Title

Reconsidering Linear Transmit Signal Processing in 1-Bit Quantized Multi-User MISO Systems

Journal

IEEE Transactions on Wireless Communications, 18(1)

ISSN

1536-1276

Authors

De Candido, Oliver
Jedda, Hela
Mezghani, Amine
et al.

Publication Date

2019

DOI

10.1109/twc.2018.2879106

Reconsidering Linear Transmit Signal Processing in 1-Bit Quantized Multi-User MISO Systems

Oliver De Candido, *Student Member, IEEE,* Hela Jedda, *Student Member, IEEE,*
Amine Mezghani, *Member, IEEE,* A. Lee Swindlehurst, *Fellow, IEEE,* and Josef A. Nossek, *Life Fellow, IEEE*

***Abstract*—In this contribution, we investigate a coarsely quantized Multi-User (MU)-Multiple Input Single Output (MISO) downlink communication system, where we assume 1-Bit Digital-to-Analog Converters (DACs) at the Base Station (BS) antennas. First, we analyze the achievable sum rate lower-bound using the Bussgang decomposition. In the presence of the non-linear quantization, our analysis indicates the potential merit of reconsidering traditional signal processing techniques in coarsely quantized systems, i.e., reconsidering transmit covariance matrices whose rank is equal to the rank of the channel. Furthermore, in the second part of this paper, we propose a linear precoder design which achieves the predicted increase in performance compared with a state of the art linear precoder design. Moreover, our linear signal processing algorithm allows for higher-order modulation schemes to be employed.**

***Index Terms*—1-bit digital-to-analog converters, downlink scenario, energy efficiency, multi-user multiple-input-single-output, quantized Wiener filter, superposition modulation, sum rate.**

I. Introduction and Motivation

In recent years, the demand for higher data rates has drastically increased as the number of personal and interconnected devices continuously increases (e.g., Internet of Things). These demands should be fulfilled by 5$^{\text{th}}$ Generation Wireless Systems (5G) under similar cost and energy constraints as current wireless communication systems. To this end, two complimentary technologies have been introduced at the forefront of research to provide the required data rates for 5G. First, the use of a large number of antennas at the Base Station (BS), referred to as massive Multiple Input Multiple Output (MIMO), has been investigated. Due to the inherent high antenna diversity and array gain, these systems have shown improvements in data throughput, spectral and radiated energy efficiency, using relatively simple processing, see e.g., [1]–[3]. Second is the use of millimeter wave (mmWave) carrier frequencies, where the large amount of available bandwidth will allow for higher data rates, e.g., [4], [5].

O. De Candido, H. Jedda and J. A. Nossek are with the Associate Professorship of Signal Processing, Department of Electrical and Computer Engineering, Technische Universität München, 80290 Munich, Germany (e-mail: {oliver.de-candido, hela.jedda, josef.a.nossek}@tum.de).

A. Mezghani is with the Wireless Networking and Communications Group, University of Texas at Austin, Austin, TX 78712, USA (e-mail: amine.mezghani@utexas.edu).

A. L. Swindlehurst is, and A. Mezghani was, with the Center for Pervasive Communications and Computing, University of California Irvine, Irvine, CA 92697, USA (e-mail: swindle@uci.edu).

J. A. Nossek is also with the Department of Teleinformatics Engineering, Universidade Federal do Ceará, Fortaleza 60020-181, Brasil.

Manuscript submitted to IEEE Transactions on Wireless Communications on February 27, 2018

If future communications systems are equipped with many BS antennas (hundreds or even thousands), and/or are working at higher sampling rates, the requirement for power and cost efficient components in the Radio Frequency (RF) chain at each antenna is evident. Currently, the most power hungry component in the RF chains are the Power Amplifiers (PAs), [6], [7]. PAs are most energy efficient when operated in their saturation region; however, in this region, they introduce non-linear distortions to the transmit signal. These distortions can be avoided when constant envelope input signals are employed, i.e., signals with a constant magnitude, and thus such amplitude distortions can be ignored. Furthermore, the power consumption of the Digital-to-Analog Converters (DACs)/Analog-to-Digital Converters (ADCs) in the RF chains increases exponentially with their resolution (in bits) and linearly with the sampling frequency, i.e., $P_{\text{diss}} \propto 2^b \cdot f_s$, [8]–[10]. Thus, a simple solution to reduce power consumption and chip area, whilst simultaneously employing constant envelope modulation, is to use low-resolution (or coarsely-quantized) DACs/ADCs. In this paper, we focus on the downlink scenario with the coarsest form of quantization, i.e., systems where the BS is equipped with 1-bit DACs.

It has been shown, see e.g., [11]–[13], that systems employing oversampling at the transmitters/receivers can improve the performance limitations introduced by the 1-bit DACs/ADCs. Furthermore, the issue of spectral shaping with 1-bit DACs and oversampling was investigated in [14], where it was shown that despite the low-resolution quantization, sufficient spectral confinement can be achieved. As we only consider spatial filtering and discrete-time processing, we focus on symbol-sampled models.

A. Existing Work

Recent research into the topic of coarsely-quantized MIMO systems can be categorized as focusing on either the uplink or the downlink scenario, where the BS is assumed to have low-resolution ADCs or DACs, respectively.

1) Uplink: The capacity of coarsely-quantized MIMO systems was originally investigated in [15], which showed only a small loss in capacity comparing quantized and unquantized MIMO systems. However, [15] and [16] show that coding becomes an issue, since traditional channel coding methods are unsuitable for quantized MIMO systems. In [17], the Taylor expansion of the mutual information up to the second-order is derived, which shows a $2/\pi$ loss in achievable rate at low-Signal-to-Noise Ratio (SNR). This loss, due to the use of

symmetric threshold quantizers, was also reported in [18]. Moreover, a mutual information lower-bound was derived in [19], based on the Bussgang theorem [20], which confirms the $2/\pi$ loss at low-SNR.

In [21] and [22], a closed-form expression for the capacity of the Single Input Single Output (SISO) and Multiple Input Single Output (MISO) uplink scenarios is derived, assuming perfect Channel State Information (CSI). Moreover, capacity bounds are found for the general MIMO scenario, and the mutual information lower-bound from [19] was shown to be tight at low-SNR but loose at high-SNR. Furthermore, [23] shows that higher-order modulation is possible with 1-bit quantized ADCs.

2) Downlink: A lower-bound for the achievable rate in quantized MIMO systems was derived in [24], assuming matched filter precoding and estimated CSI. Moreover, in [24], for single-antenna users, it was shown that roughly 2.5 times more BS antennas are required to achieve the same rates as in unquantized systems for maximum ratio precoding. In [25], the validity of traditional signal processing techniques was questioned for quantized single-user MISO systems, i.e., whether proper signaling and transmit covariance matrices whose rank is equal the rank of the channel matrix (channel rank) are still optimal in the presence of the non-linear quantization.

Recent research has also focused on linear and non-linear transmit signal processing techniques in quantized MIMO downlink systems; one of the first linear signal processing designs taking quantization into account was introduced in [26]. Therein, a Quantized Transmit Wiener Filter (TxWFQ) was designed using the optimal quantization step-size and linearizing the quantization operation. In [27], a linear precoder and an analog power allocation matrix were designed to minimize the Mean Squared Error (MSE) using a gradient projection algorithm. It should be noted that a precoder design using the optimal quantization step-size (e.g., [26]) is equivalent to using constant step-sizes and introducing an analog real-valued diagonal power allocation matrix (e.g., [27]). A linear precoder designed to maximize the weighted sum rate in a Multi-User (MU)-MISO system was introduced in [28], where the weighted sum rate is derived using a lower-bound on the achievable rate similar to [19]. In [29] an asymptotic analysis of MIMO scenarios is provided where the number of antennas and users increase to infinity. Moreover, [29] employs the Zero Forcing (ZF) precoder as a benchmark; an asymptotic achievable rate lower-bound is provided based on the Bussgang decomposition [20], and the authors show that reasonable performance can be obtained if the ratio of antennas to users is large enough. Applying simple perturbations to the solutions obtained by standard quantized linear precoders has also been shown to improve system performance in [30].

Non-linear precoders, which map the source symbols to the transmit vector in a general way, outperform linear precoders whose outputs are simply truncated by the one-bit quantization, however this comes at the price of higher computational complexity when designing the precoder. The first non-linear precoder design for low resolution quantized MIMO systems was introduced in [31], where the Tomlinson-Harashima Precoding method was extended to take the quantization into account. A novel, non-linear precoder design which optimizes the transmit signal vector by generating lookup-tables for each channel realization was introduced in [32]. Furthermore, in [33], an optimization to reduce the probability of detection error of Phase Shift Keying (PSK) symbols was introduced. This optimization is based on linear programming, and significantly reduces complexity compared to the lookup-table optimization in [32]. In [34], linear and non-linear precoding methods are investigated; it is shown that linear precoding methods only require 3 or 4 bit DACs to achieve performance similar to unquantized systems. Furthermore, three different non-linear algorithms which minimize the squared error are introduced, and only show a 3 dB loss compared with unquantized systems.

In [35] two non-linear precoder designs based on a biconvex relaxation of the MSE minimization is introduced, whereby the second algorithm is optimized to be scalable and have low-complexity with increasing number of BS antennas. A multi-step non-linear precoder design was introduced in [36] to reduce complexity for Quadrature Phase Shift Keying (QPSK) input symbols. First, a quantized linear precoder is applied, a subset of transmit antennas are selected and an exhaustive search over a subset-codebook is performed to optimize a criterion similar to [32]. A branch-and-bound approach to maximize the minimum distance to the decision boundary at the receivers is introduced in [37].

Finally, the non-linear algorithms described in [32] and [34] were extended to allow higher-order modulation schemes in [38] and [39]. In [38], two lookup-tables are generated per channel realization to allow for a superpositioning of the transmit symbols. In [39], linear and non-linear algorithms for higher order modulation schemes are introduced, including two algorithms to estimate the receiver scaling factor. These results show the potential of using higher-order modulation schemes despite the constraint of low-resolution DACs.

The aforementioned non-linear precoder designs optimize the transmit vector symbol-by-symbol at the sampling frequency, which greatly increases their computational complexity. Therefore, despite the performance gains of the non-linear methods, we are interested in investigating whether linear precoding methods can be improved.

B. Motivation

The motivation behind this work stems from the same question we asked in [25]: Are traditional signal processing techniques optimal in coarsely quantized MU-MISO systems?

1) Improper Signaling: First, traditional signal processing techniques often assume that all signals are circular Gaussian distributed, [40], i.e., $s \sim \mathcal{CN}(0, \sigma_s^2)$ with $\sigma^2_{\Re\{s\}} + \sigma^2_{\Im\{s\}} = \sigma_s^2$, where $\sigma^2_{\Re\{s\}} = \sigma^2_{\Im\{s\}}$ represent the variance of the real and imaginary parts, respectively. Moreover, the real and imaginary parts of s are assumed to be uncorrelated.

