Title

Massive MIMO 1-Bit DAC Transmission: A low-complexity symbol scaling approach

Permalink

https://escholarship.org/uc/item/90x0j8nk

Journal

IEEE Transactions on Wireless Communications, 17(11)

ISSN

1536-1276

Authors

Li, A
Masouros, C
Liu, F
et al.

Publication Date

2018-11-01

DOI

10.1109/TWC.2018.2868369

Massive MIMO 1-Bit DAC Transmission: A Low-Complexity Symbol Scaling Approach

Ang Li, *Student Member, IEEE*, Christos Masouros, *Senior Member, IEEE*, Fan Liu, *Student Member, IEEE*, and A. Lee Swindlehurst, *Fellow, IEEE*

***Abstract*—We study multi-user massive multiple-input single-output (MISO) systems and focus on downlink transmission, where the base station (BS) employs a large antenna array with low-cost 1-bit digital-to-analog converters (DACs). The direct combination of existing beamforming schemes with 1-bit DACs is shown to lead to an error floor at medium-to-high SNR regime, due to the coarse quantization of the DACs with limited precision. In this paper, based on the constructive interference we consider both a quantized linear beamforming scheme where we analytically obtain the optimal beamforming matrix, and a non-linear mapping scheme where we directly design the transmit signal vector. Due to the 1-bit quantization, the formulated optimization for the non-linear mapping scheme is shown to be non-convex. To solve this problem, the non-convex constraints of the 1-bit DACs are firstly relaxed, followed by an element-wise normalization to satisfy the 1-bit DAC transmission. We further propose a low-complexity symbol scaling scheme that consists of three stages, in which the quantized transmit signal on each antenna element is selected sequentially. Numerical results show that the proposed symbol scaling scheme achieves a comparable performance to the optimization-based non-linear mapping approach, while its corresponding complexity is negligible compared to that of the non-linear scheme.**

***Index Terms*—Massive MIMO, 1-bit quantization, beamforming, constructive interference, Lagrangian, low-complexity scheme.**

I. Introduction

TOWARDS the fifth generation (5G) and future wireless communication systems, massive multiple-input multiple-output (MIMO) systems [1] have received increasing research attention in recent years as they are able to greatly improve the spectral efficiency. It has also been shown that low-complexity linear precoding approaches such as zero-forcing (ZF) [2] and regularized ZF (RZF) [3] achieve close-to-optimal performance in the massive MIMO regime. Nevertheless, with a large number of antennas employed at the BS, the large number of radio frequency (RF) chains and corresponding digital-to-analog converters (DACs) that need to be employed at the BS pose a significant practical challenge. This increase in the hardware complexity and resulting power consumption hinders the practical implementation of massive MIMO. To achieve a compromise between the performance, hardware complexity and the consequent power consumption in practical massive MIMO systems, hybrid analog digital beamforming [4], [5] has attracted research interest as a means of reducing the number of RF chains.

In addition to the hybrid structures, another potential approach, which is the focus of this paper, is to reduce the cost and power consumption per RF chain by employing very low-resolution digital-to-analog converters (DACs) instead of high-precision DACs. It has been shown in [6] that DACs are one of the dominant power-consuming hardware components in the downlink, whose power consumption grows exponentially with the resolution and linearly with the bandwidth. In the traditional MIMO downlink, each transmit signal is generated by a pair of high-resolution (usually more than 8-bit) DACs that are connected to the RF chain. However, in the case of massive MIMO with hundreds of antennas employed at the BS, a large number of DACs are required and the resulting power consumption will be prohibitively high. Therefore, employing low-resolution DACs, especially 1-bit DACs, can greatly reduce the power consumption per RF chain and the resulting total power consumed at the BS. When 1-bit DACs are employed, the output signal at each antenna element is equivalent to the constant-envelope symbol from a QPSK constellation, which enables the use of low-cost power amplifiers (PAs) and can further reduce the hardware complexity.

