael Coldrey, *Member, IEEE*, Tom Goldstein, *Member, IEEE*, and Christoph Studer, *Senior Member, IEEE*

**Abstract**—Massive multiuser (MU) multiple-input multiple-output (MIMO) is foreseen to be one of the key technologies in fifth-generation wireless communication systems. In this paper, we investigate the problem of downlink precoding for a narrow-band massive MU-MIMO system with low-resolution digital-to-analog converters (DACs) at the base station (BS). We analyze the performance of linear precoders, such as maximal-ratio transmission and zero-forcing, subject to coarse quantization. Using Bussgang’s theorem, we derive a closed-form approximation on the rate achievable under such coarse quantization. Our results reveal that the performance attainable with infinite-resolution DACs can be approached using DACs having only 3 to 4 bits of resolution, depending on the number of BS antennas and the number of user equipments (UEs). For the case of 1-bit DACs, we also propose novel nonlinear precoding algorithms that significantly outperform linear precoders at the cost of an increased computational complexity. Specifically, we show that nonlinear precoding incurs only a 3 dB penalty compared to the infinite-resolution case for an uncoded bit error rate of  $10^{-3}$ , in a system with 128 BS antennas that uses 1-bit DACs and serves 16 single-antenna UEs. In contrast, the penalty for linear precoders is about 8 dB.

**Index Terms**—Massive multi-user multiple-input multiple-output, digital-to-analog converter, Bussgang’s theorem, minimum mean-square error precoding, convex optimization, semidefinite relaxation, Douglas-Rachford splitting, sphere precoding.

## I. INTRODUCTION

Massive multiuser (MU) multiple-input multiple-output (MIMO) wireless systems, where the base station (BS) is equipped with several hundreds of antenna elements, promises significant improvements in spectral efficiency, energy efficiency, reliability, and coverage compared to traditional cellular systems [1]–[3]. Increasing the number of radio frequency (RF) chains at the BS could, however, lead to significant increases in hardware complexity, system costs, and circuit power

S. Jacobsson is with Ericsson Research and Chalmers University of Technology, Gothenburg, Sweden (e-mail: sven.jacobsson@ericsson.com).

G. Durisi is with Chalmers University of Technology, Gothenburg, Sweden (e-mail: durisi@chalmers.se).

M. Coldrey is with Ericsson Research, Gothenburg, Sweden (e-mail: mikael.coldrey@ericsson.com)

T. Goldstein is with the Department of Computer Science, University of Maryland, College Park, MD (e-mail: tomg@cs.umd.edu).

C. Studer is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY (e-mail: studer@cornell.edu).

The work of S. Jacobsson and G. Durisi was supported by the Swedish Foundation for Strategic Research under grant ID14-0022, and by the Swedish Governmental Agency for Innovation Systems (VINNOVA) within the competence center ChaseOn.

The work of T. Goldstein was supported in part by the US National Science Foundation (NSF) under grant CCF-1535902 and by the US Office of Naval Research under grant N00014-17-1-2078.

The work of C. Studer was supported in part by Xilinx Inc. and by the US NSF under grants ECCS-1408006, CCF-1535897, and CAREER CCF-1652065.

The system simulator for the precoders studied in this paper is available on GitHub: [https://github.com/quantizedmassivemimo/1bit\\_precoding](https://github.com/quantizedmassivemimo/1bit_precoding).

consumption. Therefore, practical massive MU-MIMO systems may require low-cost and power-efficient hardware components at the BS. In this paper, we consider the downlink of massive MU-MIMO system, where the BS is equipped with low-resolution digital-to-analog converters (DACs) and transmits data to multiple, independent user equipments (UEs) in the same time-frequency resource.

For the quantization-free case (infinite-resolution DACs), the capacity region of the MU downlink Gaussian channel has been characterized in [4]–[7]. When channel state information (CSI) is known noncausally at the BS, dirty-paper coding (DPC) [8] is known to achieve the sum-rate capacity [6]. Several precoding algorithms to approach the DPC performance have been proposed (see, e.g., [9]–[12]). Most of these precoding methods are, however, computationally demanding, and their complexity scales unfavorably with the number of BS antennas, preventing their use in massive MU-MIMO. Linear precoding, on the other hand, is an attractive low-complexity approach to massive MU-MIMO downlink precoding, which offers competitive performance to DPC for large antenna arrays [13], [14].

These results assume that the RF circuitry connected to each antenna port at the BS is ideal. The impact of RF hardware impairments at the transmit side has been investigated in, e.g., [15]–[18]. Some of these results indicate that massive MU-MIMO exhibits a certain degree of resilience against RF impairments. The crude aggregate models used for characterizing such hardware impairments, however, are unable to accurately capture the distortion caused by low-resolution DACs.

### A. What are the Benefits of Quantized Massive MU-MIMO?

One of the dominant sources of power consumption in massive MU-MIMO systems are the data converters at the BS. In the downlink, the transmit baseband signal at each RF chain is generated by a pair of DACs. The power consumption of these DACs increases exponentially with the resolution (in bits) and linearly with the bandwidth [19], [20]. In traditional multi-antenna systems, each RF port is connected to a pair of high-resolution DACs (e.g., 10-bit or more). For massive MU-MIMO systems with hundreds or even thousands of antenna elements, this would lead to prohibitively high power consumption due to the large number of required DACs. Hence, the DAC resolution must be limited to keep the power budget within tolerable levels. Furthermore, an often overlooked issue in massive MU-MIMO is the vast amount of data that must be exchanged between the baseband-processing unit and the radio unit (where the DACs are located). To make matters worse, in many deployment scenarios, these two units are separated by a large distance. Hence, lowering the DAC resolution is a potential solution to mitigate the data-rate bottleneck on the fronthaul.

## B. Relevant Prior Art

1) *Quantized Receivers*: Reducing the fronthaul throughput at the BS can be achieved by using low-resolution DACs in the downlink and low-resolution analog-to-digital converters (ADCs) in the uplink. Several recent contributions have studied the use of low-resolution ADCs in the massive MU-MIMO uplink. In particular, there has been a significant interest in the 1-bit ADC case. For frequency-flat channels, the performance of 1-bit ADCs followed by linear detectors was analyzed in, e.g., [21]–[24], where it was shown that large achievable sum rates are supported. Similar conclusions were made in [25] for the frequency-selective case. Nonlinear detection algorithms for frequency-selective channels were studied in, e.g., [26]. These results suggest that the number of ADC bits can be reduced significantly compared to today’s systems.

2) *Quantized Precoding*: In contrast to the uplink case, there has only been a small number of contributions that consider the massive MU-MIMO downlink with low-resolution DACs at the BS. In [27], the authors design a linear-quantized precoder based on the minimum mean-square error (MMSE) criterion, taking into account the distortion caused by the DACs. For DACs with 4 to 6 bits resolution, the precoder proposed in [27] is shown to outperform conventional linear-quantized precoders for small-to-moderate-sized MIMO systems at high signal-to-noise ratio (SNR). Massive MU-MIMO systems with 1-bit DACs are investigated in [28], where it is shown that maximal ratio transmission (MRT) precoding results in manageable distortion levels. Again for the case of 1-bit DACs, the authors of [29] analyze the performance of zero-forcing (ZF) precoding on a Rayleigh-fading channel. Interestingly, it is shown that the received signal can be made proportional to the transmitted signal when the number of BS antennas tend to infinity. This implies that the severe per-antenna distortion caused by the 1-bit DACs averages out when many transmit antennas are available. A linear precoder where the 1-bit quantized outcomes are rescaled in the analog domain was presented in [30]. There, the authors use the gradient projection algorithm to find a precoder that yields improved performance over the one reported in [27]. In [31], it is shown that, in the presence of transceiver nonlinearities (e.g., finite-resolution DACs), the achievable rate can be improved by minimizing the MSE between the transmitted symbols and the received signal prior to decoding. This result which, as we shall see, is related to the approach taken in this paper, relies on the assumptions of Gaussian inputs and nearest-neighbor decoding.

3) *Low-PAR and Constant-Envelope Precoding*: Other types of hardware-aware precoding have previously been considered for massive MU-MIMO systems, with the goal of reducing the linearity requirements at the BS. In [32], joint MU precoding and peak-to-average power ratio (PAR) reduction was achieved by solving a convex optimization problem. Constant-envelope precoding, which minimizes the PAR by transmitting constant-modulus signals only, was studied in [33], [34]. Note that the 1-bit DAC precoding problem can be seen as a special (or extreme) case of constant-envelope precoding, where the phase of the transmitted signal is limited to only four different values.

## C. Contributions

We consider quantized precoding for the massive MU-MIMO downlink over frequency-flat channels. Similarly to [28]–[30], we consider DACs operating at symbol rate sampling frequency. However, in contrast to [28]–[30], we do not restrict ourselves to 1-bit DACs and linear precoding. Specifically, we consider both *linear-quantized precoders*, where a linear precoder is followed by a finite-resolution DAC, and *nonlinear precoders*, where the data vector together with the CSI is used to directly generate the DAC outputs. Our contributions can be summarized as follows.

- We formulate the MMSE-optimal linear-quantized precoding problem and present low complexity, suboptimal linear-quantized precoders that yield approximate solutions to this problem. We use Busgang’s theorem to develop simple closed-form approximations for the rate achievable with linear-quantized precoding and low-resolution DACs. Through numerical simulations, we validate the accuracy of these approximations, and we show that only a small number of quantization bits are sufficient to close the performance gap to the infinite-resolution case. For the special case of 1-bit DACs, we obtain a firm lower bound on the achievable rate with linear precoding.
- For the 1-bit case, we develop a variety of low-complexity nonlinear precoders that achieve near-optimal performance. We show that the MMSE-optimal downlink precoding problem can be relaxed to a convex problem that can be solved in a computationally-efficient manner. We propose computationally efficient algorithms based on semidefinite relaxation, squared- $\ell_\infty$  norm relaxation, and sphere decoding, and discuss advantages and limitations of each of these methods. Through numerical simulations, we demonstrate the superiority of nonlinear precoding over linear-quantized precoding.
- We investigate the sensitivity of the proposed precoders to channel-estimation errors and demonstrate that the proposed precoders are robust to imperfect CSI at the BS.

Our results reveal that massive MU-MIMO enables the use of low-resolution DACs at the BS without a significant performance loss in terms of error-rate performance and information-theoretic rates.

## D. Notation

Lowercase and uppercase boldface letters designate column vectors and matrices, respectively. For a matrix  $\mathbf{A}$ , we denote its complex conjugate, transpose, and Hermitian transpose by  $\mathbf{A}^*$ ,  $\mathbf{A}^T$ , and  $\mathbf{A}^H$ , respectively. The entry on the  $k$ th row and on the  $\ell$ th column of the matrix  $\mathbf{A}$  is denoted as  $[\mathbf{A}]_{k,\ell}$ . For a vector  $\mathbf{a}$ , the  $k$ th entry is denoted as  $[\mathbf{a}]_k$ . We use  $\mathbf{A} \succeq \mathbf{0}$  to indicate that the matrix  $\mathbf{A}$  is positive semidefinite. The trace and the main diagonal of  $\mathbf{A}$  are  $\text{tr}(\mathbf{A})$  and  $\text{diag}(\mathbf{A})$ , respectively. The  $M \times M$  identity matrix and the all-zeros matrix are denoted by  $\mathbf{I}_M$  and  $\mathbf{0}_{M \times N}$ , respectively. The real and imaginary parts of a complex vector  $\mathbf{a}$  are  $\Re\{\mathbf{a}\}$  and  $\Im\{\mathbf{a}\}$ , respectively. We use  $\|\mathbf{a}\|_2$  and  $\|\mathbf{a}\|_\infty$  to denote the  $\ell_2$ -norm and the  $\ell_\infty$ -norm of  $\mathbf{a}$ , respectively. We use  $\text{sgn}(\cdot)$  to denote the signum function, which is applied entry-wise to vectors and

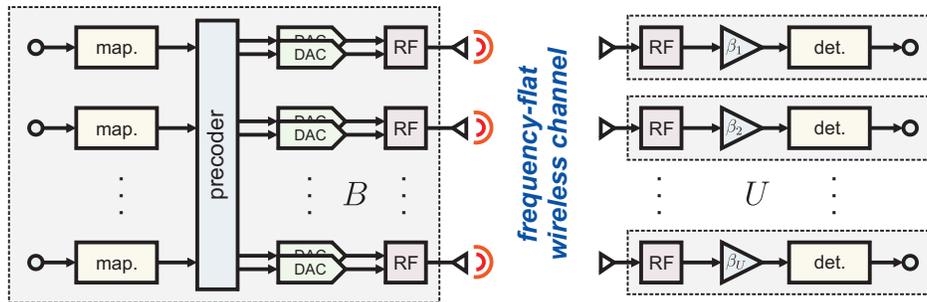


Fig. 1. Overview of the proposed quantized massive MU-MIMO downlink system. Left:  $B$  antenna massive MU-MIMO BS that performs quantized precoding to enable the use of low-resolution DACs; Right:  $U$  single-antenna UEs.

defined as  $\text{sgn}(a) = +1$  if  $a \geq 0$  and  $\text{sgn}(a) = -1$  if  $a < 0$ . We further use  $\mathbb{1}_{\mathcal{A}}(a)$  to denote the indicator function, which is defined as  $\mathbb{1}_{\mathcal{A}}(a) = 1$  for  $a \in \mathcal{A}$  and  $\mathbb{1}_{\mathcal{A}}(a) = 0$  for  $a \notin \mathcal{A}$ . The multivariate complex-valued circularly-symmetric Gaussian probability density function (PDF) with covariance matrix  $\mathbf{K}$  is denoted by  $\mathcal{CN}(\mathbf{0}, \mathbf{K})$ . We use  $f(\cdot)$  to denote PDFs and  $\mathbb{E}_{\mathbf{x}}[\cdot]$  to denote expectation with respect to the random vector  $\mathbf{x}$ . The mutual information between two random vectors  $\mathbf{x}$  and  $\mathbf{y}$  is written as  $\mathcal{I}(\mathbf{x}; \mathbf{y})$ .