To motivate the question of whether circular Gaussian signaling is still optimal, we consider the following symmetrical non-convex optimization problem

$$\min_{x_1,x_2} \{|x_1| + |x_2|\} \quad \text{s.t.} \quad x_1^2 + x_2^2 = 1. \tag{1}$$

Despite the fact that (1) is symmetric with respect to (w.r.t.) the variables x_1 and x_2, i.e., exchanging the variables does not change the objective function nor the constraint, yet the extreme points, $x_{1,\text{opt}} = \pm 1$ and $x_{2,\text{opt}} = 0$ or $x_{1,\text{opt}} = 0$ and $x_{2,\text{opt}} = \pm 1$ are not equal, i.e., $x_{1,\text{opt}} \neq x_{2,\text{opt}}$. We can imagine that the constraint in (1) is the variance of the real and imaginary part of a complex signal s, i.e., $x_1 = \sigma_{\Re\{s\}}$ and $x_2 = \sigma_{\Im\{s\}}$. This would imply that the extreme points allow for unequal power allocation.

In general, optimization problems in quantized MIMO systems are non-convex due to the non-linearities and constraints introduced by the quantization. This simple example, of a symmetrical non-convex optimization problem motivated us to question the optimality of circular Gaussian distributed signals and proper signaling in quantized MIMO systems.

2) Higher-Rank Transmit Covariance Matrix: Second, traditional linear signal processing techniques typically assume that the transmit covariance matrix has the same, or lower, rank than the channel. As an example, we consider Fig. 1, in which an abstract, real-valued, noiseless quantized single-user MISO scenario is depicted, with the rank one channel vector $\boldsymbol{h}^{\mathrm{T}} = [1, \ldots, 1] \in \mathbb{R}^{1\times N_{\mathrm{t}}}$. The input signal $\boldsymbol{s}$ is fed into the linear precoder matrix $\boldsymbol{P}$, which we assume can be either a vector $\boldsymbol{p} \in \mathbb{R}^{N_{\mathrm{t}}}$ (a beamforming vector) or a full matrix $\boldsymbol{P} \in \mathbb{R}^{N_{\mathrm{t}}\times N_{\mathrm{t}}}$. Note, we consider linear precoders, the rank of the transmit covariance matrix $\boldsymbol{R}_{\boldsymbol{x}}$ is determined by the rank of the precoder matrix. The transmit signal is then passed through the 1-bit non-linear quantizers at all transmit antennas; these merely take the sign of the transmit signal, i.e., $\boldsymbol{Q}(\boldsymbol{x}) : \mathbb{R}^{N_{\mathrm{t}}} \to \{\pm 1\}^{N_{\mathrm{t}}}$.

Next, we consider the signal at the transmit antennas after the 1-bit DACs; here we see that in total we have $2^{N_{\mathrm{t}}}$ distinct transmit signals. However, if we assume a linear, channel rank precoder vector then the transmit signal is given by: $\boldsymbol{p}\cdot s \in \mathbb{R}^{N_{\mathrm{t}}}$, where only the sign of s affects the transmit signal. Thus, we restrict the system to only use 2 of the available $2^{N_{\mathrm{t}}}$ distinct transmit signals. This implies that the receive constellation yields only two points, $y \in \mathcal{Y} = \{\pm N_{\mathrm{t}}\}$, and the achievable rate is $I(x; y) \leq 1$ bit(s) per channel use (bpcu).

If, however, we increase the number of streams available, i.e., $\boldsymbol{s} = [s_1, \ldots, s_{\mathcal{R}}]$ with independent symbols s_i, and combine them with an augmented precoder matrix $\boldsymbol{P} \in \mathbb{R}^{N_{\mathrm{t}}\times\mathcal{R}}$ where the columns of $\boldsymbol{P}$ are linearly independent, then we can obtain more distinct transmit signals. In other words, by increasing the rank of the precoder matrix to $\operatorname{rank}(\boldsymbol{P}) = \mathcal{R}$, the maximum number of distinct receive constellation points becomes $|\mathcal{Y}| = \mathcal{R}+1$. Thus, using a full rank precoder matrix, i.e., $\boldsymbol{P} \in \mathbb{R}^{N_{\mathrm{t}}\times N_{\mathrm{t}}}$, the receive constellation has $|\mathcal{Y}| = N_{\mathrm{t}} + 1$ points and, in turn, we can achieve a rate closer to the capacity of the channel, $C \leq \log_2(N_{\mathrm{t}} + 1)$ bpcu, assuming a uniform distributed input signal.

This simple example motivated our study of whether higher-rank transmit covariance matrices can increase the system performance in the presence of the non-linear quantizers.

C. Contributions

In this paper, we analyze and investigate the optimality of two aspects of traditional signal processing in 1-bit quantized MU-MISO systems: (i) proper signaling, and/or (ii) channel rank transmit covariance matrices. We summarize our contributions presented in this paper for the downlink scenario as follows:

1) In the first part of the paper, we investigate the structure of the transmit covariance matrix $\boldsymbol{R}_{\tilde{\boldsymbol{x}}}$. We provide an achievable rate analysis, applying the Bussgang decomposition to investigate the sum rate lower-bound in 1-bit quantized MU-MISO systems. We investigate whether (i) improper signaling and/or (ii) higher-rank transmit covariance matrices maximize the sum rate lower-bound. This analysis indicates that higher-rank transmit covariance matrices can improve the sum rate lower-bound, whereas improper signaling may only marginally improve the system performance.
2) In the second part of the paper, we focus on the optimization of the linear precoder taking the results from our achievable rate analysis into account. In the end, we provide a gradient-projection algorithm to design a sub-optimal, higher-rank linear precoder. To obtain a higher-rank linear precoder, we introduce the idea of a linear superposition matrix which allows for linear superposition coding in 1-bit quantized MU-MISO systems. Moreover, our higher-rank linear precoder shows the predicted performance increase due to the increase in rank, compared with the linear precoder TxWFQ from [26].

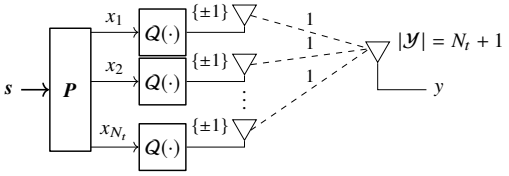

Fig. 1: Motivation: Higher-Rank Transmit Covariance Matrix

Our results indicate that indeed there are benefits in reconsidering signal processing methods for 1-bit quantized MU-MISO scenarios. Moreover, we provide an algorithm to design a linear precoder TxWFQ–$\boldsymbol{\Pi}$ whose rank is higher than the rank of the channel.

D. Paper Structure

The rest of the paper is structured as follows. In Section II the system model is introduced, including the necessary mathematical tools required for our analysis. In Section III we delve into the achievable rate analysis, and in Section IV we introduce our novel transmit precoder design taking the results from Section III into account. At the end of Section IV, we show the performance improvements introduced by our linear precoder design, investigate the complexity of our algorithm, and analyze the robustness of the algorithm against channel estimation errors. Finally, in Section V we conclude the paper by summarizing our main results and providing an outlook onto further work.

E. Notation

Scalars, vectors and matrices are denoted by italic letters, bold italic lowercase letters and bold italic uppercase letters,

respectively. The operators $(\cdot)^{\mathrm{T}}$, $\mathrm{tr}(\cdot)$, $\mathrm{E}[\cdot]$ $\Re\{\cdot\}$, $\Im\{\cdot\}$ represent the transpose, trace, expected value, real part and imaginary part, respectively. The notation $\mathrm{diag}(\boldsymbol{A})$ represents a diagonal matrix with the diagonal elements of $\boldsymbol{A}$, while $\mathrm{nondiag}(\boldsymbol{A})$ represents the matrix $\boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$. The matrix operation $\boldsymbol{A}^{\circ n}$ defines the Hadamard product to the nth power, i.e., $\boldsymbol{A} \circ \ldots \circ \boldsymbol{A}$, where $[\boldsymbol{A}^{\circ n}]_{i,j} = a_{i,j}^n$, which represents element-wise multiplication. The Kronecker product of two matrices is represented by $\boldsymbol{A} \otimes \boldsymbol{B}$. We use $\boldsymbol{I}_N$ and $\boldsymbol{0}_N$ to represent an $N \times N$ identity matrix and all-zero matrix, respectively.

Moreover, we introduce Widely-Linear (WL) notation (see e.g., [40]–[42]) to accommodate our analysis of whether proper signaling is optimal in quantized MU-MISO systems. To this end, we introduce the following definitions:

Definition 1 (Widely-Linear Vector): Taking a complex vector $\boldsymbol{a} \in \mathbb{C}^N$, we can express it in WL notation as

$$\bar{\boldsymbol{a}} = \begin{bmatrix} \Re\{\boldsymbol{a}\} \\ \Im\{\boldsymbol{a}\} \end{bmatrix} \in \mathbb{R}^{2N}. \tag{2}$$

Definition 2 (Strictly Linear Transformation): A transformation in the complex domain is strictly linear, i.e., $\boldsymbol{c} = \boldsymbol{B}\boldsymbol{a} \in \mathbb{C}^M \Leftrightarrow \bar{\boldsymbol{c}} = \bar{\boldsymbol{B}}\bar{\boldsymbol{a}} \in \mathbb{R}^{2M}$, if and only if (iff), in the real domain, the matrix $\bar{\boldsymbol{B}}$ has the following structure

$$\bar{\boldsymbol{B}} = \begin{bmatrix} \Re\{\boldsymbol{B}\} & -\Im\{\boldsymbol{B}\} \\ \Im\{\boldsymbol{B}\} & \Re\{\boldsymbol{B}\} \end{bmatrix} \in \mathbb{R}^{2M \times 2N}. \tag{3}$$

We define the real-valued covariance matrix of the arbitrary signal $\bar{\boldsymbol{a}}$ in WL notation as

$$\boldsymbol{R}_{\bar{\boldsymbol{a}}} = \begin{bmatrix} \mathrm{E}[\Re\{\boldsymbol{a}\}\Re\{\boldsymbol{a}^{\mathrm{T}}\}] & \mathrm{E}[\Re\{\boldsymbol{a}\}\Im\{\boldsymbol{a}^{\mathrm{T}}\}] \\ \mathrm{E}[\Im\{\boldsymbol{a}\}\Re\{\boldsymbol{a}^{\mathrm{T}}\}] & \mathrm{E}[\Im\{\boldsymbol{a}\}\Im\{\boldsymbol{a}^{\mathrm{T}}\}] \end{bmatrix}. \tag{4}$$

Definition 3 (Proper Signals): The signal $\boldsymbol{a}$ is proper iff both of the following conditions hold:

$$\mathrm{E}\left[\Re\{\boldsymbol{a}\}\,\Re\left\{\boldsymbol{a}^{\mathrm{T}}\right\}\right] = \mathrm{E}\left[\Im\{\boldsymbol{a}\}\,\Im\left\{\boldsymbol{a}^{\mathrm{T}}\right\}\right], \tag{5}$$

$$\mathrm{E}\left[\Re\{\boldsymbol{a}\}\,\Im\left\{\boldsymbol{a}^{\mathrm{T}}\right\}\right] = -\mathrm{E}\left[\Im\{\boldsymbol{a}\}\,\Re\left\{\boldsymbol{a}^{\mathrm{T}}\right\}\right]. \tag{6}$$

II. System Model

We consider the downlink scenario of a single-cell, coarsely quantized MU-MISO system as depicted in Fig. 2. The BS has N_{t} transmit antennas, each equipped with two 1-bit quantized DACs for the in-phase and quadrature signal components. The BS serves K single-antenna users simultaneously, and we assume the ADCs at the users have infinite quantization resolution. Furthermore, we assume that the BS and the users are fully synchronized, with their DACs and ADCs working at the same sampling frequency.