### E. Paper Outline

The rest of the paper is organized as follows. In Section II, we introduce the system model and formulate the MMSE-optimal quantized precoding problem. In Section III, we investigate linear-quantized precoders for massive MU-MIMO systems. Section IV deals with nonlinear precoding algorithms for the case of 1-bit DACs. In Section V, we provide numerical simulation results and we analyze the robustness of the developed algorithms to channel-estimation errors. We conclude the paper in Section VI.

## II. SYSTEM MODEL AND QUANTIZED PRECODING

### A. System Model

We consider the downlink of a single-cell massive MU-MIMO system as illustrated in Fig. 1. The system consists of a BS with  $B$  antennas that serves  $U$  single-antenna UEs simultaneously and in the same time-frequency resource. For simplicity, we assume that all RF hardware (e.g., local oscillators, mixers, power amplifiers, etc.) are ideal and that the ADCs at the UEs have infinite resolution. We also assume that the sampling rate of the DACs at the BS is equal to the sampling rate of the ADCs at the UEs and that the system is perfectly synchronized. Finally, we assume that the *reconstruction stage* (see, e.g., [35]) of the DACs consists only of a zero-order hold circuit (no filtering stage).<sup>1</sup> Under these assumptions, the input-output relation of the downlink channel can be modeled as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \quad (1)$$

<sup>1</sup>Symbol-rate sampling combined with low-resolution DACs may yield undesired out-of-band emissions, which may be mitigated by using analog filters. Such filters, however, may in turn cause inter-symbol-interference. In this work, we shall ignore the out-of-band emissions caused by the low-resolution DACs and no filter will be considered.

Here, the vector  $\mathbf{y} = [y_1, \dots, y_U]^T$  contains the received signals at all users, with  $y_u \in \mathbb{C}$  denoting the signal received at the  $u$ th UE. The matrix  $\mathbf{H} \in \mathbb{C}^{U \times B}$  models the downlink channel, and it is assumed to be perfectly known to the BS.<sup>2</sup> We shall also assume that the entries of  $\mathbf{H}$  are independent circularly-symmetric complex Gaussian random variables with unit variance, i.e.,  $h_{u,b} = [\mathbf{H}]_{u,b} \sim \mathcal{CN}(0, 1)$ , for  $u = 1, \dots, U$ , and  $b = 1, \dots, B$ . The vector  $\mathbf{n} \in \mathbb{C}^U$  in (1) models additive noise. We assume the noise to be i.i.d. circularly-symmetric complex Gaussian with variance  $N_0$  per complex entry, i.e.,  $n_u \sim \mathcal{CN}(0, N_0)$ , for  $u = 1, \dots, U$ . We shall also assume that the noise level is known perfectly at the BS.<sup>3</sup>

The precoded vector is denoted by  $\mathbf{x} \in \mathcal{X}^B$ , where the set  $\mathcal{X}$  is the transmit alphabet; this set coincides with the set  $\mathbb{C}$  of complex numbers in the case of infinite-resolution DACs. In real-world BS architectures with finite-resolution DACs, the set  $\mathcal{X}$  is, however, a finite-cardinality alphabet. Specifically, we denote the set of possible real-valued DAC outputs (quantization labels) as  $\mathcal{L} = \{\ell_0, \dots, \ell_{L-1}\}$ . We refer to  $L = |\mathcal{L}|$  and  $Q = \log_2 L$  as the number of quantization levels and the number of quantization bits per real dimension, respectively. For each BS antenna, we assume the same quantization alphabet for the real part and the imaginary part. Hence, the set of complex-valued DAC outputs at each antenna is  $\mathcal{X} = \mathcal{L} \times \mathcal{L}$ . Under these assumptions, the  $b$ th entry of the precoded vector  $\mathbf{x}$  is  $x_b = \ell_R + j\ell_I$  where  $\ell_R, \ell_I \in \mathcal{L}$ .

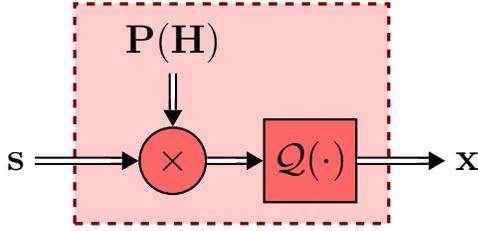
### B. Precoding

Let  $s_u \in \mathcal{O}$  for  $u = 1, \dots, U$  be the constellation point at the BS intended for the UE  $u$ ; here,  $\mathcal{O}$  is the set of constellation points (e.g., QPSK). The BS uses the available CSI, namely the knowledge of the realization of the channel matrix  $\mathbf{H}$ , to precode the symbol vector  $\mathbf{s} = [s_1, \dots, s_U]^T$  into a  $B$ -dimensional precoded vector  $\mathbf{x} = \mathcal{P}(\mathbf{s}, \mathbf{H})$ . Here, the function  $\mathcal{P}(\cdot, \cdot) : \mathcal{O}^U \times \mathbb{C}^{U \times B} \rightarrow \mathcal{X}^B$  represents the precoder. The precoded vector  $\mathbf{x}$  must satisfy the average power constraint

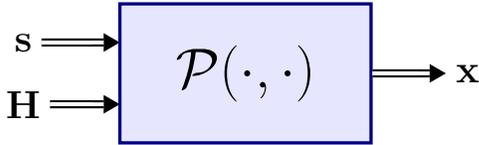
$$\mathbb{E}_{\mathbf{s}} \left[ \|\mathbf{x}\|_2^2 \right] \leq P. \quad (2)$$

<sup>2</sup>In Section V-B, we will relax this assumption by investigating the impact of imperfect CSI to the robustness of the proposed quantized precoding algorithms.

<sup>3</sup>Knowledge of  $N_0$  at the BS can be obtained by explicit feedback from the UEs to the BS.



(a) Linear-quantized precoders: the precoding matrix  $\mathbf{P}$  is designed based on  $\mathbf{H}$ . The transmit vector is the quantized version of  $\mathbf{P}\mathbf{s}$ , i.e.,  $\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s})$ . Here,  $\mathcal{Q}(\cdot)$  denotes the quantizer.



(b) Nonlinear precoders: the quantized transmit vector  $\mathbf{x} \in \mathcal{X}^B$  is a nonlinear function of  $\mathbf{s}$  and  $\mathbf{H}$ , i.e.,  $\mathbf{x} = \mathcal{P}(\mathbf{s}, \mathbf{H})$ .

Fig. 2. Illustration of linear-quantized (a) and nonlinear (b) precoders.

We define  $\rho = P/N_0$  as the SNR.

Coherent transmission of data using multiple BS antennas leads to an *array gain*, which depends on the realization of the fading channel. We shall assume that the  $u$ th UE is able to rescale the received signal  $y_u$  by a factor  $\beta_u \in \mathbb{R}$  to compute an estimate  $\hat{s}_u \in \mathbb{C}$  of the transmitted symbol  $s_u \in \mathcal{O}$  as follows:

$$\hat{s}_u = \beta_u y_u. \quad (3)$$

The problem of downlink precoding has been studied extensively in the literature. Broadly speaking, the goal is to increase the array gain to the intended UE while simultaneously reducing MU interference (MUI) [36]. There exist multiple formulations of this optimization problem based on different performance metrics (e.g., sum-rate throughput, worst-case throughput, error probability, etc.). We refer the interested reader to the tutorial [37] for a comprehensive overview.

Our specific goal is to design a precoder that minimizes the MSE between the received signal and the transmitted symbol vector  $\mathbf{s}$  under the power constraint (2). This problem has been studied extensively for the case of infinite-resolution DACs (see, e.g., [38]–[40]). If the BS is equipped with finite-resolution DACs, then the UEs will experience additional distortion compared to the infinite-resolution case, due to finite cardinality of the set  $\mathcal{X}^B$  of possible precoder outputs.

Finding the MMSE-optimal precoder for BS architectures with finite-resolution DACs is a formidable task due to the finite cardinality of  $\mathcal{X}^B$ . In what follows, we present novel algorithms that efficiently compute approximate solutions to the quantized precoding problem. More specifically, we investigate two approaches: *linear-quantized precoding* in Section III and *nonlinear-quantized precoding* for the special case of 1-bit DACs in Section IV. As illustrated in Fig. 2, linear-quantized precoders perform linear processing (matrix-vector multiplication) followed by quantization; in contrast, nonlinear precoders use the transmit vector  $\mathbf{s}$  together with

the available CSI in order to directly compute the precoded vector  $\mathbf{x}$ . As it will be shown in Section V, nonlinear precoders outperform (often significantly) linear-quantized precoders in terms of error-rate performance at the cost of higher computational complexity.

### III. LINEAR-QUANTIZED PRECODERS

In the infinite-resolution case, linear precoders multiply the  $U$ -dimensional symbol vector  $\mathbf{s}$  with a precoding matrix  $\mathbf{P} \in \mathbb{C}^{B \times U}$  so that  $\mathbf{x} = \mathbf{P}\mathbf{s}$ . This approach is particularly attractive for massive MU-MIMO systems due to (i) the relatively low computational complexity and (ii) the fact that even the simplest linear precoder, namely the MRT precoder, achieves virtually optimal performance in the large-antenna limit (see, e.g., [1]). Linear-quantized precoders inherit the first of these two advantages. Indeed, quantizing the precoded vector implies no additional computational complexity. For linear-quantized precoders, the precoded vector  $\mathbf{x} \in \mathcal{X}^B$  is given by

$$\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s}). \quad (4)$$

Here,  $\mathcal{Q}(\cdot) : \mathbb{C}^B \rightarrow \mathcal{X}^B$  denotes the *quantizer-mapping* function, which is a nonlinear function that describes the joint operation of the  $2B$  DACs at the BS.

The remainder of this section is organized as follows. We start by formulating the MMSE quantized precoding problem for linear-quantized precoders. We then describe the operation of the DACs and define the quantizer-mapping function. We then use Bussgang's theorem [41] to derive a lower bound on the sum-rate capacity for the case of 1-bit DACs at the BS. Finally, we derive a simple closed-form approximation of the rate achievable with Gaussian inputs for the more general case of  $Q$ -bit DACs.

#### A. The Linear-Quantized Precoding Problem

By restricting ourselves to linear-quantized precoding (LQP), we can formulate the quantized precoding problem as follows:

$$(\text{LQP}) \quad \begin{cases} \text{minimize} & \mathbb{E}_{\mathbf{s}} \left[ \|\mathbf{s} - \beta \mathbf{H} \mathcal{Q}(\mathbf{P}\mathbf{s})\|_2^2 \right] + \beta^2 U N_0 \\ \text{subject to} & \mathbb{E}_{\mathbf{s}} \left[ \|\mathbf{x}\|_2^2 \right] \leq P \text{ and } \beta > 0. \end{cases} \quad (5)$$

The resulting precoding matrix  $\mathbf{P}^{\text{LQP}}$  and the associated *precoding factor*  $\beta^{\text{LQP}}$  will be referred to as the optimal solution to the problem (LQP). Here, we have introduced the scalar  $\beta \in \mathbb{R}$  to account for the array gain at the UEs (as commonly done in the MMSE precoding literature; see, e.g., [27], [39]). By solving (5), we find the precoded vector  $\mathbf{x}^{\text{LQP}}$  that minimizes the per-channel MSE between the transmitted symbols  $\mathbf{s}$  and the vector  $\beta \mathbf{y}$ . Indeed, note that

$$\mathbb{E}_{\mathbf{s}} \left[ \|\mathbf{s} - \beta \mathbf{y}\|_2^2 \right] = \mathbb{E}_{\mathbf{s}} \left[ \|\mathbf{s} - \beta \mathbf{H} \mathbf{x}\|_2^2 \right] + \beta^2 U N_0 \quad (6)$$

and recall that  $\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s})$ . Next, we provide more insights on the role of the precoding factor  $\beta$ . We seek a precoded vector  $\mathbf{x}$  that makes the received signal proportional to the transmitted symbol vector  $\mathbf{s}$ , i.e.,  $\mathbf{s} \approx \beta \mathbf{y}$ . To lessen the adverse impact of the noise vector  $\mathbf{n}$  in (1), we look for a design that maximizes the received signal power at the UEs. The cost function in (5)

accomplishes exactly this goal by favoring solutions with a smaller  $\beta$ . Unfortunately, the introduction of the precoding factor  $\beta$  (which is not known to the UEs) may complicate decoding at the UEs.<sup>4</sup>

Solving (5) in closed form is challenging due to the nonlinear operation of the DACs, which is captured by the quantizer-mapping function  $\mathcal{Q}(\cdot)$ . An approximate solution to (5) was given in [27]. This solution is obtained by approximating the statistics of the distortion caused by the DACs. We shall consider here a different approach. Specifically, we design linear precoders that assume infinite-resolution DACs at the BS, and then quantize the resulting precoded vector. Such linear-quantized precoders have the advantage that the precoding matrix  $\mathbf{P}$  does not depend on the resolution of the DACs. Furthermore, as we shall see in Section V-A, the difference in error-rate performance between the precoders found using our approach and the precoder presented in [27] is negligible. We next review a selection of linear precoding algorithms for the case of infinite-resolution DACs.