Thus, assuming narrowband channels, we can collect the real-valued baseband received signals at each user into a single vector representation

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{H}}^{\mathrm{T}}\tilde{\boldsymbol{D}}\bar{\boldsymbol{t}} + \bar{\boldsymbol{\eta}} \in \mathbb{R}^{2K}. \tag{7}$$

The vector $\bar{\boldsymbol{y}} \in \mathbb{R}^{2K}$ contains the received signals of all users, where $[\bar{y}_k, \bar{y}_{k+K}]^{\mathrm{T}} = [\Re\{y_k\}, \Im\{y_k\}]^{\mathrm{T}} \in \mathbb{R}^2$ represents the received signal of user k in WL notation. The strictly linear (see Def. 2) downlink channel matrix is denoted by $\bar{\boldsymbol{H}}^{\mathrm{T}} \in \mathbb{R}^{2K \times 2N_{\mathrm{t}}}$ in WL notation. We assume perfect CSI at the BS[1], i.e., the matrix $\bar{\boldsymbol{H}}^{\mathrm{T}}$ is perfectly known. Furthermore, we assume that the complex channel elements are circular symmetric independent and identically distributed (i.i.d.) Gaussian random variables with $h_{k,n} = [\boldsymbol{H}]_{k,n} \sim C\mathcal{N}_{\mathbb{C}}(0,1), \forall k,n$. The quantized transmit signal is $\bar{\boldsymbol{t}} \in \{\pm 1\}^{2N_{\mathrm{t}}}$, where we assume the output of the uniform 1-bit DACs is either ± 1.

The diagonal, real-valued power allocation matrix is denoted by $\tilde{\boldsymbol{D}} \in \mathbb{R}^{2N_{\mathrm{t}} \times 2N_{\mathrm{t}}}$. We assume that $\tilde{\boldsymbol{D}}$ does not necessarily have the strictly linear structure defined in Def. 2, so that the power can be allocated freely between the real and imaginary parts, which allows for improper signaling. Improper signaling can also be achieved by introducing correlation between the real and imaginary parts of the signals before the DACs, i.e., breaking the circular symmetry of the complex signals. Despite the fact that the power allocation matrix must be updated for every channel realization, one could still achieve constant envelope modulation per channel by feeding back a distinct scalar to the PAs at each antenna, which adjusts the supply voltage at each PA for a given channel, e.g., employing envelope tracking PAs (see, e.g., [43]). These supply voltages remain constant during each channel coherence time.

Finally, the Additive White Gaussian Noise (AWGN) is assumed to have the following distribution for all users: $\bar{\boldsymbol{\eta}} \sim C\mathcal{N}(0, \sigma_\eta^2/2 \cdot \boldsymbol{I}_{2K})$. The scaling factor $\beta \in \mathbb{R}_+$ at the receivers, introduced in [44], can be interpreted as an automatic gain control which is required to amplify the received symbol at each user such that it lies the correct decision region. We assume the scaling factor β changes relatively slowly and thus can be (perfectly) estimated over multiple received symbols by each user using blind methods [38], [39], allowing the users to employ minimum distance decoding.

The real-valued input signal $\bar{\boldsymbol{s}} \in \mathbb{R}^{2\mathcal{R}_{\mathrm{tot}}}$ contains the $2\mathcal{R}_{\mathrm{tot}}$ input symbols of all users. We introduce $\mathcal{R}_{\mathrm{tot}} = \sum_{k=1}^{K} \mathcal{R}_k$ as the sum of the number of streams per user. The variable $\mathcal{R}_k$ represents the number of streams each user receives, which, in turn, determines the increase in rank of the precoder matrix. It should be noted that if $\mathcal{R}_k = 1$, the precoder matrix will have the same rank as the channel. Moreover, $\mathcal{R}_k \leq N_{\mathrm{t}}$ holds because the maximum number of streams per user is upper bounded by the number of transmit antennas. Unless otherwise stated, the input signal is assumed to be Gaussian distributed with $\bar{\boldsymbol{s}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{\bar{\boldsymbol{s}}})$.

The transmit signal $\bar{\boldsymbol{x}} \in \mathbb{R}^{2N_{\mathrm{t}}}$ is the output of the precoder with input $\bar{\boldsymbol{s}}$, i.e., $\bar{\boldsymbol{x}} = \mathcal{P}(\bar{\boldsymbol{s}})$, where $\mathcal{P}(\cdot)$ is bijective but otherwise arbitrary and can be linear or non-linear. If we assume it to be a linear function, we do not restrict it to be strictly linear in the complex domain as defined in Def. 2, i.e., $\tilde{\boldsymbol{P}} \neq \bar{\boldsymbol{P}}$. This allows $\boldsymbol{R}_{\bar{\boldsymbol{x}}}$ to have the arbitrary structure as in (4).

We define the non-linear quantization function in the 1-bit case to take the sign of the input signal $\mathrm{sign}(\bar{\boldsymbol{x}})$, i.e.,

$$Q_t : \mathbb{R}^{2N_{\mathrm{t}}} \to \{\pm 1\}^{2N_{\mathrm{t}}}, \quad \bar{\boldsymbol{x}} \mapsto Q_t(\bar{\boldsymbol{x}}) = \mathrm{sign}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{t}}. \tag{8}$$

where the non-linear function $\mathrm{sign}(\cdot)$ is applied element-wise. Therefore, the total power across all transmit antennas after quantization is $\sum_{i=1}^{2N_{\mathrm{t}}} \mathrm{E}[|\bar{t}_i|^2] = 2N_{\mathrm{t}}$. If we define the total

[1]The impact of imperfect CSI will be studied later in the numerical results.

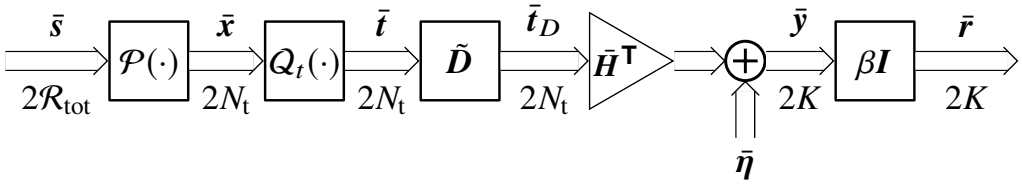

Fig. 2: Abstract Downlink Quantized MU-MISO System Model

available transmit power as E_{Tx}, and assume the power is equally allocated over all antennas, then the power allocation matrix must be a scaled identity matrix, i.e.,

$$\tilde{\boldsymbol{D}} = \sqrt{\frac{E_{\text{Tx}}}{2N_{\text{t}}}}\boldsymbol{I}_{2N_{\text{t}}}. \tag{9}$$

With this power allocation matrix we allow for improper signaling by introducing correlation between the transmit signal at different antennas. Thus, the total power after the power allocation matrix is equal to $\sum_{i=1}^{2N_{\text{t}}} \text{E}[|\bar{t}_{D,i}|^2] = E_{\text{Tx}}$.

A. Bussgang Decomposition

Similar to previous work, e.g., [19], [29], we model the quantization function using the Bussgang decomposition [20]. According to the Bussgang theorem, the cross-correlation between two Gaussian distributed input signals remains the same when one signal is subjected to non-linear distortion, except for a scaling factor. This implies that a non-linear function with Gaussian inputs can be modeled by a linear transformation and the addition of some distortion which is uncorrelated with the inputs.

The transmit signal becomes approximately Gaussian distributed as the number of users increases due to the central limit theorem. Therefore, we assume $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{\bar{\boldsymbol{x}}})$ when K is large enough. Thus, using the Bussgang theorem the quantization function in (8) can be modeled as

$$\bar{\boldsymbol{t}} = Q_t(\bar{\boldsymbol{x}}) = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{q}, \tag{10}$$

where the quantization error, $\boldsymbol{q}$ is uncorrelated with the input signal $\bar{\boldsymbol{x}}$. From the latter criterion we see that

$$\text{E}\left[\boldsymbol{q}\bar{\boldsymbol{x}}^{\text{T}}\right] = \boldsymbol{0}_{2N_{\text{t}}} \Rightarrow \boldsymbol{A} = \boldsymbol{R}_{\bar{t}\bar{x}}\boldsymbol{R}_{\bar{x}}^{-1}. \tag{11}$$

Thus, we observe that the matrix $\boldsymbol{A}$ is simply a linear Minimum Mean Squared Error (MMSE) estimate of the quantized signal $\bar{\boldsymbol{t}}$ from the unquantized input signal $\bar{\boldsymbol{x}}$. Moreover, the Bussgang decomposition depends on the covariance matrix between the quantized signal $\bar{\boldsymbol{t}}$ and the unquantized signal $\bar{\boldsymbol{x}}$.

Finally, we can express the covariance matrix of the quantization error as

$$\boldsymbol{R}_{\boldsymbol{q}} + \boldsymbol{A}\boldsymbol{R}_{\bar{x}}\boldsymbol{A}^{\text{T}} \overset{(11)}{=} \boldsymbol{R}_{\bar{t}} - \boldsymbol{R}_{\bar{t}\bar{x}}\boldsymbol{R}_{\bar{x}}^{-1}\boldsymbol{R}_{\bar{x}\bar{t}}. \tag{12}$$

B. Price's Theorem – Quantized Covariance Matrices

To calculate the covariance matrices $\boldsymbol{R}_{\bar{\boldsymbol{t}}} = \text{E}\left[\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}^{\text{T}}\right]$ and $\boldsymbol{R}_{\bar{t}\bar{x}} = \text{E}\left[\bar{\boldsymbol{t}}\bar{\boldsymbol{x}}^{\text{T}}\right]$ we apply Price's theorem [45], (for more details see e.g., [46, Sec. II]). To this end, the covariance matrix of the quantized output signal is, e.g., [47, p. 307],

$$\boldsymbol{R}_{\bar{t}} = \frac{2}{\pi}\arcsin\left(\text{diag}\left(\boldsymbol{R}_{\bar{\boldsymbol{x}}}\right)^{-1/2}\boldsymbol{R}_{\bar{\boldsymbol{x}}}\text{diag}\left(\boldsymbol{R}_{\bar{\boldsymbol{x}}}\right)^{-1/2}\right), \tag{13}$$

where the factor $2/\pi$ comes from the fixed quantization levels and the real-valued function $\arcsin(\boldsymbol{A})$ is defined element-wise on the matrix argument $\boldsymbol{A}$. The covariance matrix between the input and output of the 1-bit quantizer can equally be calculated by applying Price's theorem

$$\boldsymbol{R}_{\bar{t}\bar{\boldsymbol{x}}} = \sqrt{\frac{2}{\pi}}\text{diag}\left(\boldsymbol{R}_{\bar{\boldsymbol{x}}}\right)^{-1/2}\boldsymbol{R}_{\bar{\boldsymbol{x}}}. \tag{14}$$

Moreover, due to the real-valued WL notation, the following relationship holds: $\boldsymbol{R}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{t}}} = \boldsymbol{R}_{\bar{t}\bar{\boldsymbol{x}}}^{\text{T}}$.

III. Achievable Rate Analysis

In this section, we investigate the achievable rate in 1-bit quantized MU-MISO systems by looking at the structure of the transmit covariance matrices $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}$. For our achievable rate analysis, we assume that the total transmit power is constant $E_{\text{Tx}} = 2N_{\text{t}}$, and the SNR at the receiver is varied by changing the noise variance σ_η^2. With equal power allocation, we have $\tilde{\boldsymbol{D}} = \boldsymbol{I}_{2N_{\text{t}}}$. Moreover, since we are investigating the mutual information and not the signal processing techniques in this section, we assume the receiver scaling factor to be one, $\beta = 1$. We assume that the CSI is perfectly known at the BS and at the users.

Using the Bussgang decomposition defined in Section II-A and the covariance matrices defined in Section II-B, we can express the real-valued received signal at user k from (7) as

$$\begin{aligned}\bar{\boldsymbol{y}}_k &= \bar{\boldsymbol{H}}_k^{\text{T}}(\boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{q}) + \bar{\boldsymbol{\eta}}_k \\ &= \bar{\boldsymbol{H}}_{\text{eff},k}^{\text{T}}\bar{\boldsymbol{x}} + \tilde{\boldsymbol{\eta}}_k,\end{aligned} \tag{15}$$

where $\bar{\boldsymbol{H}}_k^{\text{T}}$ represents the strictly linear channel matrix of user k. We recall that the received signal at each single-antenna user and the total transmit signal are expressed in WL notation from Def. 1.

Moreover, we introduce the effective channel $\bar{\boldsymbol{H}}_{\text{eff},k}^{\text{T}} = \bar{\boldsymbol{H}}_k^{\text{T}}\boldsymbol{A} \overset{(11)}{=} \bar{\boldsymbol{H}}_k^{\text{T}}\boldsymbol{R}_{\bar{t}\bar{x}}\boldsymbol{R}_{\bar{x}}^{-1}$, and the effective noise $\tilde{\boldsymbol{\eta}}_k = \bar{\boldsymbol{H}}_k^{\text{T}}\boldsymbol{q} + \bar{\boldsymbol{\eta}}_k$, which is no longer Gaussian due to the quantization error, with $\boldsymbol{R}_{\bar{x}}$ and $\boldsymbol{R}_{\bar{t}\bar{x}}$ defined in (13) and (14), respectively. Furthermore, we express the transmit signal $\bar{\boldsymbol{x}}$ as the sum of the transmit signals intended for each user, and we only consider coding schemes where the transmit signals for each user are independent, i.e.,

$$\bar{\boldsymbol{x}} = \sum_{k=1}^{K}\bar{\boldsymbol{x}}_k \quad \Rightarrow \quad \boldsymbol{R}_{\bar{\boldsymbol{x}}} = \sum_{k=1}^{K}\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}, \tag{16}$$

where $\bar{\boldsymbol{x}}_k$ and $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}$ represent the transmit signal and transmit covariance matrix intended for user k prior to the DACs, respectively.

A. Sum Rate Lower-Bound

Now, we aim to calculate the mutual information between the signal intended for user k and the signal that user receives, i.e., $I(\bar{\boldsymbol{x}}_k; \bar{\boldsymbol{y}}_k) = h(\bar{\boldsymbol{y}}_k) - h(\bar{\boldsymbol{y}}_k|\bar{\boldsymbol{x}}_k)$, with the continuous entropy function $h(\cdot)$. We assume perfect knowledge of the CSI at the BS and at the users, and no cooperation between the users. The encoding at the BS does not use the non-causally known interference of the user's signals, i.e., we do not employ dirty paper coding, and the users decode the received signal by treating the Multi-User Interference (MUI) as noise.

We first focus on the second continuous entropy term

$$\begin{aligned} h(\bar{\boldsymbol{y}}_k|\bar{\boldsymbol{x}}_k) &= h\left(\bar{\boldsymbol{H}}_{\text{eff},k}^{\text{T}} \sum_{k=1}^{K} \bar{\boldsymbol{x}}_k + \tilde{\boldsymbol{\eta}}_k \middle| \bar{\boldsymbol{x}}_k\right) \\ &\overset{(a)}{\leq} h\left(\bar{\boldsymbol{H}}_{\text{eff},k}^{\text{T}} \sum_{\substack{l=1 \\ l\neq k}}^{K} \bar{\boldsymbol{x}}_l + \tilde{\boldsymbol{\eta}}_k\right), \end{aligned} \tag{17}$$

where inequality (a) comes from the fact that conditioning cannot increase entropy, [48, Th. 2.6.5], and holds with equality if $\tilde{\boldsymbol{\eta}}_k$ and $\bar{\boldsymbol{x}}_k$ are statistically independent. Moreover, the addition of a constant term does not change the entropy, i.e., $h(\bar{\boldsymbol{x}}_k|\bar{\boldsymbol{x}}_k) = 0$. However, despite the fact that $\boldsymbol{q}$ and $\bar{\boldsymbol{x}}_k$ are uncorrelated, they may still be dependent. The total noise is $\bar{\boldsymbol{H}}_{\text{eff},k}^{\text{T}} \sum_{l=1,l\neq k}^{K} \bar{\boldsymbol{x}}_l + \tilde{\boldsymbol{\eta}}_k$, which contains the MUI, quantization error $\boldsymbol{q}$, and the AWGN.

Furthermore, in [49] (see also [24]), it was shown that for a given noise covariance matrix, Gaussian distributed noise minimizes the mutual information in a given system. Thus, assuming Gaussian distributed inputs and total noise from (17), we can write the instantaneous mutual information lower-bound of the Gaussian system as

$$I(\bar{\boldsymbol{x}}_k; \bar{\boldsymbol{y}}_k) \geq \frac{1}{2} \log_2 \det(\boldsymbol{I}_2 + \text{SQINR}_k), \tag{18}$$

where the Signal-to-Quantization-plus-Interference-plus-Noise Ratio $(\text{SQINR})_k$ defined in (19) (on the next page) shows the contribution of the MUI, Quantization Error (QE) and AWGN. The identity matrix $\boldsymbol{I}_2$ comes from the fact that we consider the real and imaginary parts separately.

With the mutual information lower-bound defined per user in (18), we can now express the instantaneous sum rate lower-bound by summing over all $k = 1, \ldots, K$:

$$\sum_{k=1}^{K} I(\bar{\boldsymbol{x}}_k; \bar{\boldsymbol{y}}_k) \geq \frac{1}{2} \sum_{k=1}^{K} \log_2 \det(\boldsymbol{I}_2 + \text{SQINR}_k). \tag{20}$$

Finally, we optimize for the transmit covariance matrices which maximize the sum rate lower-bound from (20)

$$\boldsymbol{R}_{\bar{\boldsymbol{x}}_k,\text{opt}} = \underset{\boldsymbol{R}_{\bar{\boldsymbol{x}}_k} \succeq \boldsymbol{0}, \forall k}{\arg\max} \left\{ \sum_{k=1}^{K} \log_2 \det(\boldsymbol{I}_2 + \text{SQINR}_k) \right\}, \tag{21}$$

where the optimization is performed over all positive semi-definite transmit covariance matrices for all users, i.e., $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k} \succeq \boldsymbol{0}, \forall k$. Since we assume perfect CSI at the BS and at the users, we calculate the ergodic achievable sum rate as the average of the maximum sum rates achieved per channel realization, i.e., we average the sum rate lower-bounds over the channel realizations [50, Sec. II-C1].

Since the argument of the $\log_2 \det(\cdot)$ function in the sum rate depends non-linearly on the user's transmit covariance matrix (see (13)), we wish to investigate whether traditional signal processing techniques still maximize (21), i.e., whether channel rank transmit covariance matrices and proper signaling are still optimal.

B. Cholesky Decomposition

First, we note that

$$\text{rank}\left(\boldsymbol{R}_{\bar{\boldsymbol{x}}}\right) = \text{rank}\left(\sum_{k=1}^{K} \boldsymbol{R}_{\bar{\boldsymbol{x}}_k}\right) \leq \sum_{k=1}^{K} \text{rank}\left(\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}\right), \tag{22}$$

and we define the Cholesky decompositon of $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}$ to be $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k} = \boldsymbol{L}_k(\mathcal{R}_k)\boldsymbol{L}_k^{\text{T}}(\mathcal{R}_k)$, where $\mathcal{R}_k$ denotes the number of streams for user k, which determines the rank of $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}$. The rank of the transmit covariance matrix can be varied by observing the structure of the Cholesky factor:

$$\boldsymbol{L}_k(\mathcal{R}_k) = \begin{bmatrix} l_{1,1} & 0 & \ldots & 0 & \ldots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots & \ddots & \vdots \\ l_{3,1} & \ddots & l_{2\mathcal{R}_k,2\mathcal{R}_k} & 0 & \ldots & 0 \\ l_{4,1} & \ddots & l_{2(\mathcal{R}_k+1),2\mathcal{R}_k} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ l_{2N_{\text{t}},1} & \ldots & l_{2N_{\text{t}},2\mathcal{R}_k} & 0 & \ldots & 0 \end{bmatrix}, \tag{23}$$

where $l_{i,i} > 0$ for $i \leq \mathcal{R}_k \leq N_{\text{t}}$ and $l_{i,j} \in$ for $i > j$, $j \leq \mathcal{R}_k \leq N_{\text{t}}$. Finally, we can restate (21) as

$$\boldsymbol{L}_{k,\text{opt}}(\mathcal{R}_k) = \underset{\boldsymbol{L}_k(\mathcal{R}_k), \forall k}{\arg\max} \left\{ \frac{1}{2} \sum_{k=1}^{K} \log_2 \det(\boldsymbol{I}_2 + \text{SQINR}_k) \right\}, \tag{24}$$

where the SQINR_k is now parameterized in terms of $\boldsymbol{L}_k(\mathcal{R}_k)$, for a given $\mathcal{R}_k$, instead of the transmit covariance matrix $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k}$. When considering proper signaling, we can also restrict user k to employ proper signaling by further restricting the covariance matrix $\boldsymbol{R}_{\bar{\boldsymbol{x}}_k} = \boldsymbol{L}_k(\mathcal{R}_k)\boldsymbol{L}_k(\mathcal{R}_k)^{\text{T}}$ to fulfill (5) and (6) from Def. 3.