1) *WF precoding*: For the case when the BS is equipped with infinite-resolution DACs, the solution to (5) is the Wiener filter (WF) precoder [39]:

$$\mathbf{P}^{\text{WF}} = \frac{1}{\beta^{\text{WF}}} \mathbf{H}^H \left( \mathbf{H}\mathbf{H}^H + \frac{UN_0}{P} \mathbf{I}_U \right)^{-1} \quad (7)$$

where

$$\beta^{\text{WF}} = \frac{1}{\sqrt{P}} \text{tr} \left( \left( \mathbf{H}\mathbf{H}^H + \frac{UN_0}{P} \mathbf{I}_U \right)^{-1} \mathbf{H}\mathbf{H}^H \left( \mathbf{H}\mathbf{H}^H + \frac{UN_0}{P} \mathbf{I}_U \right)^{-1} \right)^{-1/2}. \quad (8)$$

We write the resulting precoded vector as  $\mathbf{x}^{\text{WF}} = \mathcal{Q}(\mathbf{P}^{\text{WF}}\mathbf{s})$ .

2) *ZF precoding*: With ZF precoding, the BS nulls the MUI by choosing as precoding matrix the pseudoinverse of the channel matrix. The ZF precoding matrix is obtained from (7) by setting the noise variance  $N_0$  to zero, which yields  $\mathbf{P}^{\text{ZF}} = \frac{1}{\beta^{\text{ZF}}} \mathbf{H}^\dagger$ , where  $\mathbf{H}^\dagger = \mathbf{H}^H (\mathbf{H}\mathbf{H}^H)^{-1}$  is the pseudoinverse of the channel matrix  $\mathbf{H}$ , and  $\beta^{\text{ZF}} = \frac{1}{\sqrt{P}} \sqrt{\text{tr}((\mathbf{H}\mathbf{H}^H)^{-1})}$ . The resulting precoded vector is  $\mathbf{x}^{\text{ZF}} = \mathcal{Q}(\mathbf{P}^{\text{ZF}}\mathbf{s})$ .

3) *MRT precoding*: The MRT precoder maximizes the power directed towards each UE, ignoring MUI. The precoding matrix can be obtained from (7) by letting the noise variance  $N_0$  tend to infinity, which yields  $\mathbf{P}^{\text{MRT}} = \frac{1}{\beta^{\text{MRT}}B} \mathbf{H}^H$  and  $\beta^{\text{MRT}} = \frac{1}{B\sqrt{P}} \sqrt{\text{tr}(\mathbf{H}\mathbf{H}^H)}$ . The resulting precoded vector is  $\mathbf{x}^{\text{MRT}} = \mathcal{Q}(\mathbf{P}^{\text{MRT}}\mathbf{s})$ .

## B. Uniform Quantization of a Complex-Valued Vector

For simplicity, we shall model the DACs as symmetric uniform quantizers with step size  $\Delta$ . When a signal is quantized, the average power in the signal is in general not preserved. Therefore, we further assume that the output of the quantizer is scaled by a constant  $\alpha \in \mathbb{R}$ , to ensure that the transmit

power constraint (2) is satisfied. We start by defining a set of quantization labels  $\mathcal{L} = \{\ell_0, \dots, \ell_{L-1}\}$  with entries

$$\ell_i = \alpha \Delta \left( i - \frac{L-1}{2} \right), \quad i = 0, \dots, L-1. \quad (9)$$

Furthermore, let  $\mathcal{T} = \{\tau_0, \dots, \tau_L\}$ , where  $-\infty = \tau_0 < \tau_1 < \dots < \tau_{L-1} < \tau_L = \infty$  specify the set of  $L+1$  quantization thresholds. For uniform quantizers, the quantization thresholds are given by

$$\tau_i = \Delta \left( i - \frac{L}{2} \right), \quad i = 1, \dots, L-1. \quad (10)$$

The quantizer-mapping function  $\mathcal{Q}(\cdot)$  can be uniquely described by the set of quantization labels  $\mathcal{L}$  and the set of quantization thresholds  $\mathcal{T}$ . The DACs map  $\mathbf{z} \in \mathbb{C}$  with entries  $\{z_b\}$  into the quantized output  $\mathbf{x}$  with entries  $\{x_b\}$  in the following way: if  $\Re\{z_b\} \in [\tau_k, \tau_{k+1})$  and  $\Im\{z_b\} \in [\tau_l, \tau_{l+1})$ , then  $x_b = \ell_k + j\ell_l$ .

The step size  $\Delta$  of the quantizers should be chosen to minimize the distortion between the quantized and nonquantized vector. The optimal step size  $\Delta$  depends on the distribution of the input [42], which in our case depends on both the precoder and the signaling scheme. For simplicity, we set the step size so as to minimize the distortion under the assumption that the per-antenna input to the quantizers is  $\mathcal{CN}(0, P/B)$ -distributed. This step size can be found numerically (see e.g., [43] for details).

In the extreme case of 1-bit DACs, the quantizer-mapping function reduces to

$$\mathcal{Q}(\mathbf{z}) = \sqrt{\frac{P}{2B}} (\text{sgn}(\Re\{\mathbf{z}\}) + j \text{sgn}(\Im\{\mathbf{z}\})). \quad (11)$$

Here, we have chosen the set of possible complex-valued quantization outcomes per antenna to be  $\mathcal{X} = \{\sqrt{P/(2B)} (\pm 1 \pm j)\}$ , which ensures that the power constraint in (2) is satisfied with equality.

## C. Signal Decomposition using Busgang's Theorem

Quantizing the precoded signal causes a distortion  $\mathcal{Q}(\mathbf{P}\mathbf{s}) - \mathbf{P}\mathbf{s}$  that is correlated with the input  $\mathbf{P}\mathbf{s}$  to the DACs. For Gaussian inputs, Busgang's theorem [41] allows us to decompose the quantized signal into a linear function of the input to the quantizers and a distortion term that is *uncorrelated* with the input to the quantizers [31], [44]. This allows us to characterize the rates achievable with Gaussian inputs. We start by stating Busgang's theorem [41], [44].

*Theorem 1*: Consider two zero-mean jointly complex Gaussian random variables  $x$  and  $y$ . Assume that  $x$  is passed through a nonlinear function  $g(\cdot) : \mathbb{C} \rightarrow \mathbb{C}$  that acts independently on the real and the imaginary components of  $x$ . The covariance between  $g(x)$  and  $y$  is given by

$$\mathbb{E}[g(x)y^*] = \frac{\mathbb{E}[g(x)x^*]}{\mathbb{E}[xx^*]} \mathbb{E}[xy^*]. \quad (12)$$

Busgang's theorem has recently been used to analyze the massive MU-MIMO uplink with 1-bit ADCs (see, e.g., [24], [25]). It has also been used in [28] to approximate the distortion levels caused by MRT precoding and 1-bit quantization in the massive MIMO downlink. We shall use Theorem 1 to

<sup>4</sup>We shall elaborate on this point in Section IV-D.

characterize the performance of linear-quantized precoders for the case of  $Q$ -bit uniform DACs. As a first step, we establish the following result, whose proof is given in Appendix A.

*Theorem 2:* Let  $\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s})$  denote the output from a set of uniform quantizers described by the quantizer-mapping function  $\mathcal{Q} : \mathbb{C}^B \rightarrow \mathcal{X}^B$ . Assume that  $\mathbf{P} \in \mathbb{C}^{B \times U}$  and that  $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_U)$ . The quantized vector  $\mathbf{x}$  can be decomposed as

$$\mathbf{x} = \mathbf{G}\mathbf{P}\mathbf{s} + \mathbf{d} \quad (13)$$

where the distortion  $\mathbf{d}$  and the signal  $\mathbf{s}$  are uncorrelated. Furthermore,  $\mathbf{G} \in \mathbb{R}^{B \times B}$  is the following diagonal matrix:

$$\mathbf{G} = \frac{\alpha\Delta}{\sqrt{\pi}} \text{diag}(\mathbf{P}\mathbf{P}^H)^{-1/2} \sum_{i=1}^{L-1} \exp\left(-\Delta^2\left(i - \frac{L}{2}\right)^2 \text{diag}(\mathbf{P}\mathbf{P}^H)^{-1}\right). \quad (14)$$

Here,  $L$  and  $\Delta$  denote the number of levels and the step size of the DACs, respectively.

The following corollary provides a well-known result for the case of 1-bit quantization (see, e.g., [24], [25]). Its proof follows by setting  $L = 2$  and  $\alpha\Delta = \sqrt{2P/B}$  in (14) to satisfy the power constraint (2) with equality.

*Corollary 3:* For the case of 1-bit DACs, the matrix  $\mathbf{G}$  in (14) reduces to

$$\mathbf{G} = \sqrt{\frac{2P}{\pi B}} \text{diag}(\mathbf{P}\mathbf{P}^H)^{-1/2}. \quad (15)$$

Let now  $\mathbf{h}_u^T$  denote the  $u$ th row of the channel matrix  $\mathbf{H}$ , let  $\mathbf{p}_u$  be the  $u$ th column of the precoding matrix  $\mathbf{P}$ , and  $n_u$  be the  $u$ th entry of the noise vector  $\mathbf{n}$ . Using (13), we can express the received signal  $y_u$  at UE  $u$  as follows:

$$y_u = \mathbf{h}_u^T \mathbf{G}\mathbf{P}\mathbf{s} + n_u \quad (16)$$

$$= \mathbf{h}_u^T \mathbf{G}\mathbf{p}_u s_u + \sum_{v \neq u} \mathbf{h}_u^T \mathbf{G}\mathbf{p}_v s_v + \mathbf{h}_u^T \mathbf{d} + n_u \quad (17)$$

$$= \mathbf{h}_u^T \mathbf{G}\mathbf{p}_u s_u + e_u + n_u. \quad (18)$$

Here, the error term  $e_u = \sum_{v \neq u} \mathbf{h}_u^T \mathbf{G}\mathbf{p}_v s_v + \mathbf{h}_u^T \mathbf{d}$  captures both the MUI and the distortion caused by the finite-resolution DACs. Note that  $e_u$  and  $s_u$  are uncorrelated. Indeed,

$$\mathbb{E}_{\mathbf{s}}[e_u s_u^*] = \sum_{v \neq u} \mathbf{h}_u^T \mathbf{G}\mathbf{p}_v \mathbb{E}_{\mathbf{s}}[s_v s_u^*] + \mathbf{h}_u^T \mathbb{E}_{\mathbf{s}}[\mathbf{d} s_u^*] = 0. \quad (19)$$

We shall next use the decomposition in (16) to analyze the performance of linear-quantized precoders.

#### D. Achievable Rate Lower Bound for 1-bit DACs

We assume that each UE scales its received signal  $y_u$  by the scalar  $\beta_u = (\mathbf{h}_u^T \mathbf{G}\mathbf{p}_u)^{-1}$  (which is assumed to be known at the  $u$ th UE) to obtain the following estimate:

$$\hat{s}_u = \beta_u y_u = s_u + \beta_u (e_u + n_u). \quad (20)$$

The nonlinearity introduced by the DACs prevents one to characterize the probability distribution of the error term  $e_u$  in closed form, which makes it difficult to compute the achievable rates. One can, however, lower-bound the achievable rate using the so-called ‘‘auxiliary-channel lower bound’’ [45, p. 3503],

which gives the rates achievable with a mismatched decoder (see [46, ch. 1] for a recent review on the subject). As auxiliary channel, we take the one with output

$$\tilde{s}_u = s_u + \beta_u (\tilde{e}_u + n_u), \quad (21)$$

where  $\tilde{e}_u \sim \mathcal{CN}(0, \mathbb{E}_{\mathbf{s}}[|e_u|^2])$  has the same variance as the actual error term  $e_u$  but is Gaussian distributed. Assuming Gaussian inputs, by standard manipulations of the mutual information, we can bound the achievable rate  $R_u$  for UE  $u = 1, 2, \dots, U$  as follows:

$$R_u = \mathbb{E}_{\mathbf{H}}[\mathcal{I}(s_u; \hat{s}_u | \mathbf{H})] \quad (22)$$

$$= \mathbb{E}_{s_u, \hat{s}_u, \mathbf{H}} \left[ \log_2 \left( \frac{f_{\hat{s}_u | s_u, \mathbf{H}}(\hat{s}_u | s_u, \mathbf{H})}{f_{\hat{s}_u | \mathbf{H}}(\hat{s}_u | \mathbf{H})} \right) \right] \quad (23)$$

$$\geq \mathbb{E}_{s_u, \hat{s}_u, \mathbf{H}} \left[ \log_2 \left( \frac{f_{\hat{s}_u | s_u, \mathbf{H}}(\hat{s}_u | s_u, \mathbf{H})}{f_{\hat{s}_u | \mathbf{H}}(\hat{s}_u | \mathbf{H})} \right) \right] \quad (24)$$

$$= \mathbb{E}_{\mathbf{H}}[\log_2(1 + \gamma_u)] \quad (25)$$

where

$$\gamma_u = \frac{|\mathbf{h}_u^T \mathbf{G}\mathbf{p}_u|^2}{\sum_{v \neq u} |\mathbf{h}_u^T \mathbf{G}\mathbf{p}_v|^2 + \mathbf{h}_u^T \mathbf{C}_{\mathbf{d}\mathbf{d}} \mathbf{h}_u^* + N_0} \quad (26)$$

is the signal-to-interference-noise-and-distortion ratio (SINDR) at the  $u$ th UE.<sup>5</sup> Here,  $\mathbf{C}_{\mathbf{d}\mathbf{d}} = \mathbb{E}_{\mathbf{s}}[\mathbf{d}\mathbf{d}^H]$  denotes the covariance of the distortion  $\mathbf{d}$ . It is worth pointing out that the choice of the auxiliary channel (21) corresponds to the use of mismatched nearest-neighbor decoding at the UEs [48], [49].

Next, we use (13) to write the covariance matrix  $\mathbf{C}_{\mathbf{d}\mathbf{d}}$  in (26) as

$$\mathbf{C}_{\mathbf{d}\mathbf{d}} = \mathbf{C}_{\mathbf{x}\mathbf{x}} - \mathbf{G}\mathbf{P}\mathbf{P}^H \mathbf{G}^H \quad (27)$$

where  $\mathbf{C}_{\mathbf{x}\mathbf{x}} = \mathbb{E}_{\mathbf{s}}[\mathbf{x}\mathbf{x}^H]$  is the covariance matrix of the quantized signal  $\mathbf{x} = \mathcal{Q}(\mathbf{P}\mathbf{s})$ . In the special case of 1-bit DACs,  $\mathbf{C}_{\mathbf{x}\mathbf{x}}$  can be written in closed-form as [50], [51]

$$\mathbf{C}_{\mathbf{x}\mathbf{x}} = \frac{P}{\pi B} \left( \sin^{-1} \left( \text{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}} \Re\{\mathbf{P}\mathbf{P}^H\} \text{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}} \right) + j \sin^{-1} \left( \text{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}} \Im\{\mathbf{P}\mathbf{P}^H\} \text{diag}(\mathbf{P}\mathbf{P}^H)^{-\frac{1}{2}} \right) \right). \quad (28)$$

Thus, using (27) and (28), we can express the SINDR in (26) in closed form for the case of 1-bit DACs. Substituting (26) in (25), one obtains a lower bound on the per-user achievable rate with Gaussian signaling for the 1-bit DAC case. Unfortunately, no closed-form expression for  $\mathbf{C}_{\mathbf{x}\mathbf{x}}$  is available for the multi-bit DAC case. We address this problem in the next section.

#### E. Achievable Rate Approximation for Multi-Bit DACs

In this section, we provide an approximation of (26) for the multi-bit DAC case, which is derived under the assumption that both  $B$  and  $U$  are large and that the error term  $e_u$  in (16)

<sup>5</sup>One can establish (25) also by noting that Gaussian noise is the worst noise for Gaussian inputs [47].

is a Gaussian random variable. The approximation relies on standard random matrix theory arguments. Specifically, let

$$G = \alpha \Delta \sqrt{\frac{B}{\pi P}} \sum_{i=1}^{L-1} \exp\left(-\frac{B\Delta^2}{P} \left(i - \frac{L}{2}\right)^2\right) \quad (29)$$

where the normalization by  $\alpha$ , given by

$$\alpha = \left(2B\Delta^2 \left(\left(\frac{L-1}{2}\right)^2 - 2 \sum_{i=1}^{L-1} \left(i - \frac{L}{2}\right) \Phi\left(\sqrt{2B\Delta^2} \left(i - \frac{L}{2}\right)\right)\right)\right)^{-1/2} \quad (30)$$

ensures that the power constraint (2) is satisfied. In (29), the function  $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$  is the cumulative distribution function of a Gaussian random variable. Let also  $\bar{\rho}$  be defined as follows:

$$\bar{\rho} = \frac{G^2 \rho}{(1 - G^2)\rho + 1}. \quad (31)$$

Following the same approach as in [52]–[54], one can show that, for the three linear-quantized precoders (WF, ZF, and MRT) introduced in Section III-A, the SINDR  $\gamma_u$  in (26) can be approximated for large  $B$  and  $U$  by

$$\begin{aligned} \bar{\gamma}^{\text{WF}} &= \frac{\bar{\rho}}{2} \left(\frac{B}{U} - 1\right) - \frac{1}{2} \\ &\quad + \frac{1}{2} \sqrt{\bar{\rho}^2 \left(\frac{B}{U} - 1\right)^2 + 2\bar{\rho} \left(\frac{B}{U} + 1\right) + 1} \end{aligned} \quad (32)$$

$$\bar{\gamma}^{\text{ZF}} = \bar{\rho} \left(\frac{B}{U} - 1\right) \quad (33)$$

$$\bar{\gamma}^{\text{MRT}} = \frac{\bar{\rho} B}{\bar{\rho}(U-1) + U}. \quad (34)$$

Substituting (32)–(34) into (25), one gets an approximation of the achievable rate with Gaussian signaling and nearest-neighbor decoding that is valid for large  $B$  and  $U$ . In Section V, we verify through numerical simulations that this approximation is accurate already for realistic values of  $B$  and  $U$ .

#### IV. NONLINEAR PRECODERS FOR 1-BIT DACS

We now investigate nonlinear precoders that seek approximate solutions to the MMSE-optimal problem detailed in Section II-B. We shall focus on the extreme case of 1-bit DACs, for which the problem simplifies and efficient numerical algorithms can be developed.

We start by noting that, in the 1-bit case, all DAC outcomes have equal amplitude, and that  $\|\mathbf{x}\|_2^2 = P$  if one sets  $\alpha\Delta = \sqrt{2P/B}$  in (9). This observation allows us to formulate the 1-bit quantized precoding (QP) problem as follows:

$$\text{(QP)} \quad \begin{cases} \text{minimize} & \|\mathbf{s} - \beta \mathbf{H} \mathbf{x}\|_2^2 + \beta^2 U N_0 \\ \text{subject to} & \beta > 0. \end{cases} \quad (35)$$

Here,  $\mathcal{X} = \{\sqrt{P/(2B)}(\pm 1 \pm j)\}$ . The resulting precoded vector  $\mathbf{x}^{\text{QP}}$  and the associated precoding factor  $\beta^{\text{QP}}$  are referred to as the optimal solution to the problem (35).

Compared to the problem (LQP) in (5), where we minimize the MSE averaged over both the symbol vector  $\mathbf{s}$  and the noise vector  $\mathbf{n}$  (for a given  $\mathbf{H}$ ), in (QP) we minimize the MSE averaged over the noise vector  $\mathbf{n}$  (for a given  $\mathbf{s}$  and  $\mathbf{H}$ ). Since the optimization problem is solved for a given  $\mathbf{s}$ , the precoding factor  $\beta$  depends on  $\mathbf{s}$ ; this is in contrast to the linear-quantized case, where  $\beta$  depends only on  $\mathbf{H}$ .<sup>6</sup>

We note that (QP) in (35) resembles an  $\ell_2$ -norm regularized closest-vector problem (CVP), with the unique feature that the discrete set of vectors is parametrized by the continuous precoding factor  $\beta$ . This prevents the straightforward use of conventional algorithms to approximate CVPs [55], [56]. Since the objective function in (35) is a quadratic function in  $\beta$ , we can compute the optimal value of  $\beta$  as

$$\hat{\beta}(\mathbf{x}) = \frac{\Re\{\mathbf{s}^H \mathbf{H} \mathbf{x}\}}{\|\mathbf{H} \mathbf{x}\|_2^2 + U N_0} = \frac{\Re\{\mathbf{s}^H \mathbf{H} \mathbf{x}\}}{\mathbf{x}^H (\mathbf{H}^H \mathbf{H} + \frac{U N_0}{P} \mathbf{I}_B) \mathbf{x}} \quad (36)$$

which depends on  $\mathbf{x}$ . Inserting (36) into the objective function in (35), we obtain the following equivalent formulation of the QP problem:

$$\text{minimize}_{\mathbf{x} \in \mathcal{X}^B} \left\| \mathbf{s} - \hat{\beta}(\mathbf{x}) \mathbf{H} \mathbf{x} \right\|_2^2 + \hat{\beta}(\mathbf{x})^2 U N_0. \quad (37)$$

To obtain  $\beta^{\text{QP}}$ , we can then simply evaluate (36) for the optimal vector  $\mathbf{x}^{\text{QP}}$ . We emphasize that a straightforward exhaustive search to solve (QP) requires the evaluation of  $|\mathcal{X}|^B = 4^B$  candidate vectors, a quantity that grows exponentially with the number of BS antennas  $B$ . For a system with  $B = 128$  antennas at the BS, this approach would require us to evaluate the objective function more than  $10^{77}$  times (more than 10 quattuorvigintillions times). In fact, for a fixed value of  $\beta$ , the problem (QP) is a closest vector problem that is NP-hard [57]. This implies that there are no known algorithms to solve such problems efficiently for large values of  $B$ .<sup>7</sup> Hence, alternative algorithms that solve a lower complexity version of the QP problem are required for massive MU-MIMO systems.

In order to develop such computationally efficient algorithms, we start by defining the auxiliary vector  $\mathbf{b} = \beta \mathbf{x}$  and rewrite (35) in the following equivalent form:

$$\text{minimize}_{\mathbf{b} \in \mathcal{B}^B} \left\| \mathbf{s} - \mathbf{H} \mathbf{b} \right\|_2^2 + \frac{U N_0}{P} \|\mathbf{b}\|_2^2. \quad (38)$$

Here,  $\mathcal{B} = \{\sqrt{P/(2B)}(\pm \beta \pm j\beta)\}$ , for all  $\beta > 0$ . To obtain (38), we have used that  $\beta^2 = \|\mathbf{b}\|_2^2 / P$ . Let  $\mathbf{b}^{\text{QP}}$  be the solution to (38). The resulting precoding vector is obtained by scaling each entry of  $\mathbf{b}^{\text{QP}}$  so that it belongs to the set  $\mathcal{X}$ . Clearly,  $1/\beta^{\text{QP}}$  is the scaling parameter.

It turns out convenient to transform the complex-valued problem (38) into an equivalent real-valued problem using the

<sup>6</sup>We shall discuss how the dependence of  $\beta$  on  $\mathbf{s}$  affects decoding the receiver side in Section IV-D.

<sup>7</sup>As we will show in Section IV-C, we can—in some cases—design branch-and-bound methods (such as sphere-decoding methods) that allow us to solve the quantized precoding problem efficiently for moderately-sized problems. For massive MU-MIMO systems with hundreds of antennas, however, such methods still exhibit prohibitive computational complexity.

following definitions:

$$\mathbf{b}_\mathbb{R} = \begin{bmatrix} \Re\{\mathbf{b}\} \\ \Im\{\mathbf{b}\} \end{bmatrix}, \mathbf{s}_\mathbb{R} = \begin{bmatrix} \Re\{\mathbf{s}\} \\ \Im\{\mathbf{s}\} \end{bmatrix}, \text{ and } \mathbf{H}_\mathbb{R} = \begin{bmatrix} \Re\{\mathbf{H}\} & -\Im\{\mathbf{H}\} \\ \Im\{\mathbf{H}\} & \Re\{\mathbf{H}\} \end{bmatrix}.$$

These definitions enable us to rewrite (38) as

$$\underset{\mathbf{b}_\mathbb{R} \in \mathcal{B}_\mathbb{R}^{2B}}{\text{minimize}} \quad \|\mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2 + \frac{UN_0}{P} \|\mathbf{b}_\mathbb{R}\|_2^2 \quad (39)$$

where  $\mathcal{B}_\mathbb{R} = \{\pm \sqrt{P/(2B)}\beta, \text{ for all } \beta > 0\}$  is the set of scaled antipodal outputs of each 1-bit DAC. We shall next develop a variety of nonlinear precoding methods that find approximate solutions to the problem (39).