C. Simulation Results: Achievable Rate

In this section, we provide numerical results for our achievable rate analysis in 1-bit quantized MU-MISO systems. We simulated a downlink scenario with a $N_{\text{t}} = 16$ antenna BS and $K = 2$ single antenna users, solving the optimization in (24) numerically for both proper and improper transmit covariance matrices. In this scenario, the rank of the user's transmit covariance matrices must be $\mathcal{R}_k \leq N_{\text{t}} = 16$. In our simulations we assumed the rank of each users' transmit covariance matrix was equal, i.e., $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}$, and set $\mathcal{R} = 1$ and 2. We observed that if we increased the rank beyond 2, the additional streams per user led to worse performance. We plot the ergodic sum rate lower-bound by averaging the sum rate

$$\mathrm{SQINR}_k = \left(\underbrace{\bar{\boldsymbol{H}}_{\mathrm{eff},k}^{\mathrm{T}} \sum_{l=1,l\neq k}^{K} \boldsymbol{R}_{\bar{\boldsymbol{x}}_l} \bar{\boldsymbol{H}}_{\mathrm{eff},k}}_{\text{MUI}} + \underbrace{\bar{\boldsymbol{H}}_k^{\mathrm{T}} \boldsymbol{R}_{\boldsymbol{q}} \bar{\boldsymbol{H}}_k}_{\text{QE}} + \underbrace{\boldsymbol{R}_{\bar{\boldsymbol{\eta}}_k}}_{\text{AWGN}} \right)^{-1} \bar{\boldsymbol{H}}_{\mathrm{eff},k}^{\mathrm{T}} \boldsymbol{R}_{\bar{\boldsymbol{x}}_k} \bar{\boldsymbol{H}}_{\mathrm{eff},k} \tag{19}$$

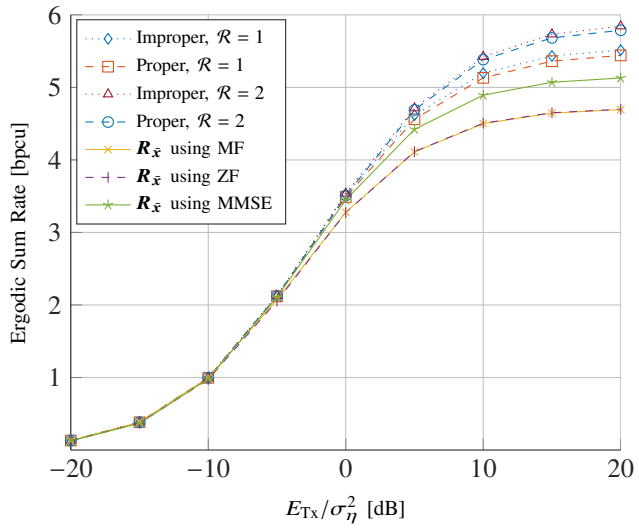

Fig. 3: MU-MISO Downlink lower-bound of the ergodic sum rate with $N_{\mathrm{t}} = 16$ and $K = 2$, averaged over 200 i.i.d. channels.

lower-bounds over 200 i.i.d. channel realizations where we set $E_{\mathrm{Tx}} = 2N_{\mathrm{t}}$ and varied the noise variance $\sigma_\eta^2 \in \{-20, \ldots, 20\}$ dB.

We plot the performance of different transmit covariance matrices in Fig. 3, comparing our optimized covariance matrices, $\boldsymbol{R}_{\bar{\boldsymbol{x}}_{k,\mathrm{opt}}} = \boldsymbol{L}_{k,\mathrm{opt}} \boldsymbol{L}_{k,\mathrm{opt}}^{\mathrm{T}}$, with the traditional strictly linear Matched Filter (MF), ZF and MMSE precoders. We observe that at low-SNR our optimized transmit covariance matrices converge to those employing traditional signal processing techniques. This indicates that traditional signal processing methods, i.e., channel rank transmit covariance matrices and proper signaling, may be optimal at low-SNR in the MU scenario, similar to the Single-User (SU) scenario (see [25, Th. 1]).

At mid- to high-SNR we observe that the optimized covariance matrices diverge from the traditional signal processing techniques. The transmit covariance matrix using the MMSE precoder shows a higher sum rate lower-bound at high-SNR compared with the other two traditional signal processing techniques, but our optimized covariance matrices provide better performance. The gain at higher-SNR for the improper and proper solutions with channel rank, i.e., $\mathcal{R} = 1$, is due to the fact that our solutions further mitigate the MUI.

Furthermore, we observe that when the rank of both users' transmit covariance matrices is higher than the rank of the channel, i.e., $\mathcal{R} = 2$, the sum rate lower-bound is the highest. This indicates that higher-rank transmit covariance matrices can achieve better performance in 1-bit quantized MU-MISO scenarios. These results concur with our results in the SU-MISO scenario in [25, Sec. V]. Moreover, improper signaling only seems to marginally improve the performance.

We can summarize the results as follows: (i) higher-rank transmit covariance matrices maximize the sum rate lower-bound, and (ii) improper signaling only marginally improves the sum rate lower-bound. Thus, in the following we will attempt to optimize a linear precoder matrix which has a rank higher than the rank of the channel, and further investigate whether improper signaling can improve the uncoded-Bit Error Rate (BER) or MSE performance.

IV. Transmit Signal Processing

In this section, we move on from our investigation of the transmit covariance matrix and introduce a linear precoder design taking the results from Section III into account. Therefore, we assume that the precoder function is a linear precoder matrix $\tilde{\boldsymbol{P}} \in \mathbb{R}^{2N_{\mathrm{t}} \times 2\mathcal{R}_{\mathrm{tot}}}$. Note that the WL precoder matrix can have an arbitrary structure and not the Strictly-Linear (SL) structure defined in Def. 2, i.e., $\tilde{\boldsymbol{P}} \neq \bar{\boldsymbol{P}}$. To this end, we express the received signal as

$$\bar{\boldsymbol{r}} = \beta \bar{\boldsymbol{y}} = \beta \left(\bar{\boldsymbol{H}}^{\mathrm{T}} \tilde{\boldsymbol{D}} \boldsymbol{Q}_t (\tilde{\boldsymbol{P}} \bar{\boldsymbol{s}}) + \bar{\boldsymbol{\eta}} \right) \in \mathbb{R}^{2K}. \tag{25}$$

A. Superposition Matrix

With the definition of the received signal in (25) we can define the MSE, ε, as

$$\varepsilon = \mathrm{E}\left[\| \bar{\boldsymbol{r}} - \boldsymbol{\Pi} \bar{\boldsymbol{s}} \|_2^2 \right], \tag{26}$$

where we introduce the linear superposition matrix $\boldsymbol{\Pi}$ which allows the users to receive symbols from higher-order constellations than those transmitted (see e.g., [38], [51]–[53]).

In our transmitter signal processing design we take the results from [39, Fig. 2(a)] into account which show that with a linear precoder (ZF) and 1-bit DACs at the BS, QPSK transmit symbols show the best uncoded-BER performance. Therefore, we assume that the input signal in our system are QPSK for all users.

Assuming all users receive the same constellation, i.e., $\mathcal{R}_k = \mathcal{R}\ \forall k$, the linear superposition matrix describing higher order Quadrature Amplitude Modulation (QAM) based on QPSK is defined as

$$\boldsymbol{\Pi} = \boldsymbol{I}_{2K} \otimes \boldsymbol{\tau}^{\mathrm{T}} \in \mathbb{R}^{2K \times 2K\mathcal{R}} \tag{27}$$

and the superposition row vector $\boldsymbol{\tau}^{\mathrm{T}}$ is defined as

$$\boldsymbol{\tau}^{\mathrm{T}} = \begin{bmatrix} 2^{\mathcal{R}-1} & 2^{\mathcal{R}-2} & \ldots & 2^1 & 2^0 \end{bmatrix} \in \mathbb{R}^{1 \times \mathcal{R}}. \tag{28}$$

The superposition vector $\boldsymbol{\tau}^{\mathrm{T}} \in \{1, 2, 4, \ldots, 2^{\mathcal{R}-1}\}^{1 \times \mathcal{R}}$ has a length $\mathcal{R}$ per user which determines the rank of each user's

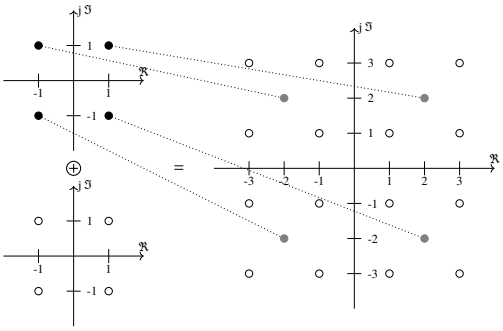

Fig. 4: Linear Superposition Matrix – Two QPSK symbols to one 16–QAM symbol

precoder. The maximum rank of the precoder is equal to the number of transmit antennas, i.e., $\mathcal{R} \leq N_\mathrm{t}$.

To clarify how our linear superposition matrix works, assume $\mathcal{R} = 2$ which implies that each user receives 16-QAM symbols, and we have $\boldsymbol{\tau}^\mathrm{T} = [2, 1]$ as per (28). We observe in Fig. 4 how the superposition vector $\boldsymbol{\tau}^\mathrm{T} = [2, 1]$ works; (i) the first symbol (solid points) is multiplied by a factor 2 and defines which quadrant the received symbol should lie in, (ii) the second symbol (hollow points) is added to the first and defines which 16-QAM symbol should be received. Since we assume QPSK input symbols and the specific superposition matrix defined in (27), the superimposed received symbols will be M-QAM, where M depends on the chosen number of streams per user, $\mathcal{R}_k$.

B. MSE Definition

With the superposition matrix defined in (27), we can express the MSE as

$$\begin{aligned}\varepsilon =&\beta^2 \frac{2}{\pi}\mathrm{tr}\left(\bar{\boldsymbol{H}}^\mathrm{T}\tilde{\boldsymbol{D}}\arcsin(\boldsymbol{P}'\boldsymbol{P}'^\mathrm{T})\tilde{\boldsymbol{D}}\bar{\boldsymbol{H}}\right) + \beta^2\mathrm{tr}\left(\boldsymbol{R}_{\bar{\boldsymbol{\eta}}}\right) + \mathrm{tr}\left(\boldsymbol{\Pi}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\boldsymbol{\Pi}^\mathrm{T}\right)\\ &- 2\beta\sqrt{\frac{2}{\pi}}\mathrm{tr}\left(\bar{\boldsymbol{H}}^\mathrm{T}\tilde{\boldsymbol{D}}\boldsymbol{P}'\boldsymbol{R}_{\bar{\boldsymbol{s}}}^{1/2}\boldsymbol{\Pi}^\mathrm{T}\right),\end{aligned} \tag{29}$$

where we have inserted the covariance matrices of the quantized signals defined in (13) and (14); also we define the normalized precoding matrix $\boldsymbol{P}'$ as

$$\boldsymbol{P}' = \mathrm{diag}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right)^{-1/2}\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}^{1/2}. \tag{30}$$

C. Transmit Wiener Filter Design

In this subsection we introduce our algorithm to calculate a higher-rank version of the TxWFQ from [26]. First, we define the optimization problem as follows

$$\{\tilde{\boldsymbol{P}}_\mathrm{opt}, \beta_\mathrm{opt}, \tilde{\boldsymbol{D}}_\mathrm{opt}\} = \underset{\tilde{\boldsymbol{P}},\beta,\tilde{\boldsymbol{D}}}{\arg\min}\,\{\varepsilon\} \;\text{ s.t. }\; \begin{array}{l}\mathrm{E}\left[\left\|\bar{\boldsymbol{t}}_D\right\|_2^2\right] \leq E_\mathrm{Tx},\\ \tilde{\boldsymbol{D}} \in \mathbb{R}^{2N_\mathrm{t}\times 2N_\mathrm{t}} \text{ is diagonal},\end{array} \tag{31}$$

with ε defined in (29) and the sum power constraint is applied after the power allocation matrix $\tilde{\boldsymbol{D}}$ (see Fig. 2).