### A. Semidefinite Relaxation

Semidefinite relaxation (SDR) is a well-established technique to develop approximate algorithms for a variety of discrete programming problems [58]. For example, SDR is commonly used to find near-ML solutions for the MU-MIMO detection problem (see, e.g., [58], [59]). For the case when the BS is equipped with infinite-resolution DACs, SDR has been used for downlink precoding in [60], [61]. We next show how SDR can be used to find approximate solutions to (35).

In our context, SDR involves relaxing (39) to a semidefinite program (SDP) as follows. We start by writing the real-valued problem (39) in the following equivalent form [58]:

$$\underset{\mathbf{b}_\mathbb{R} \in \mathbb{R}^{2B}, \psi \in \{\pm 1\}}{\text{minimize}} \quad \|\psi \mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2 + \frac{UN_0}{P} \|\mathbf{b}_\mathbb{R}\|_2^2 \quad (40)$$

$$\text{subject to} \quad [\mathbf{b}_\mathbb{R}]_1^2 = [\mathbf{b}_\mathbb{R}]_b^2, b = 2, \dots, 2B.$$

If  $\psi = 1$  then  $\mathbf{b}_\mathbb{R}$  is the solution to (40); if  $\psi = -1$ , the solution is  $-\mathbf{b}_\mathbb{R}$ . Next, let the  $(2B+1) \times (2B+1)$  matrix  $\mathbf{T}_\mathbb{R}$  be defined as follows:

$$\mathbf{T}_\mathbb{R} = \begin{bmatrix} \mathbf{H}_\mathbb{R}^T \mathbf{H}_\mathbb{R} + \frac{UN_0}{P} \mathbf{I}_{2B} & -\mathbf{H}_\mathbb{R}^T \mathbf{s}_\mathbb{R} \\ -\mathbf{s}_\mathbb{R}^T \mathbf{H}_\mathbb{R} & \|\mathbf{s}_\mathbb{R}\|_2^2 \end{bmatrix}. \quad (41)$$

Also, let  $\mathbf{B}_\mathbb{R} = [\mathbf{b}_\mathbb{R}^T \ \psi]^T [\mathbf{b}_\mathbb{R}^T \ \psi]$ . Following steps similar to those in [58], we rewrite the objective function in (40) as

$$\|\psi \mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2 + \frac{UN_0}{P} \|\mathbf{b}_\mathbb{R}\|_2^2 = \text{tr}(\mathbf{T}_\mathbb{R} \mathbf{B}_\mathbb{R}). \quad (42)$$

The problem (40) can now be reformulated as

$$\underset{\mathbf{B}_\mathbb{R} \in \mathbb{S}^{2B+1}}{\text{minimize}} \quad \text{tr}(\mathbf{T}_\mathbb{R} \mathbf{B}_\mathbb{R})$$

$$\text{subject to} \quad \begin{aligned} & [\mathbf{B}_\mathbb{R}]_{1,1} = [\mathbf{B}_\mathbb{R}]_{b,b} \text{ for } b = 2, \dots, 2B, \\ & [\mathbf{B}_\mathbb{R}]_{2B+1, 2B+1} = 1, \mathbf{B}_\mathbb{R} \succeq \mathbf{0}, \\ & \text{and } \text{rank}(\mathbf{B}_\mathbb{R}) = 1. \end{aligned} \quad (43)$$

Here,  $\mathbb{S}^{2B+1}$  denotes the set of real and symmetric  $(2B+1) \times (2B+1)$  matrices. To see why (40) and (43) are equivalent, remember that  $\mathbf{B}_\mathbb{R} = [\mathbf{b}_\mathbb{R}^T \ \psi]^T [\mathbf{b}_\mathbb{R}^T \ \psi]$ , which implies that  $\mathbf{B}_\mathbb{R}$  has rank 1, and that  $[\mathbf{B}_\mathbb{R}]_{b,b} = [\mathbf{b}_\mathbb{R}]_b^2$  for  $b = 1, \dots, 2B$ , and  $[\mathbf{B}_\mathbb{R}]_{2B+1, 2B+1} = \psi^2 = 1$ .

Unfortunately, the rank-1 constraint in (43) is nonconvex, which makes this problem just as hard to solve as the original QP problem in (35). Nevertheless, we can use SDR to relax the problem in (43) by omitting the rank-1 constraint, which

results in the following SDP:

$$\text{(SDR-QP)} \quad \begin{cases} \text{minimize} & \text{tr}(\mathbf{T}_\mathbb{R} \mathbf{B}_\mathbb{R}) \\ & \mathbf{B}_\mathbb{R} \in \mathbb{S}^{2B+1} \\ \text{subject to} & [\mathbf{B}_\mathbb{R}]_{1,1} = [\mathbf{B}_\mathbb{R}]_{b,b}, b = 2, \dots, 2B, \\ & [\mathbf{B}_\mathbb{R}]_{2B+1, 2B+1} = 1, \text{ and } \mathbf{B}_\mathbb{R} \succeq \mathbf{0}. \end{cases} \quad (44)$$

This problem can be solved efficiently using standard methods from convex optimization [62]. If the solution matrix  $\mathbf{B}_\mathbb{R}^{\text{SDR-QP}}$  has rank one, then (SDR-QP) finds the exact solution to the problem (QP) in (39). If, however, the rank exceeds one, we have to extract a precoding vector  $\mathbf{x}^{\text{SDR-QP}}$  that belongs to the discrete set  $\mathcal{X}^B$ . As commonly done, one can obtain such vector by first performing an eigenvalue-decomposition of  $\mathbf{B}_\mathbb{R}^{\text{SDR-QP}}$  and by then quantizing the first  $2B$  entries of the leading eigenvector  $\mathbf{u}_\mathbb{R}$ . To this end, let  $\mathbf{x}_\mathbb{R}^{\text{SDR-QP}}$  denote the real-valued counterpart of  $\mathbf{x}^{\text{SDR-QP}}$ , whose  $b$ th entry ( $b = 1, \dots, 2B$ ) is given by

$$[\mathbf{x}_\mathbb{R}^{\text{SDR-QP}}]_b = \sqrt{\frac{P}{2B}} \text{sgn}([\mathbf{u}_\mathbb{R}]_{2B+1}) \text{sgn}([\mathbf{u}_\mathbb{R}]_b). \quad (45)$$

The multiplication by  $\text{sgn}([\mathbf{u}_\mathbb{R}]_{2B+1})$  takes into account the potential sign change caused by  $\psi$ . The  $b$ th entry of the resulting complex-valued precoded vector  $\mathbf{x}^{\text{SDR-QP}}$  is obtained as follows:

$$[\mathbf{x}^{\text{SDR-QP}}]_b = [\mathbf{x}_\mathbb{R}^{\text{SDR-QP}}]_b + j [\mathbf{x}_\mathbb{R}^{\text{SDR-QP}}]_{B+b} \quad (46)$$

for  $b = 1, \dots, B$ . We refer to this approach as SDR with a rank-one approximation (SDR1). Alternatively, we can obtain a precoding vector in  $\mathcal{X}^B$  using more sophisticated randomized procedures. See the survey article [58] for more details. We refer to this approach as SDR with randomization (SDRr).

SDR enables the computation of approximate solutions to the NP-hard problem (QP) in polynomial time. Specifically, the worst-case complexity scales with  $B^{4.5}$  [58]. However, SDR lifts the problem to a higher dimension: from  $2B$  dimensions to  $(2B+1)^2$  dimensions. Furthermore, implementing the corresponding numerical solvers entails high hardware complexity [63]. Recently, a hardware-friendly *approximate* SDR solver for problems of dimension up to  $B = 16$  was proposed in [63]. However, the complexity of this solver still prevents its use for massive MU-MIMO systems with hundreds of antennas. Hence, we conclude that SDR is a suitable technique only for small to moderately-sized systems (e.g., 16 BS antennas or less). For larger antenna arrays, alternative methods are necessary. One such method is described next.

### B. Squared $\ell_\infty$ -Norm Relaxation

We next present a novel method to approximately solving (35), which avoids lifting the problem to a higher dimension and requires low complexity. We start by rewriting the real-valued optimization problem (39) as

$$\underset{\mathbf{b}_\mathbb{R} \in \mathbb{R}^{2B}}{\text{minimize}} \quad \|\mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2 + \frac{2BUN_0}{P} \|\mathbf{b}_\mathbb{R}\|_\infty^2 \quad (47)$$

$$\text{subject to} \quad [\mathbf{b}_\mathbb{R}]_1^2 = [\mathbf{b}_\mathbb{R}]_b^2, b = 2, \dots, 2B$$

where we used that  $\|\mathbf{b}_\mathbb{R}\|_2^2 = 2B\|\mathbf{b}_\mathbb{R}\|_\infty^2$  under the constraint that  $[\mathbf{b}_\mathbb{R}]_1^2 = [\mathbf{b}_\mathbb{R}]_b^2$  for  $b = 2, \dots, 2B$ . By dropping the

nonconvex constraints  $[\mathbf{b}_\mathbb{R}]_1^2 = [\mathbf{b}_\mathbb{R}]_b^2$  for  $b = 2, \dots, 2B$ , we obtain the following convex relaxation of (47):

$$(\ell_\infty^2\text{-QP}) \quad \underset{\mathbf{b}_\mathbb{R} \in \mathbb{R}^{2B}}{\text{minimize}} \quad \|\mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2 + \frac{2BUN_0}{P} \|\mathbf{b}_\mathbb{R}\|_\infty^2 \quad (48)$$

which, as we shall see, can be solved efficiently. To extract a feasible precoding vector  $\mathbf{x}^{\ell_\infty^2\text{-QP}} \in \mathcal{X}^B$  from the solution  $\mathbf{b}_\mathbb{R}^{\ell_\infty^2\text{-QP}}$  to the problem (48), we quantize the entries of the vector to the quaternary set  $\mathcal{X}$  by computing

$$\left[ \mathbf{x}_\mathbb{R}^{\ell_\infty^2\text{-QP}} \right]_b = \sqrt{\frac{P}{2B}} \operatorname{sgn} \left( \left[ \mathbf{b}_\mathbb{R}^{\ell_\infty^2\text{-QP}} \right]_b \right) \quad (49)$$

for  $b = 1, \dots, 2B$ , where  $\mathbf{x}_\mathbb{R}^{\ell_\infty^2\text{-QP}}$  is the real-valued counterpart of  $\mathbf{x}^{\ell_\infty^2\text{-QP}}$ . As in (46), we then obtain the complex-valued precoded vector as follows:

$$\left[ \mathbf{x}^{\ell_\infty^2\text{-QP}} \right]_b = \left[ \mathbf{x}_\mathbb{R}^{\ell_\infty^2\text{-QP}} \right]_b + j \left[ \mathbf{x}_\mathbb{R}^{\ell_\infty^2\text{-QP}} \right]_{B+b} \quad (50)$$

for  $b = 1, \dots, B$ . There exist several numerical optimization methods that are capable of solving problems of the form of  $(\ell_\infty^2\text{-QP})$  in a computationally efficient manner. The most prominent methods are forward-backward splitting (FBS) [64], [65] and Douglas-Rachford (DR) splitting [66], [67]. In what follows, we develop a DR splitting method, which we refer to as squared-infinity norm Douglas-Rachford splitting (SQUID). We define the two convex functions  $g(\mathbf{b}_\mathbb{R}) = \|\mathbf{s}_\mathbb{R} - \mathbf{H}_\mathbb{R} \mathbf{b}_\mathbb{R}\|_2^2$  and  $f(\mathbf{b}_\mathbb{R}) = \frac{2BUN_0}{P} \|\mathbf{b}_\mathbb{R}\|_\infty^2$ , and solve

$$\underset{\mathbf{b}_\mathbb{R} \in \mathbb{R}^{2B}}{\text{minimize}} \quad g(\mathbf{b}_\mathbb{R}) + f(\mathbf{b}_\mathbb{R}). \quad (51)$$