Intuitively, we understand that the power allocated to the transmit antennas by the precoder is normalized back to unit power by the 1-bit DACs. Therefore, we choose $\tilde{\boldsymbol{D}}$ to restore the desired power allocation of the precoder by setting

$$\tilde{\boldsymbol{D}}_\mathrm{opt} = \mathrm{diag}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right)^{1/2}. \tag{32}$$

With the choice of the power allocation matrix $\tilde{\boldsymbol{D}}_\mathrm{opt}$ defined in (32), we can rewrite the optimization problem as

$$\{\tilde{\boldsymbol{P}}_\mathrm{opt}, \beta_\mathrm{opt}\} = \underset{\tilde{\boldsymbol{P}},\beta}{\arg\min}\,\{\varepsilon\} \quad \text{s.t.} \quad \begin{array}{l}\mathrm{tr}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right) \leq E_\mathrm{Tx},\\ \tilde{\boldsymbol{D}}_\mathrm{opt} = \mathrm{diag}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right)^{1/2},\end{array} \tag{33}$$

where the sum-power constraint comes from the fact that $\bar{\boldsymbol{t}}_D = \tilde{\boldsymbol{D}}_\mathrm{opt}\bar{\boldsymbol{t}}$ and using the optimal power allocation matrix from (32).

D. Arcsine Approximation

We note that due to the non-linear matrix function $\arcsin(\cdot)$ in the MSE expression (29), the derivative of ε w.r.t. $\tilde{\boldsymbol{P}}$ proves difficult to solve for in closed form. Therefore, we use the second-order Taylor expansion of the off-diagonal elements defined as: $\arcsin(x) \approx x + 1/6 \cdot x^3$. The diagonal elements are given by: $\arcsin\left(\mathrm{diag}\left(\boldsymbol{P}'\boldsymbol{P}'^\mathrm{T}\right)\right) = \arcsin(\boldsymbol{I}_{2N_\mathrm{t}}) = \pi/2 \cdot \boldsymbol{I}_{2N_\mathrm{t}}$. Thus, the matrix $\arcsin(\cdot)$ function can be approximated as

$$\arcsin(\boldsymbol{P}'\boldsymbol{P}'^\mathrm{T}) \approx \boldsymbol{P}'\boldsymbol{P}'^\mathrm{T} + \frac{1}{6}\left(\boldsymbol{P}'\boldsymbol{P}'^\mathrm{T}\right)^{\circ 3} + \left(\frac{\pi}{2} - \frac{7}{6}\right)\boldsymbol{I}_{2N_\mathrm{t}}, \tag{34}$$

where $\boldsymbol{A}^{\circ n}$ represents the matrix Hadamard product to the power n. We use the second-order Taylor expansion since we want to retain the non-linearities introduced by the coarse quantization to observe the performance gains from higher-rank transmit covariance matrices. Therefore, we can substitute the optimal power allocation matrix from (32) and the approximation from (34) into the MSE expression from (29) to arrive at

$$\begin{aligned}\varepsilon \approx&\beta^2\frac{2}{\pi}\mathrm{tr}\left(\bar{\boldsymbol{H}}^\mathrm{T}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T} + \left(\frac{\pi}{2} - \frac{7}{6}\right)\mathrm{diag}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right)\right)\bar{\boldsymbol{H}}\right)\\ &+ \beta^2\frac{2}{\pi}\frac{1}{6}\mathrm{tr}\left(\bar{\boldsymbol{H}}^\mathrm{T}\left(\tilde{\boldsymbol{D}}_\mathrm{opt}^{-2}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^\mathrm{T}\right)^{\circ 3}\tilde{\boldsymbol{D}}_\mathrm{opt}^{-2}\right)\bar{\boldsymbol{H}}\right)\\ &- 2\beta\sqrt{\frac{2}{\pi}}\mathrm{tr}\left(\bar{\boldsymbol{H}}^\mathrm{T}\tilde{\boldsymbol{P}}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\boldsymbol{\Pi}^\mathrm{T}\right) + \beta^2\mathrm{tr}\left(\boldsymbol{R}_{\bar{\boldsymbol{\eta}}}\right) + \mathrm{tr}\left(\boldsymbol{\Pi}\boldsymbol{R}_{\bar{\boldsymbol{s}}}\boldsymbol{\Pi}^\mathrm{T}\right).\end{aligned} \tag{35}$$

E. Gradient Projection Algorithm

Since the optimization problem in (33) is non-convex and non-linear w.r.t. the precoder matrix $\tilde{\boldsymbol{P}}$, solving (33) in closed-form is intractable. Consequently, we use a gradient-projection algorithm [54, p. 466] to iteratively solve for a locally optimal solution, with a projection back onto the feasible set. The gradient-projection algorithm we implement is outlined in Algorithm 1.

First, we initialize our algorithm with a random full rank matrix $\tilde{\boldsymbol{P}}_{(0)}$ and calculate the initial scaling factor $\beta_{(0)}$ using the function $g^*(\tilde{\boldsymbol{P}})$. This function finds the optimal scaling factor for a given precoder matrix and will be introduced in Appendix A (see (42)). We use an initial constant step-size of $\gamma = 10$, but we allow for backtracking in Step 9, where we

Algorithm 1 Gradient Projection Algorithm to Solve for the Higher-Rank, WL TxWFQ–**Π**

1: **Initialization:**
2: $\tilde{\boldsymbol{P}}_{(0)}$, $\beta_{(0)} \leftarrow g^*(\tilde{\boldsymbol{P}}_{(0)})$, $\gamma = 10$ and $n = 0$
3: **repeat**
4: **if** $\varepsilon_{(n+1)} \leq \varepsilon_{(n)}$ **then**
5: $\tilde{\boldsymbol{P}}_{(n+1)} \leftarrow \mathcal{P}_C\left(\tilde{\boldsymbol{P}}_{(n)} - \gamma \frac{\partial \varepsilon\left(\tilde{\boldsymbol{P}}_{(n)}, \beta^*_{(n)}\right)}{\partial \tilde{\boldsymbol{P}}}\right)$
6: $\beta^*_{(n+1)} \leftarrow g^*\left(\tilde{\boldsymbol{P}}_{(n)}\right)$ ▷ defined in Appendix A
7: $n \leftarrow n + 1$
8: **else**
9: $\gamma \leftarrow \gamma/2$
10: **end if**
11: **until** $|\varepsilon_{(n+1)} - \varepsilon_{(n)}|/\varepsilon_{(n)} \leq \delta$

halve the gradient step-size if the MSE in iteration $(n+1)$ is larger than in the current iteration (n), which is checked in Step 4.

In Step 5 we update the precoder by taking a step in the direction of the MSE gradient w.r.t. the precoder, where the derivative term is defined in Appendix A. Here, the projection function $\mathcal{P}_C(\cdot)$ ensures that the sum-power constraint $\mathrm{tr}\left(\tilde{\boldsymbol{P}}\boldsymbol{R}_{\tilde{\boldsymbol{s}}}\tilde{\boldsymbol{P}}^{\mathrm{T}}\right) \leq E_{\mathrm{Tx}}$ is fulfilled in each iteration. The optimal scaling factor is updated in Step 6 using the function $g^*(\tilde{\boldsymbol{P}})$ defined in Appendix A. Our algorithm runs until the stopping criterion is met, which is triggered once the relative difference in MSE from the previous iteration is less than a predefined threshold, δ.

F. Simulation Results: Signal Processing

In this subsection, we present simulation results for the TxWFQ–**Π** precoder introduced above. We compare our linear precoder design with the TxWFQ design from [26]. We note the following facts about the precoder design from [26]: (i) it has channel rank, (ii) it is strictly linear, and (iii) the authors further optimize the quantization output step-sizes, i.e., the outputs of the DACs are not uniform. In the end, the resulting power allocation in [26] is equivalent to the optimal power allocation matrix $\tilde{\boldsymbol{D}}_{\mathrm{opt}}$ we introduced in (32). Additionally, we plot the optimal Transmit Wiener Filter (TxWF) introduced in [44], simulating the TxWF in an unquantized scenario, i.e., assuming the DACs at the BS have infinite quantization resolution, which we will refer to as TxWF (unq.).

We assume that the BS has $N_{\mathrm{t}} = 128$ transmit antennas, which serves $K = 4$ single antenna users. Our simulation results assume a constant noise covariance matrix $\boldsymbol{R}_{\boldsymbol{\eta}} = \boldsymbol{I}_K$, and vary the transmit power $E_{\mathrm{Tx}} \in \{0, \ldots, 21\}$ dB. In our simulations, we used a block length of $N_{\mathrm{b}} = 10,000$ symbols and averaged over 200 i.i.d. channel realizations. We terminate our algorithms with the value $\delta = 10^{-4}$.

We plot two solutions for our precoder design, a WL and a SL solution. The SL solution has the same structure as defined in Def. 2. To obtain a SL solution we use the fact that the SL structure in Def. 2 is maintained under multiplication, addition, transposition and inversion as shown in [42] and [55]. Therefore, if we initialize our algorithm with a SL matrix, i.e., $\tilde{\boldsymbol{P}}_{(0)} = \bar{\boldsymbol{P}}_{(0)}$ with the structure from Def. 2, then the resulting solution will be SL.

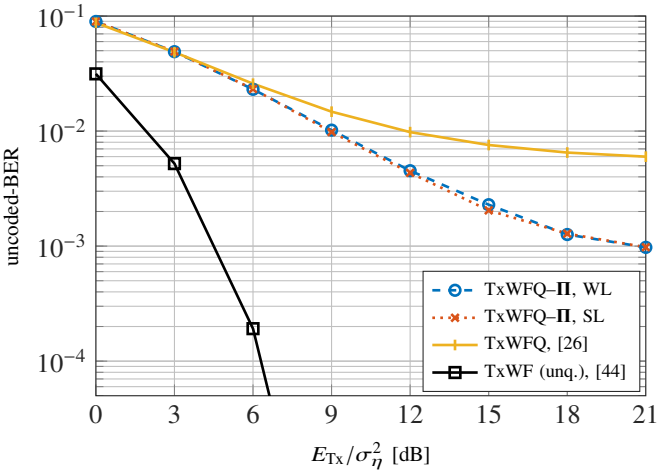

Fig. 5: MU-MISO Downlink 16-QAM uncoded-BER using Alg. 1 with $N_{\mathrm{t}} = 128$ and $K = 4$, averaged over 200 i.i.d. channels.

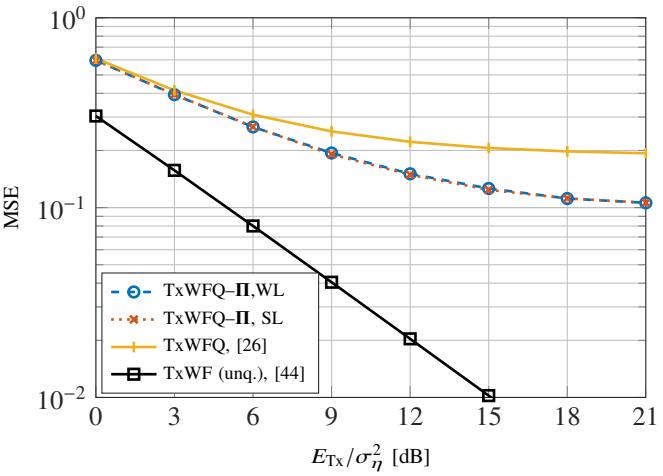

Fig. 6: MU-MISO Downlink 16-QAM MSE using Alg. 1 with $N_{\mathrm{t}} = 128$ and $K = 4$, averaged over 200 i.i.d. channels.