Let

$$\operatorname{prox}_h(\mathbf{w}) = \arg \min_{\mathbf{b}_\mathbb{R} \in \mathbb{R}^{2B}} h(\mathbf{b}_\mathbb{R}) + \frac{1}{2} \|\mathbf{b}_\mathbb{R} - \mathbf{w}\|_2^2 \quad (52)$$

define the proximal operator for the function  $h(\cdot)$  [64]. By initializing  $\mathbf{b}_\mathbb{R}^{(0)} = \mathbf{0}_{2B \times 1}$  and  $\mathbf{c}_\mathbb{R}^{(0)} = \mathbf{0}_{2B \times 1}$ , SQUID performs the following iterative procedure for  $t = 1, 2, \dots$  until convergence or until a maximum number of iterations has been reached:

$$\mathbf{a}_\mathbb{R}^{(t)} = \operatorname{prox}_g(2\mathbf{b}_\mathbb{R}^{(t-1)} - \mathbf{c}_\mathbb{R}^{(t-1)}) \quad (53)$$

$$\mathbf{b}_\mathbb{R}^{(t)} = \operatorname{prox}_f(\mathbf{c}_\mathbb{R}^{(t-1)} - \mathbf{a}_\mathbb{R}^{(t)} - \mathbf{b}_\mathbb{R}^{(t-1)}) \quad (54)$$

$$\mathbf{c}_\mathbb{R}^{(t)} = \mathbf{c}_\mathbb{R}^{(t-1)} + \mathbf{a}_\mathbb{R}^{(t)} - \mathbf{b}_\mathbb{R}^{(t-1)}. \quad (55)$$

The proximal operator  $\operatorname{prox}_g(\cdot)$  in (53) has the following simple<sup>8</sup> expression:

$$\operatorname{prox}_g(\mathbf{w}) = (\mathbf{H}_\mathbb{R}^T \mathbf{H}_\mathbb{R} + \frac{1}{2} \mathbf{I}_{2B \times 2B})^{-1} (\mathbf{H}_\mathbb{R}^T \mathbf{s}_\mathbb{R} + \frac{1}{2} \mathbf{w}). \quad (56)$$

While the proximal operator for the  $\ell_\infty$ -norm is well known in the literature [64], the proximal operator  $\operatorname{prox}_f(\cdot)$  for the *squared*  $\ell_\infty$ -norm, needed in (54), appears to be novel. The following theorem details an efficient procedure for computing this proximal operator. The proof is given in Appendix B.

*Theorem 4:* Let  $\lambda > 0$ . Then, the squared  $\ell_\infty$ -norm

<sup>8</sup>One can further accelerate the evaluation of this proximal operator by using the Woodbury matrix identity (which reduces the dimension of the matrix inverse), and by precomputing certain constant quantities, such as  $\mathbf{H}_\mathbb{R}^T \mathbf{s}_\mathbb{R}$ .

---

### Algorithm 1 Proximal operator for the $\ell_\infty^2$ -norm

---

```

1: inputs:  $\mathbf{z} \in \mathbb{R}^N$ ,  $\lambda \in (0, \infty)$ 
2:  $\mathbf{a} \leftarrow \operatorname{abs}(\mathbf{z})$ 
3:  $\mathbf{s} \leftarrow \operatorname{sort}(\mathbf{a}, \text{'descending'})$ 
4: for  $k = 1, \dots, N$  do
5:    $c_k \leftarrow \frac{1}{2\lambda+k} \sum_{i=1}^k s_i$ 
6: end for
7:  $\alpha \leftarrow \max\{0, \max_k \{c_k\}\}$ 
8: for  $k = 1, \dots, N$  do
9:    $u_k \leftarrow \min\{a_k, \alpha\} \operatorname{sgn}(z_k)$ 
10: end for
11: return  $\mathbf{u}$ 

```

---

proximal operator

$$\mathbf{u} = \operatorname{prox}_{\lambda \ell_\infty^2}(\mathbf{z}) = \arg \min_{\mathbf{u} \in \mathbb{R}^N} \lambda \|\mathbf{u}\|_\infty^2 + \frac{1}{2} \|\mathbf{z} - \mathbf{u}\|_2^2 \quad (57)$$

can be computed using the procedure summarized in Algorithm 1.

In summary, SQUID enables us to solve the relaxed problem in (48) in a computationally efficient manner. Indeed, each iteration requires only simple matrix and vector operations, and the evaluation of the proximal operator in Algorithm 1. The performance of SQUID is investigated in Section V where we demonstrate that this low-complexity algorithm achieves performance comparable to SDR, which is a far more demanding algorithm in terms of computational complexity.

### C. Sphere Precoding

Sphere decoding (SD) is a common method to solve CVPs exactly but at lower average computational complexity than a naïve exhaustive search [55], [56], [68]. The idea of SD is to constrain the search for possible optimal solutions to a hypersphere of radius  $r$ . By transforming the optimal CVP into a tree-search problem, one can then perform a depth-first branch-and-bound procedure and prune branches that exceed the radius constraint to reduce the number of candidate vectors. While SD reduces (often significantly) the average complexity compared to an exhaustive search, it was shown to exhibit exponential complexity in the number of variables for data detection in multi-antenna wireless systems [69], [70].

To adapt SD to 1-bit quantized precoding (we call this adaptation *sphere precoding* (SP)), we proceed as follows. Assume that the optimal precoding factor  $\beta$  is known. Then, we can rewrite the objective function in (37) as follows:

$$\|\mathbf{s} - \beta \mathbf{H} \mathbf{x}\|_2^2 + \beta^2 UN_0 = \|\mathbf{s} - \beta \mathbf{H} \mathbf{x}\|_2^2 + \beta^2 \frac{UN_0}{P} \|\mathbf{x}\|_2^2 \quad (58)$$

$$= \|\bar{\mathbf{s}} - \beta \bar{\mathbf{H}} \mathbf{x}\|_2^2. \quad (59)$$

In (58), we used that  $\|\mathbf{x}\|_2^2 = P$  in the 1-bit case; in (59) we set  $\bar{\mathbf{s}} = [\mathbf{s}^T \mathbf{0}_B^T]^T$  and  $\bar{\mathbf{H}} = [\mathbf{H}^T \sqrt{UN_0/P} \mathbf{I}_B]^T$ . Hence, we can write the precoding problem as

$$\underset{\mathbf{x} \in \mathcal{X}^B}{\text{minimize}} \quad \|\bar{\mathbf{s}} - \beta \bar{\mathbf{H}} \mathbf{x}\|_2 \quad (60)$$

which can be solved using SD. More specifically, by computing the QR factorization  $\bar{\mathbf{H}} = \mathbf{Q} \mathbf{R}$ , where  $\mathbf{Q} \in \mathbb{C}^{(U+B) \times B}$

with  $\mathbf{Q}^H \mathbf{Q} = \mathbf{I}_B$  and  $\mathbf{R} \in \mathbb{C}^{B \times B}$  is upper triangular with non-negative diagonal entries, we obtain the equivalent problem

$$(\text{SP}) \quad \underset{\mathbf{x} \in \mathcal{X}^B}{\text{minimize}} \quad \|\mathbf{Q}^H \hat{\mathbf{s}} - \beta \mathbf{R} \mathbf{x}\|_2. \quad (61)$$

The triangular structure of this problem allows us to deploy standard SD methods, as the one in [55].

In practice, the optimal precoding factor  $\beta$  is unknown. We therefore propose the following alternating optimization approach. At iteration  $t = 1$ , we initialize the algorithm with the precoding factor obtained from WF precoding. Specifically, we use (36) and set  $\beta_1 = \hat{\beta}(\mathbf{x}^{\text{WF}})$ . We then solve (SP) to obtain  $\mathbf{x}_t^{\text{SP}}$  and compute an improved precoding factor  $\beta_{t+1} = \hat{\beta}(\mathbf{x}_t^{\text{SP}})$  using (36). We repeat this procedure for  $t = 2, 3, \dots$  until convergence or until a maximum number of iterations is reached. Our simulations have shown that this procedure usually converges in only 1 to 3 iterations and achieves near-optimal performance for small to moderately-sized MIMO systems (in Section V-A, we present numerical results for the case of  $B = 8$  antennas). We note that a plethora of SD-related methods can be used to solve SP. However, the exponential complexity of SD prevents its use for massive MIMO systems with hundreds of antennas.

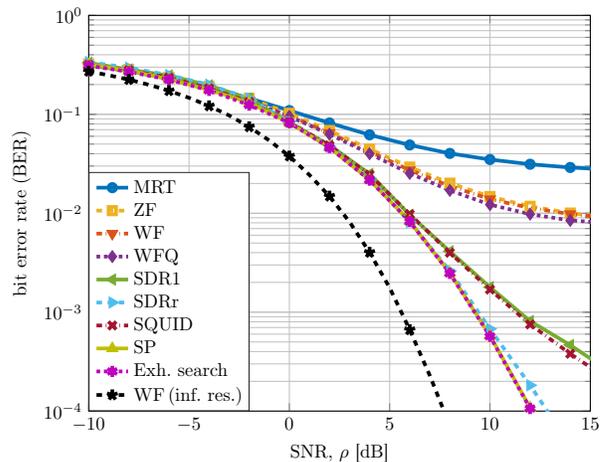
#### D. Decoding at the UEs

As for the case of linear-quantized precoders, we assume that the  $u$ th UE is able to scale the received signal by some scaling factor  $\beta_u$ . Note that the scaling factor  $\beta_u$  can not directly be chosen to be equal to the precoding factor  $\beta$ , since  $\beta$  depends in the nonlinear case on the instantaneous transmit vector  $\mathbf{s}$  and cannot be estimated at the UEs. It is worth noting that for the special case in which the entries of  $\mathbf{s}$  are taken from of a constant-modulus constellation (e.g.,  $M$ -PSK) and the receiver employs symbol-wise nearest-neighbor decoding (i.e., each UE maps its estimate  $\hat{s}_u$  in (3) to the nearest constellation point, which implies that both the residual MUI and the quantization error are treated as Gaussian noise, although they are not Gaussian), the scaling factor  $\beta_u$  chosen by the receiver does not affect performance because the decision regions are circular sectors in the complex plane. In the simulation results in Section V, we shall focus on QPSK modulation for which no scaling is needed. In a follow-up work [71], we presented simulation results for the case of higher-order constellations that do not satisfy the constant-modulus assumption (e.g., 16-QAM). In this case, it is sufficient to modify the precoding problem (35) so that a single value of  $\beta$  is chosen for a block of transmit symbols whose length does not exceed the channel coherence time. This allows the UEs to estimate  $\beta$  through pilot transmissions or blind estimation techniques.

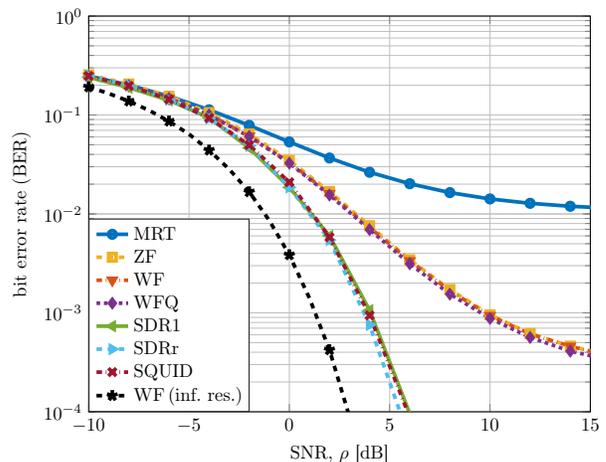
### V. NUMERICAL RESULTS

We now present numerical simulations for the quantized precoders introduced in Section III and Section IV. Due to space constraints, we shall focus on a limited set of system parameters.<sup>9</sup>

<sup>9</sup>Our simulation framework is available for download from GitHub ([https://github.com/quantizedmassivemimo/1bit\\_precoding](https://github.com/quantizedmassivemimo/1bit_precoding)). The purpose is to enable interested readers to perform their own simulations with different system parameters and also to test alternative algorithms.



(a)  $B = 8$  and  $U = 2$ .



(b)  $B = 128$  and  $U = 16$ .

Fig. 3. Uncoded BER with QPSK signaling for 1-bit DACs as a function of the SNR,  $\rho$ , for the precoders introduced in Section III and in Section IV. The performance of the WFQ precoder proposed in [27] is also illustrated for comparison.

#### A. Error-Rate Performance

We start by comparing the performance of the developed precoders in terms of uncoded bit error rate (BER). In what follows, we assume that the UEs perform symbol-wise nearest-neighbor decoding.

In Fig. 3, we compare the BER with QPSK signaling and 1-bit DACs for the linear precoders presented in Section III (namely, WF, ZF and, MRT) and the nonlinear precoding algorithms presented in Section IV (namely, SDR1, SDRr, SQUID and SP). For comparison, we also report the performance of the WF-quantized (WFQ) precoder proposed in [27], and the performance of the WF precoder for the infinite-resolution case.

In Fig. 3a, we consider the case  $B = 8$  BS antennas and  $U = 2$  UEs (moderately-sized MIMO system). For this case, one can find the optimal solution to (QP) in (35) by exhaustive search. We find that the gap between the performance of the optimal nonlinear precoder and the performance of the infinite-resolution WF precoder is remarkably small: about 4 dB for a target BER of  $10^{-3}$ . Furthermore, the SP algorithm achieves near-optimal performance, as does SDRr. SQUID

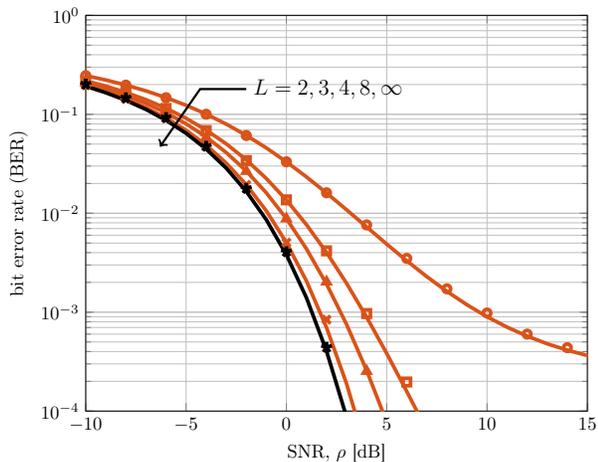


Fig. 4. Uncoded BER with QPSK signaling and WF precoding with multi-bit DACs;  $B = 128$  and  $U = 16$ . Here,  $L$  denotes the number of quantization levels. The markers correspond to simulated values and the solid lines correspond to closed-form approximations.

and SDR1 follow closely the optimal curve up to a BER of  $10^{-2}$  but then their performance degrades. The linear-quantized precoders, on the other hand, are adversely impacted by the coarse 1-bit quantization. Indeed, the BER for linear-quantized precoding saturates at  $10^{-2}$  or above. Hence, in contrast to recently reported findings [29], our results suggest that nonlinear precoding offers significant advantages in terms of BER compared to linear-quantized precoding.

In Fig. 3b, we consider a massive MIMO system with  $B = 128$  BS antennas and  $U = 16$  UEs. Exhaustive search and SP are not viable in this setup due to the exponential complexity in  $B$  that these methods entail. We note that the increased number of antennas yields a performance improvement for the linear-quantized precoders. Indeed, with ZF, WF, or WFQ one can now support error probabilities below  $10^{-3}$ . However, the nonlinear precoders still significantly outperform the linear-quantized precoders. The gap to the infinite-resolution BER with SQUID is about 3 dB for a target BER of  $10^{-3}$ . With WFQ precoding, the gap is about 8 dB for the same BER target.

It is worth pointing out that at low SNR, the error-rate performance of all precoders is comparable. In this regime, linear-quantized precoders may offer a better performance-complexity trade-off. Furthermore, linear-quantized precoders may yield satisfactory BER performance for a larger range of SNR values if the number of BS antennas is increased further.

In Fig. 4, we show the uncoded BER for WF precoding as a function of the SNR and the number of DAC levels  $L$  for a system with  $U = 16$  UEs and  $B = 128$  BS antennas. The simulated BER values in Fig. 4 are compared with closed-form approximations obtained by approximating the uncoded BER by  $1 - \Phi\left(\sqrt{\bar{\gamma}^{\text{WF}}}\right)$  where  $\bar{\gamma}^{\text{WF}}$  is given in (32). We observe that this approximation is accurate for the entire range of SNR values. We further observe that low BER probabilities can be attained with very coarse DACs. Interestingly, by only adding a zero-level in the DACs (so that  $L = 3$ ), the performance is drastically improved compared to the 1-bit case ( $L = 2$ ). Furthermore, with only 3-bit DACs ( $L = 8$ ) the performance

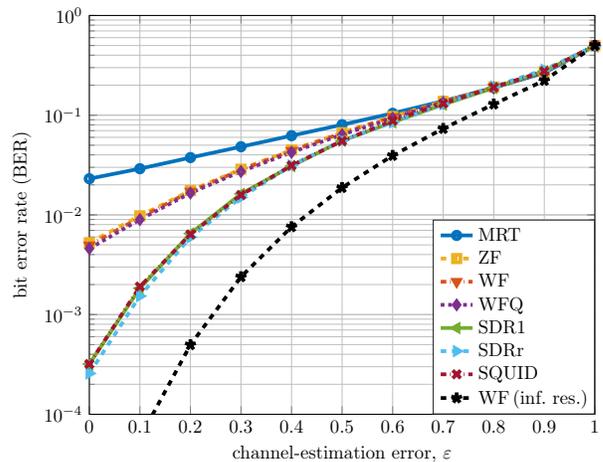


Fig. 5. Uncoded BER with QPSK signaling for 1-bit DACs as a function of the channel-estimation error,  $\varepsilon$ .

gap to the infinite-resolution case is negligible. This suggests that it is possible to significantly reduce the number of bits in the high-resolution DACs used in today's systems.

### B. Robustness to Channel-Estimation Errors

So far, we have assumed that the BS has access to perfect CSI. In this section we shall relax this assumption to investigate the robustness of the developed algorithms to channel estimation errors. More specifically, we shall assume that the BS has access to a noisy version of  $\mathbf{H}$  modelled as

$$\hat{\mathbf{H}} = \sqrt{1 - \varepsilon}\mathbf{H} + \sqrt{\varepsilon}\mathbf{Z}. \quad (62)$$

Here,  $\varepsilon \in [0, 1]$  and  $\mathbf{Z}$  has  $\mathcal{CN}(0, 1)$  entries. We refer to  $\varepsilon$  as the channel-estimation error:  $\varepsilon = 0$  corresponds to perfect CSI and  $\varepsilon = 1$  corresponds to no CSI; intermediate values corresponds to partial CSI.

In Fig. 5, we show, for the 1-bit case, the uncoded BER with QPSK signaling as a function of the channel-estimation error  $\varepsilon$  for a system with  $B = 128$  BS antennas and  $U = 16$  UEs. Interestingly, the nonlinear precoders outperform the linear-quantized precoders whenever  $\varepsilon \leq 0.5$ . This implies that nonlinear precoders can be used also when only imperfect CSI is available to the BS.

### C. Achievable rate

Next, we validate the analytic results on the achievable rate with linear-quantized precoders reported in Section III by numerical simulations.

In Fig. 6, we show the achievable sum-rate with Gaussian signaling and WF precoding as a function of the SNR and the number of DAC levels. The rate approximation computed using (32) is illustrated together with the rate lower bound (25) for the 1-bit case. We also show the achievable rate computed numerically using (23) by simulating many noise and interference realizations for each channel realization and by mapping the resulting  $\hat{s}_u$  to a rectangular grid in the complex plane to estimate the probability density functions required to compute (23) (see, e.g., [23] for details). We note that the

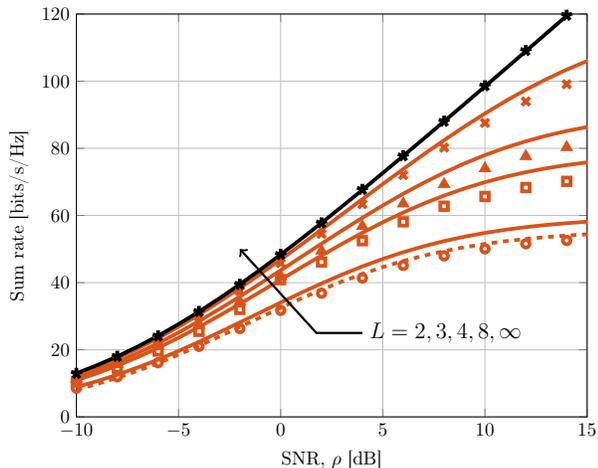


Fig. 6. Achievable sum-rate with Gaussian signaling and WF precoding with multi-bit DACs;  $B = 128$  and  $U = 16$ . Here,  $L$  denotes the number of quantization levels. The markers correspond to simulated values and the solid lines correspond to closed-form approximations. The dashed line corresponds to the lower bound (25) for 1-bit DACs.

asymptotic approximation matches well the numerical results, confirming its accuracy. We further note that, analogously to the uplink scenario [23], [24], high sum-rate throughputs can be achieved despite having low-resolution DACs at the BS.

## VI. CONCLUSIONS

We have presented novel algorithms for the problem of downlink precoding in massive MIMO systems equipped with low-resolution DACs at the BS. To handle the challenges imposed by the finite-cardinality outputs of the DACs, we have considered two distinct approaches, namely linear-quantized precoding and nonlinear precoding. We have shown that, with linear-quantized precoding, the use of DACs with 3 to 4 bits of resolution is sufficient to close the performance gap (measured in terms of both BER and achievable rate) to the infinite-resolution case. Furthermore, we have developed an asymptotic approximation of the effective SINDR, which can be used to predict the system performance accurately using simple closed-form expressions.

Linear-quantized precoders are, however, far from optimal. For the case of 1-bit DACs, we have shown that the error-rate performance can be significantly improved by allowing for nonlinear precoding. For example, we showed that for a BS with 128 BS antennas serving 16 UEs, the gap to infinite-resolution performance is about 8 dB for a target BER of  $10^{-3}$  with linear-quantized precoders, but only 3 dB with nonlinear precoders. Nonlinear precoding, however, entails increased signal-processing complexity. For small-to-moderate sized systems (e.g., 16 BS antennas or less), SDR- and SP-based precoders offer near-optimal BER-performance at tolerable complexity. For massive MIMO systems, SQUID is an efficient and hardware-friendly algorithm to find a near-optimal solution to the 1-bit quantized precoding problem. In a follow-up study [72], we recently proposed two additional precoding algorithms for massive MIMO, and provide both very-large-scale integration (VLSI) designs and field-programmable gate

array (FPGA) implementations. These designs demonstrate that nonlinear precoding algorithms can be realized in practice, with a manageable implementation complexity.

There are many avenues for future work. Extending our analysis and our algorithms to the frequency-selective case, where the use of orthogonal-frequency division multiplexing (OFDM) is assumed, is part of ongoing work. Early results for the case of OFDM and linear-quantized precoding are reported in [73]. As mentioned in Section I, the use of low-resolution DACs operating at symbol rate may result in significant out-of-band emissions and intersymbol interference. A characterization of both effects is critically required to assess the full potential of low-resolution DAC architectures in real-world MU-MIMO systems. Also, a generalization of our analysis to the case of oversampled DACs, which operate at a sampling frequency larger than the symbol rate, is of practical interest.

## APPENDIX A PROOF OF THEOREM 2

Let  $\mathbf{z} = \mathbf{P}\mathbf{s} \in \mathbb{C}^B$  and  $\mathbf{x} = \mathcal{Q}(\mathbf{z}) \in \mathcal{X}^B$ . It follows from Theorem 1 that the covariance matrices  $\mathbf{C}_{\mathbf{xz}} = \mathbb{E}_{\mathbf{s}}[\mathbf{xz}^H]$  and  $\mathbf{C}_{\mathbf{zz}} = \mathbb{E}_{\mathbf{s}}[\mathbf{zz}^H]$  are related as follows:

$$\mathbf{C}_{\mathbf{xz}} = \mathbf{G}\mathbf{C}_{\mathbf{zz}} \quad (63)$$

where  $\mathbf{G}$  is a  $B \times B$  diagonal matrix with

$$[\mathbf{G}]_{b,b} = \frac{1}{\sigma_b^2} \mathbb{E}[\mathcal{Q}(z_b)z_b^*] \quad (64)$$

where  $z_b = [\mathbf{z}]_b$  and  $\sigma_b^2 = \mathbb{E}[|z_b|^2]$  for  $b = 1, \dots, B$ . Note now that

$$\mathbf{C}_{\mathbf{zz}} = \mathbb{E}_{\mathbf{s}}[\mathbf{zz}^H] = \mathbf{P} \mathbb{E}_{\mathbf{s}}[\mathbf{ss}^H] \mathbf{P}^H = \mathbf{P}\mathbf{P}^H. \quad (65)$$

It follows from (63) that we can write the quantized signal as  $\mathbf{x} = \mathbf{G}\mathbf{z} + \mathbf{d}$ , where the distortion  $\mathbf{d}$  is uncorrelated with  $\mathbf{z}$ . Indeed, note that

$$\mathbb{E}_{\mathbf{s}}[\mathbf{d}\mathbf{z}^H] = \mathbb{E}_{\mathbf{s}}[(\mathbf{x} - \mathbf{G}\mathbf{z})\mathbf{z}^H] = \mathbf{C}_{\mathbf{xz}} - \mathbf{G}\mathbf{C}_{\mathbf{zz}} = \mathbf{0}_{B \times B} \quad (66)$$

where the last equality follows from (63). We next evaluate (64). Note that, since the real and imaginary components of the symbol vector  $\mathbf{s}$  are independent and identically distributed, so are the real and imaginary components of the precoded vector  $\mathbf{z}$ . Therefore, it holds that

$$\mathbb{E}[\mathcal{Q}(z_b)z_b^*] = 2 \mathbb{E}[\mathcal{Q}(z)z] \quad (67)$$

where we have introduced the random variable  $z \sim \mathcal{N}(0, \sigma_b^2/2)$ . For a uniform DAC, the quantizer-mapping function can be expressed as

$$\mathcal{Q}(z) = \frac{\alpha\Delta}{2}(1-L) + \alpha\Delta \sum_{i=1}^{L-1} \mathbb{1}_{[\Delta(i-\frac{1}{2}), \infty)}(z). \quad (68)$$

Inserting (67) and (68) into (64), we get that

$$\begin{aligned} [\mathbf{G}]_{b,b} &= \frac{2}{\sigma_b^2} \mathbb{E}[\mathcal{Q}(z)z] \\ &= \frac{\alpha\Delta}{\sigma_b^2}(1-L) \mathbb{E}[z] \end{aligned} \quad (69)$$

$$+ \frac{2\alpha\Delta}{\sigma_b^2} \sum_{i=1}^{L-1} \mathbb{E} \left[ \mathbb{1}_{[\Delta(i-\frac{L}{2}), \infty)}(z) z \right] \quad (70)$$

$$= \frac{2\alpha\Delta}{\sigma_b^2} \sum_{i=1}^{L-1} \int_{\Delta(i-\frac{L}{2})}^{\infty} \frac{z}{\sqrt{\pi\sigma_b^2}} \exp\left(-\frac{z^2}{\sigma_b^2}\right) dz \quad (71)$$

$$= \frac{\alpha\Delta}{\sqrt{\pi\sigma_b^2}} \sum_{i=1}^{L-1} \exp\left(-\frac{\Delta^2}{\sigma_b^2} \left(i - \frac{L}{2}\right)^2\right). \quad (72)$$

Finally, the desired result (14) follows from (72) by using that  $\sigma_b^2 = [\mathbf{P}\mathbf{P}^H]_{b,b}$ .