First, we present uncoded-BER and MSE results using 16-QAM symbols in Fig. 5 and Fig. 6, respectively. To receive 16-QAM symbols, the BS transmits two QPSK symbols to each user using the superposition vector from (28) as $\boldsymbol{\tau}^{\mathrm{T}} = [2, 1]$, implying $\mathcal{R} = 2$. Thus, the rank of each user's precoder matrix is twice the rank of the channel. In Fig. 5, we observe that our precoder design from Alg. 1 outperforms the linear TxWFQ method at higher SNR. We observe that this increase in uncoded-BER and MSE performance, compared with the TxWFQ design from [26], is due to the increase in rank of our TxWFQ–**Π** precoder design. However, there appears to be no gain from using improper signaling, i.e., the WL solution performed as well as the SL solution.

Interestingly, both the WL and the SL solutions turn out to be independent of the random initialization. Our solutions achieve an uncoded-BER of 10^{-2} at around 9 dB, which is roughly 3 dB better than TxWFQ. Compared with the unquantized TxWF we see roughly a 7 dB performance loss

$E_{\mathrm{Tx}}/\sigma_\eta^2$ [dB]	3	6	9	12	15	18	21
Alg. 1 WL	33	86	111	137	166	209	254
Alg. 1 SL	30	91	117	138	170	201	248

TABLE I: Alg. 1 with $N_{\mathrm{t}} = 128$ and $K = 4$ with 16-QAM Average Number of Iterations

due to the quantization. In Fig. 6, we also observe similar performance gains in terms of MSE for our higher-rank precoder design over the whole SNR range.

1) Complexity Analysis: In Table I, we show the average number of iterations, including the back-tracking steps, for Alg. 1 using either a WL or SL initialization. We observe that the number of iterations grows almost linearly with the SNR and at high-SNR roughly 250 iterations are required to achieve a relative MSE difference of $\delta = 10^{-4}$. Moreover, we see that, on average, both the WL and SL solutions require roughly the same number of iterations to converge.

Moreover, we take a closer look at the computational complexity of Alg. 1 and calculate an asymptotic upper bound on the number of floating-point operations (FLOPs) required. We observe that most of the computational complexity comes from calculating the derivative of the MSE w.r.t. the precoder matrix (derived in Appendix A). Due to the term in (45), we see that the asymptotic upper bound is:

$$\mathcal{O}\left(N_{\mathrm{t}}^2 \cdot K \cdot \mathcal{R}_{\mathrm{tot}}\right) \text{ FLOPs},$$

which is quadratic in the number of antennas but linear in the number of users and total number of streams per user. It should be noted that our derivation of the derivative of the MSE w.r.t. the precoder was not optimized to consider the number of FLOPs required, and there may be more efficient implementations.

2) Channel State Estimation Error: Thus far, we have assumed perfect CSI at the BS. In the following, we investigate how sensitive our algorithm is to CSI estimation errors. To this end, we introduce the estimated channel matrix $\bar{\boldsymbol{H}}_{\mathrm{est}}$ as

$$\bar{\boldsymbol{H}}_{\mathrm{est}} = \sqrt{1-\xi}\bar{\boldsymbol{H}} + \sqrt{\xi}\bar{\boldsymbol{\Gamma}}, \tag{36}$$

where $\xi \in [0,1]$ and $\left[\bar{\boldsymbol{\Gamma}}\right]_{i,j} \sim \mathcal{CN}(0,1), \forall i,j$. The variable ξ represents the variance of the channel estimation error, where a value $\xi = 0$ is equivalent to a system without estimation error, i.e., perfect CSI, and $\xi = 1$ is a fully erroneous channel estimation, i.e., where the BS has no CSI. Intermediate values of ξ represent partial CSI estimation errors.

We plot the sensitivity of our algorithm against CSI estimation error in Fig. 7 for 16-QAM received symbols at $E_{\mathrm{Tx}}/\sigma_\eta^2 = 12$ dB. We observe that our algorithm shows a slightly better performance compared with the TxWFQ solution for all ξ values, although larger performance gains are seen with smaller CSI estimation errors, since an increase in CSI estimation error can also be seen as a decrease in SNR.

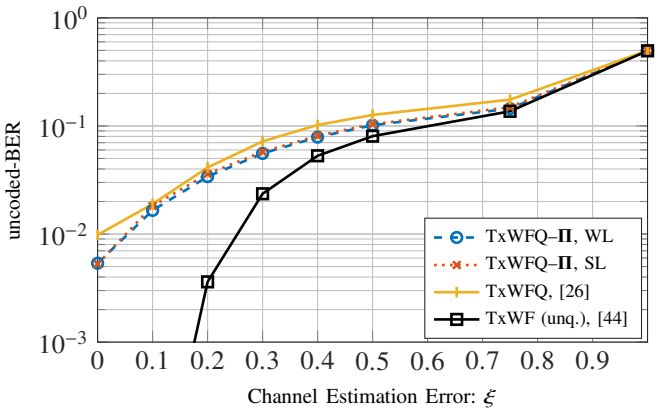

Fig. 7: MU-MISO Downlink uncoded-BER vs. Channel Estimation Error employing 16-QAM modulation at $E_{\mathrm{Tx}}/\sigma_\eta^2 = 12$ dB.

V. Conclusion

In this paper, we reconsider linear transmit signal processing methods in 1-bit quantized MU-MISO downlink scenarios using an achievable rate analysis. Our results indicate that higher-rank precoders can increase the lower-bound of the achievable sum rate. By taking these results into account, we developed an algorithm to design a higher-rank linear precoder. The derived precoder achieved performance superior to a state-of-the-art linear signal processing method with channel rank, both in terms of uncoded-BER and MSE. These gains were due to the higher rank of the linear precoders; we observed no additional gain by employing improper signaling. For 16-QAM symbols, we observe a 3 dB gain for an uncoded-BER of 10^{-2} over traditional linear signal processing techniques for a system with $N_{\mathrm{t}} = 128$ BS antennas and $K = 4$ single antenna users.

Non-linear precoding methods where the transmit vector is optimized symbol-by-symbol, (e.g., [32]–[39]), show uncoded-BER and MSE performance even closer to the unquantized TxWF. However, they require much higher computational complexity as they must work at the sampling rate and scale with the number of transmit antennas. In comparison, the linear precoder design presented here motivates reconsidering traditional linear precoder designs to improve the system performance with low complexity. Moreover, the linear precoder matrix only has to be calculated once per channel coherence time instead of for each input symbol, which drastically reduces the computational complexity.

To extend the work presented in this paper, one could analyze a system using higher resolution DACs, still assuming constant envelope modulation. Additionally, one could optimize the superposition matrix $\mathbf{\Pi}$. In the end, $\mathbf{\Pi}$ determines the increase in rank of the precoder matrix, and also depends on the users' channels which could be taken into account during the optimization. Finally, it would be interesting to extend the work to frequency selective channels employing Orthogonal Frequency-Division Multiplexing (OFDM); initial results show that OFDM can be implemented with low-resolution DACs and linear processing (e.g., [13], [56]).

Appendix
Derivations in Algorithm 1

In this Appendix we briefly derive the various functions and derivatives required in Alg. 1. Note that, in the following

derivations we will drop the iteration index (n) for notational brevity. First, we restate the MSE term from (29) as

$$\varepsilon = \beta^2 \left(a(\tilde{\boldsymbol{P}}) + b(\tilde{\boldsymbol{P}}) + d \right) - 2\beta c(\tilde{\boldsymbol{P}}) + e, \tag{37}$$

where we define the following functions

$$a(\tilde{\boldsymbol{P}}) \coloneqq \frac{2}{\pi} \mathrm{tr}\left(\bar{\boldsymbol{H}}^{\mathrm{T}} \left(\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}^{\mathrm{T}} + \left(\frac{\pi}{2} - \frac{7}{6} \right) \mathrm{diag}\left(\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}^{\mathrm{T}} \right) \right) \bar{\boldsymbol{H}} \right) \tag{38}$$

$$b(\tilde{\boldsymbol{P}}) \coloneqq \frac{2}{\pi} \frac{1}{6} \mathrm{tr}\left(\bar{\boldsymbol{H}}^{\mathrm{T}} \left(\tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-2} \left(\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}^{\mathrm{T}} \right)^{\circ 3} \tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-2} \right) \bar{\boldsymbol{H}} \right) \tag{39}$$

$$c(\tilde{\boldsymbol{P}}) \coloneqq -\sqrt{\frac{2}{\pi}} \mathrm{tr}\left(\bar{\boldsymbol{H}}^{\mathrm{T}} \tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \boldsymbol{\Pi}^{\mathrm{T}} \right) \tag{40}$$

$$d = \mathrm{tr}\left(\boldsymbol{R}_{\bar{\boldsymbol{\eta}}} \right) \text{ and } e = \mathrm{tr}\left(\boldsymbol{\Pi} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \boldsymbol{\Pi}^{\mathrm{T}} \right). \tag{41}$$

We note that the functions $a(\tilde{\boldsymbol{P}})$ and $b(\tilde{\boldsymbol{P}})$ come from the second-order Taylor expansion of the non-linear $\arcsin(\cdot)$ function (see Section IV-D). Moreover, we recall that the optimal power allocation matrix $\tilde{\boldsymbol{D}}_{\mathrm{opt}}$ is also a function of the precoder matrix.

With the approximate MSE expression from (37), we can define the function $g^*(\tilde{\boldsymbol{P}})$ by setting $\partial\varepsilon/\partial\beta = 0$, yielding

$$g^*\left(\tilde{\boldsymbol{P}}\right) \coloneqq \frac{-c(\tilde{\boldsymbol{P}})}{a(\tilde{\boldsymbol{P}}) + b(\tilde{\boldsymbol{P}}) + d}, \tag{42}$$

with $a(\tilde{\boldsymbol{P}}), b(\tilde{\boldsymbol{P}}), c(\tilde{\boldsymbol{P}})$ and d defined in (38), (39), (40) and (41), respectively.

Next, we calculate the derivative of the MSE w.r.t. the precoding matrix as

$$\frac{\partial \varepsilon}{\partial \tilde{\boldsymbol{P}}} = \beta^2 \left(\frac{\partial a(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}} + \frac{\partial b(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}} \right) - 2\beta \frac{\partial c(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}}, \tag{43}$$

with $a(\tilde{\boldsymbol{P}}), b(\tilde{\boldsymbol{P}})$ and $c(\tilde{\boldsymbol{P}})$ defined in (38), (39) and (40), respectively. Closed-form expressions of the derivative terms in (43) can be written as

$$\frac{\partial a(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}} = 2\frac{2}{\pi} \left[\bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^{\mathrm{T}} + \left(\frac{\pi}{2} - \frac{7}{6} \right) \mathrm{diag}\left(\bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^{\mathrm{T}} \right) \right] \tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tag{44}$$

$$\begin{aligned} \frac{\partial b(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}} =& 2\frac{2}{\pi} \left[\frac{1}{2} (\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}^{\mathrm{T}})^{\circ 2} \circ \mathrm{nondiag}\left(\tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-2} \bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^{\mathrm{T}} \tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-2} \right) \right. \\ & - \frac{1}{3} \mathrm{diag}\left(\left(\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}^{\mathrm{T}} \right)^{\circ 3} \tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-2} \bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^{\mathrm{T}} \right) \tilde{\boldsymbol{D}}_{\mathrm{opt}}^{-4} \\ & \left. + \frac{1}{2} \mathrm{diag}\left(\bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^{\mathrm{T}} \right) \right] \tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \end{aligned} \tag{45}$$

$$\frac{\partial c(\tilde{\boldsymbol{P}})}{\partial \tilde{\boldsymbol{P}}} = -2\sqrt{\frac{2}{\pi}} \bar{\boldsymbol{H}} \boldsymbol{\Pi} \boldsymbol{R}_{\bar{\boldsymbol{s}}}. \tag{46}$$

Finally, the projection function $\mathcal{P}_C(\cdot)$ is simply defined as the normalization: $\mathcal{P}_C\left(\tilde{\boldsymbol{P}}\right) \coloneqq \sqrt{E_{\mathrm{Tx}}/\mathrm{tr}\left(\tilde{\boldsymbol{P}} \boldsymbol{R}_{\bar{\boldsymbol{s}}} \tilde{\boldsymbol{P}}\right)} \cdot \tilde{\boldsymbol{P}}$.

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb 2014.

[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct 2014.

[3] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, Feb 2013.

[4] W. Hong, K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 63–69, Sept 2014.

[5] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[6] O. Blume, D. Zeller, and U. Barth, "Approaches to Energy Efficient Wireless Access Networks," in *Proceedings of 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*. Limassol, Cyprus: IEEE, March 2010.

[7] T. Chen, H. Kim, and Y. Yang, "Energy Efficiency Metrics For Green Wireless Communications," in *Proceedings of 2010 International Conference on Wireless Communications Signal Processing (WCSP)*. Suzhou, China: IEEE, Oct 2010.

[8] B. Murmann, "ADC Performance Survey 1997-2016," [Online]. Available: http://web.stanford.edu/ murmann/adcsurvey.html., 2015.

[9] C. Svensson, S. Andersson, and P. Bogner, "On The Power Consumption of Analog to Digital Converters," in *Proceedings of 2006 24th Norchip Conference*. Linköping, Sweden: IEEE, Nov 2006, pp. 49–52.

[10] R. H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, April 1999.

[11] S. Krone and G. Fettweis, "Capacity of communications channels with 1-bit quantization and oversampling at the receiver,," in *Proceedings of 2012 35th IEEE Sarnoff Symposium*. Newark, NJ, USA: IEEE, May 2012.

[12] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," in *Proceedings of 2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel*. Eliat, Israel: IEEE, Nov 2010.

[13] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Massive MU-MIMO-OFDM Downlink with One-Bit DACs and Linear Precoding," in *Proceedings of 2017 IEEE Global Communications Conference (GLOBECOM)*. Singapore, Singapore: IEEE, Dec 2017.

[14] H. Jedda, A. Mezghani, and J. A. Nossek, "Spectral shaping with low resolution signals," in *Proceedings of 2015 49th Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA, USA: IEEE, Nov 2015.

[15] J. A. Nossek and M. T. Ivrlač, "Capacity and Coding for Quantized MIMO Systems," in *Proceedings of 2006 International Conference on Wireless Communications and Mobile Computing (IWCMC)*, ser. IWCMC '06. New York, NY, USA: ACM, July 2006, pp. 1387–1392. [Online]. Available: http://doi.acm.org/10.1145/1143549.1143827

[16] M. T. Ivrlač and J. A. Nossek, "Challenges in Coding for Quantized MIMO Systems," in *Proceedings of 2006 IEEE International Symposium on Information Theory (ISIT)*. Seattle, WA, United States: IEEE, July 2006, pp. 2114–2118.

[17] A. Mezghani and J. A. Nossek, "On Ultra-Wideband MIMO Systems with 1-bit Quantized Outputs: Performance Analysis and Input Optimization," in *Proceedings of 2007 IEEE International Symposium on Information Theory (ISIT)*. Nice, France: IEEE, June 2007, pp. 1286–1289.

[18] T. Koch and A. Lapidoth, "At Low SNR, Asymmetric Quantizers are Better," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, Sept 2013.

[19] A. Mezghani and J. A. Nossek, "Capacity Lower Bound of MIMO Channels with Output Quantization and Correlated Noise," in *Proceedings of 2012 IEEE International Symposium on Information Theory (ISIT)*. Cambridge, MA, USA: IEEE, July 2012.

[20] J. Bussgang, "Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals," Res. Lab. of Electron., MIT, Cambridge, MA, USA, Tech. Rep. No. 216, March 1952.

[21] J. Mo and R. W. Heath, "High SNR Capacity of Millimeter Wave MIMO Systems with One-Bit Quantization," in *Proceedings of 2014 Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, Feb 2014.

[22] ——, "Capacity Analysis of One-Bit Quantized MIMO Systems with Transmitter Channel State Information," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5498–5512, Oct 2015.

[23] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-Bit Massive MIMO: Channel Estimation and High-Order Modulations," in *Proceedings of 2015 IEEE International Conference on*

Communication Workshop (ICCW). London, UK: IEEE, June 2015, pp. 1304–1309.

[24] Y. Li, C. Tao, A. L. Swindlehurst, A. Mezghani, and L. Liu, "Downlink Achievable Rate Analysis in Massive MIMO Systems with One-Bit DACs," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1669–1672, July 2017.

[25] O. De Candido, H. Jedda, A. Mezghani, A. L. Swindlehurst, and J. A. Nossek, "Are Traditional Signal Processing Techniques Rate Maximizing in Quantized SU-MISO Systems?" in *Proceedings of 2017 IEEE Global Communications Conference (GLOBECOM)*. Singapore, Singapore: IEEE, Dec 2017.

[26] A. Mezghani, R. Ghiat, and J. A. Nossek, "Transmit Processing with Low Resolution D/A-Converters," in *Proceedings of 2009 16th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. Yasmine Hammamet, Tunisia: IEEE, Dec 2009, pp. 683–686.

[27] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE Precoder for Massive MIMO Using 1-bit Quantization," in *Proceedings of 2016 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, March 2016, pp. 3381–3385.

[28] A. Kakkavas, J. Munir, A. Mezghani, H. Brunner, and J. A. Nossek, "Weighted Sum Rate Maximization for Multi-User MISO Systems with Low Resolution Digital to Analog Converters," in *Proceedings of 2016 20th International ITG Workshop on Smart Antennas (WSA)*. Munich, Germany: VDE, March 2016.

[29] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of One-Bit Quantized Precoding for the Multiuser Massive MIMO Downlink," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4624–4634, Sept 2017.

[30] A. Swindlehurst, A. Saxena, A. Mezghani, and I. Fijalkow, "Minimum Probability-Of-Error Perturbation Precoding for the One-Bit Massive MIMO Downlink," in *Proceedings of 2017 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA, USA: IEEE, March 2017, pp. 6483–6487.

[31] A. Mezghani, R. Ghiat, and J. A. Nossek, "Tomlinson Harashima Precoding for MIMO Systems with Low Resolution D/A-converters," in *Proceedings of 2008 ITG/IEEE Workshop on Smart Antennas (WSA)*. Darmstadt, Germany: IEEE, Feb 2008.

[32] H. Jedda, J. A. Nossek, and A. Mezghani, "Minimum BER Precoding in 1-bit massive MIMO Systems," in *Proceedings of 2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Rio de Janeiro, Brazil: IEEE, July 2016.

[33] H. Jedda, A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, "Massive MIMO Downlink 1-Bit Precoding with Linear Programming for PSK Signaling," *arXiv preprint arXiv:1704.06426*, 2017.

[34] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized Precoding for Massive MU-MIMO," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4670–4684, Nov 2017.

[35] O. Castañeda, S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "1-bit Massive MU-MIMO Precoding in VLSI," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 508–522, Dec 2017.

[36] O. Tirkkonen and C. Studer, "Subset-codebook precoding for 1-bit massive multiuser MIMO," in *Proceedings of 2017 51st Annual Conference on Information Sciences and Systems (CISS)*. Baltimore, MD, USA: IEEE, March 2017.

[37] L. T. N. Landau and R. C. de Lamare, "Branch-and-Bound Precoding for Multiuser MIMO Systems With 1-Bit Quantization," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 770–773, Dec 2017.

[38] D. B. Amor, H. Jedda, and J. Nossek, "16 QAM Communication with 1-Bit Transmitters," in *Proceedings of 2017 21th International ITG Workshop on Smart Antennas (WSA)*. Berlin, Germany: VDE, March 2017.

[39] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-Bit Precoding for Massive MU-MIMO with Higher-Order Modulation," in *Proceedings of 2016 50th Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA, USA: IEEE, Nov 2016, pp. 763–767.

[40] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-Valued Signal Processing: The Proper Way to Deal With Impropriety," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, Nov 2011.

[41] D. P. Mandic and V. S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*, 1st ed. John Wiley & Sons, 2009.

[42] C. Hellings and W. Utschick, "Block-Skew-Circulant Matrices in Complex-Valued Signal Processing," *IEEE Transactions on Signal Processing*, vol. 63, no. 8, pp. 2093–2107, April 2015.

[43] F. H. Raab, P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "Power Amplifiers and Transmitters for RF and Microwave," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 814–826, March 2002.

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

[45] R. Price, "A Useful Theorem For Nonlinear Devices Having Gaussian Inputs," *IRE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, June 1958.

[46] K. Roth, J. Munir, A. Mezghani, and J. A. Nossek, "Covariance based signal parameter estimation of coarse quantized signals," in *Proceedings of 2015 IEEE International Conference on Digital Signal Processing (DSP)*. Singapore, Singapore: IEEE, July 2015, pp. 19–23.

[47] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. Tata McGraw-Hill Education, 1998.

[48] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[49] S. N. Diggavi and T. M. Cover, "The Worst Additive Noise Under a Covariance Constraint," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 3072–3081, Nov 2001.

[50] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, June 2003.

[51] S. Huan, Z. Fei, L. Huang, and J. Kuang, "Cooperative Transmission Utilizing High Order Superposition Modulation with Iterative Detection," in *Proceedings of 2009 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*. Beijing, China: IEEE, Sept 2009.

[52] H. Sun, S. X. Ng, and L. Hanzo, "Superposition Coded Modulation for Cooperative Communications," in *Proceedings of 2012 IEEE Vehicular Technology Conference (VTC Fall)*. Quebec City, QC, Canada: IEEE, Sept 2012.

[53] S. Wang and B. K. Yi, "Optimizing Enhanced Hierarchical Modulations," in *Proceedings of 2008 IEEE Global Communications Conference (GLOBECOM)*. New Orleans, LA, USA: IEEE, Nov 2008.

[54] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[55] C. Hellings and W. Utschick, "Iterative Algorithms for Transceiver Design in MIMO Broadcast Channels with Improper Signaling," in *Proceedings of 2015 10th International ITG Conference on Systems, Communications and Coding (SCC)*. Hamburg, Germany: VDE, Feb 2015.

[56] J. Guerreiro, R. Dinis, and P. Montezuma, "Use of 1-bit digital-to-analogue converters in massive MIMO systems," *Electronics Letters*, vol. 52, no. 9, pp. 778–779, April 2016.