## APPENDIX B PROOF OF THEOREM 4

We start by rewriting the proximal operator in (57) as

$$\begin{aligned} \mathbf{u} = \arg \min_{\mathbf{x} \in \mathbb{R}^N, \alpha \in \mathbb{R}} \lambda \alpha^2 + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \\ \text{subject to } x_k^2 \leq \alpha^2, \quad k = 1, \dots, N \end{aligned} \quad (73)$$

and use the Karush-Kuhn-Tucker (KKT) conditions [62] to compute its solution. The Lagrangian of the optimization problem in (73) is given by

$$\mathcal{L}(\mathbf{x}, \alpha, \mathbf{u}) = \lambda \alpha^2 + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \sum_{k=1}^N u_k (x_k^2 - \alpha^2) \quad (74)$$

which yields the following two stationarity conditions:

$$\lambda - \sum_{k=1}^N u_k = 0 \quad (75)$$

$$x_k - y_k + 2x_k u_k = 0, \quad k = 1, \dots, N. \quad (76)$$

The stationarity condition (76) reveals that  $x_k = y_k / (1 + 2u_k)$ , which implies that if  $u_k = 0$ , then  $x_k = y_k$ . Complementary slackness yields  $u_k (x_k^2 - \alpha^2) = 0$ , which implies that if  $u_k \neq 0$ , then  $x_k^2 = \alpha^2$  for a given  $k$ . Hence, the values of  $x_k$  must either be  $|x_k| = \alpha$  or  $x_k = y_k$  so that  $|x_k| < \alpha$ . In words, the proximal operator in (73) clips to  $x_k = \text{sgn}(y_k)\alpha$  the values  $y_k$  whose magnitude exceeds  $\alpha$  and leaves the remaining values unaffected. Hence, we only need to determine the optimal clipping threshold  $\alpha^* > 0$ .

Assume  $x_k \neq 0$  without loss of generality (in the case  $y_k = 0$ , we have  $x_k = 0$ ). Then, the stationarity condition in (76) reveals that  $u_k = \frac{1}{2} \left( \frac{y_k}{x_k} - 1 \right)$ . Together with the stationarity condition (75), we have

$$\sum_{k=1}^N u_k = \frac{1}{2} \sum_{k=1}^N \left( \frac{y_k}{x_k} - 1 \right) = \lambda \quad (77)$$

which implies that

$$\sum_{k=1}^N \frac{y_k}{x_k} = 2\lambda + N. \quad (78)$$

We now partition the indices  $k = 1, \dots, N$  into two disjoint sets  $\Omega$  and  $\Omega^c$ . The set  $\Omega$  contains the indices of the entries  $y_k$  for which  $|y_k| \geq \alpha$ ; the set  $\Omega^c$  contains the indices of

the entries  $u_k$  for which  $|y_k| < \alpha$ . Since  $x_k = \text{sgn}(y_k)\alpha$  for  $k \in \Omega$  and  $x_k = y_k$  for  $k \in \Omega^c$ , it follows from (78) that

$$\sum_{k \in \Omega} \frac{|y_k|}{\alpha} + \sum_{k \in \Omega^c} 1 = 2\lambda + N. \quad (79)$$

Hence,

$$\sum_{k \in \Omega} \frac{|y_k|}{\alpha} = 2\lambda + N - |\Omega^c| = 2\lambda + |\Omega|. \quad (80)$$

We see from (80) that the clipping threshold  $\alpha$  must satisfy

$$\alpha = \frac{\sum_{k \in \Omega} |y_k|}{2\lambda + |\Omega|}. \quad (81)$$

To solve (73), it is convenient to sort the values  $|y_k|$  in descending order. Specifically, let us denote these values by  $r_1 \geq r_2 \geq \dots \geq r_N$ . Then one computes  $\alpha_\ell = \sum_{k=1}^{\ell} r_k / (2\lambda + \ell)$  for  $\ell = 1, 2, \dots, N$  and chooses  $\alpha^*$  as the only  $\alpha_\ell$  that satisfies  $r_{\ell+1} < \alpha_\ell \leq r_\ell$ . Simple algebraic manipulations reveal that this is equivalent to setting  $\alpha^* = \max_{\ell} \alpha_\ell$ . We then use  $\alpha^*$  to perform element-wise clipping. Algorithm 1 implements exactly this procedure in a computationally efficient manner.

## REFERENCES

- [1] F. Rusek, D. Persson, B. Kiong, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [2] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [3] L. Lu, G. Ye Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [4] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.
- [5] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.
- [6] S. Viswanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.
- [7] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.
- [8] M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
- [9] R. D. Wesel and J. M. Cioffi, "Achievable rates for Tomlinson-Harashima precoding," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 824–831, Mar. 1998.
- [10] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J. B. Huber, "Precoding in multiantenna and multiuser communications," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1305–1316, Jul. 2004.
- [11] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2057–2060, Dec. 2004.
- [12] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, Mar. 2005.
- [13] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172–179, Feb. 2013.
- [14] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Massive MU-MIMO downlink TDD systems with linear precoding and downlink pilots," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, Oct. 2013.
- [15] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Int. ITG Workshop on Smart Antennas (WSA)*, Bremen, Germany, Feb. 2010, pp. 189–196.

- [16] U. Gustavsson, C. Sanchéz-Perez, T. Eriksson, F. Athley, G. Durisi, P. Landin, K. Hausmair, C. Fager, and L. Svensson, "On the impact of hardware impairments on massive MIMO," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 294–300.
- [17] X. Zhang, M. Matthaiou, M. Coldrey, and E. Björnson, "Impact of residual transmit RF impairments on training-based MIMO systems," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2899–2911, Aug. 2015.
- [18] F. Athley, G. Durisi, and U. Gustavsson, "Analysis of massive MIMO with hardware impairments and different channel models," in *European Conf. Ant. Prop. (EUCAP)*, Lisbon, Portugal, Apr. 2015.
- [19] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.
- [20] B. Murmann, "ADC performance survey 1997-2016." [Online]. Available: <http://web.stanford.edu/~murmam/adcsurvey.html>
- [21] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," Apr. 2014. [Online]. Available: <http://arxiv.org/abs/1404.7736>
- [22] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, U.K., Jun. 2015, pp. 1304–1309.
- [23] —, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.
- [24] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.
- [25] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath Jr., "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Oct. 2016.
- [26] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.
- [27] A. Mezghani, R. Ghiat, and J. A. Nossek, "Transmit processing with low resolution D/A-converters," in *Proc. IEEE Int. Conf. Electron., Circuits, Syst. (ICECS)*, Yasmine Hammamet, Tunisia, Dec. 2009, pp. 683–686.
- [28] R. D. J. Guerreiro and P. Montezuma, "Use of 1-bit digital-to-analogue converters in massive MIMO systems," *IEEE Electron. Lett.*, vol. 52, no. 9, pp. 778–779, Apr. 2016.
- [29] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "On one-bit quantized ZF precoding for the multiuser massive MIMO downlink," in *IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, Rio de Janeiro, Brazil, Jul. 2016.
- [30] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE precoder for massive MIMO using 1-bit quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 3381–3385.
- [31] W. Zhang, "A remark on channels with transceiver distortion," in *Inf. Theory Applicat. Workshop (ITA)*, La Jolla, CA, USA, Jan.–Feb. 2016.
- [32] C. Studer and E. G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 303–313, Feb. 2013.
- [33] S. K. Mohammed and E. G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, Mar. 2013.
- [34] —, "Constant-envelope multi-user precoding for frequency-selective massive MIMO systems," *IEEE Commun. Lett.*, vol. 2, no. 5, pp. 547–550, Oct. 2013.
- [35] F. Maloberti, *Data converters*. Springer, Mar. 2007.
- [36] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.
- [37] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, 2013.
- [38] S. S. Christensen, E. C. Agarwal, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [39] 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.
- [40] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE transceiver optimization for multiuser MIMO systems: Duality and sum-MSE minimization," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5436–5446, Nov. 2007.
- [41] J. J. Busgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Elec., Cambridge, MA, Tech. Rep. 216, Mar. 1952.
- [42] 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.
- [43] N. Al-Dhahir and J. M. Cioffi, "On the uniform ADC bit precision and clip level computation for a Gaussian signal," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 434–438, Feb. 1996.
- [44] H. E. Rowe, "Memoryless nonlinearities with Gaussian inputs: Elementary results," *Bell Labs Tech. J.*, vol. 61, no. 7, pp. 1519–1525, Sep. 1982.
- [45] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.
- [46] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, University of Cambridge, U.K., Jun. 2014.
- [47] A. Lapidoth, "Mismatched decoding and the multiple access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep. 1996.
- [48] —, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1520–1529, Sep. 1996.
- [49] 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.
- [50] J. H. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proc. IEEE*, vol. 54, no. 1, pp. 2–19, Jan. 1966.
- [51] G. Jacovitti and A. Neri, "Estimation of the autocorrelation function of complex Gaussian stationary processes by amplitude clipped signals," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 239–245, Jan. 1994.
- [52] R. Coulliet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge: Cambridge Univ. Press, 2011.
- [53] S. Wagner, R. Coulliet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.
- [54] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
- [55] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.
- [56] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, no. 170, pp. 463–471, Apr. 1985.
- [57] S. Verdú, "Computational complexity of multiuser detection," *Algorithmica*, vol. 4, no. 1, pp. 303–312, 1989.
- [58] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
- [59] P. H. Tan and L. Rasmussen, "The application of semidefinite programming for detection in CDMA," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 8, pp. 1442–1449, Aug. 2001.
- [60] M. Bengtsson and B. Ottersten, "Optimal downlink beamforming using semidefinite optimization," in *Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, Sep. 1999.
- [61] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.
- [62] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.
- [63] O. Castañeda, T. Goldstein, and C. Studer, "Data detection in large multi-antenna wireless systems via approximate semidefinite relaxation," *IEEE Trans. Circuits Syst.*, vol. 63, no. 12, pp. 2334–2346, Dec. 2016.
- [64] N. Parikh and S. Boyd, *Proximal algorithms*. Now Publishers, 2013.
- [65] T. Goldstein, C. Studer, and R. Baraniuk, "A field guide to forward-backward splitting with a FASTA implementation," Feb. 2016. [Online]. Available: <https://arxiv.org/abs/1411.3406>
- [66] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, Dec. 1979.
- [67] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programming*, vol. 55, pp. 293–318, Apr. 1992.
- [68] C. Studer and H. Bölcskei, "Soft-input soft-output single tree-search sphere decoding," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4827–4842, Oct. 2010.
- [69] J. Jaldén and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.
- [70] D. Seethaler, J. Jaldén, C. Studer, and H. Bölcskei, "On the complexity distribution of sphere decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5754–5768, Sep. 2011.
- [71] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 763–767.
- [72] O. Castañeda, S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "1-bit massive MU-MIMO precoding in VLSI," Feb. 2017. [Online]. Available: <https://arxiv.org/abs/1702.03449>
- [73] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Massive MU-MIMO-OFDM downlink with one-bit DACs and linear precoding," Apr. 2017. [Online]. Available: <https://arxiv.org/abs/1704.04607>