In the existing literature, most recent studies have focused on the performance analysis for massive MIMO uplink with low-resolution analog-to-digital converters (ADCs), especially for the 1-bit case [7]-[9], where it is shown that the number of quantization bits can be reduced while a comparable performance is still achievable. For the case of downlink transmission with 1-bit DACs, there have been an increasing number of studies due to the benefits mentioned above [10]-[14]. In [10], a simple quantized ZF scheme is considered, where the transmit signal vector is obtained by a direct quantization on the ZF-precoded signals. The authors further analyze the performance of the quantized ZF scheme, and show that it outperforms the maximum likelihood (ML) encoder in the low-to-medium SNR regime. In [11], [12], the quantized linear beamforming schemes based on minimum-mean squared error (MMSE) are proposed, whose performance

A. Li and C. Masouros are with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, London, WC1E 7JE, UK (e-mail: c.masouros@ucl.ac.uk, ang.li.14@ucl.ac.uk).

F. Liu is with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, London, WC1E 7JE, UK, and also with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: liufan92@bit.edu.cn).

A. L. Swindlehurst is with the Department of Electrical Engineering and Computer Science, Henry Samueli School of Engineering, University of California, Irvine, CA 92697 USA, and also with the Institute for Advanced Study, Technical University of Munich, 80333 Munich, Germany (e-mail: swindle@uci.edu).

This work was supported by the Royal Academy of Engineering, UK, the Engineering and Physical Sciences Research Council (EPSRC) project EP/M014150/1 and the China Scholarship Council (CSC).

is shown to be superior to the quantized ZF scheme in [10]. In [13], a non-linear symbol perturbation technique is introduced in 1-bit massive MIMO downlink for QPSK modulation, while in [14] an iterative non-linear beamforming scheme is introduced via a biconvex relaxation approach, where the proposed scheme directly designs the transmit signal vector based on the MMSE criterion. Nevertheless, while operating on a symbol-by-symbol basis, these MMSE-based schemes may be sub-optimal as they ignore the fact that interference can be exploited on an instantaneous basis in [15]-[20]. Moreover, while there have been studies on the downlink beamforming schemes with 1-bit DACs, most of the these existing schemes either suffer a severe performance degradation in [10]-[12] compared to the unquantized case, or require sophisticated optimizations and iterative algorithms that are computationally inefficient [14].

In this paper, we revisit the symbol-level operations required for massive MIMO downlink transmission with 1-bit DACs to exploit the formulation of constructive interference. The symbol-by-symbol precoding operation allows us to observe the interference from an instantaneous point of view, and exploit it constructively [15]-[20]. We firstly consider a quantized linear beamforming scheme by constructing a beamforming matrix before quantization. Based on the concept of constructive interference, the optimization aims to maximize the distance between the received symbols and the detection thresholds. By mathematically analyzing the optimization problem with the Lagrangian approach, it is shown that the optimality is achieved by applying a strict phase rotation for the constructed problem in the case of massive MIMO. Due to the operation of the 1-bit quantization, the above quantized linear scheme is analytically shown to be equivalent to the quantized ZF scheme, which suffers an error floor at high SNR. To improve the performance, we then propose a non-linear mapping scheme where we directly design the quantized transmit signal vector. Nevertheless, due to the constraint on the output signals of 1-bit DACs, the resulting optimization problem is shown to be non-convex. To solve this problem, we firstly apply a relaxation on the mathematical constraint resulting from the use of 1-bit DACs, such that the optimization problem becomes convex. Then, we apply an element-wise normalization on the signal vector obtained from the relaxed optimization to meet the constraint on the output signals of 1-bit DACs.

Nevertheless, since the variable of the non-linear optimization approach is the transmit signal vector, whose dimension is equal to the number of transmit antennas, the computational complexity of the resulting optimization is high in the case of massive MIMO. Therefore, to enable the practical implementation of 1-bit DACs, we further propose a low-complexity symbol scaling scheme based on a coordinate transformation of the constructive interference problem, where we directly select the 1-bit DAC output for each antenna element on a sequential basis, and a relaxation-normalization process is therefore no longer needed. The proposed symbol scaling approach consists of three stages: an initialization stage where we decide the output signals for some antenna elements whose channel coefficients satisfy certain requirements, an allocation stage where we sequentially select the output signals for the residual antenna elements, and a refinement stage where we check whether the performance with the obtained signal vector can be further improved based on the greedy algorithm. Both the 'Sum-Max' and the 'Max-Min' criteria are considered in the allocation stage, and the output signal vector that returns the best performance is then obtained within the above two criteria. We further study the computational costs of the proposed optimization-based and symbol scaling schemes in terms of the floating operations required. Numerical results show that in the case of small-scale MIMO systems, the proposed symbol scaling scheme is shown to achieve the best performance. In the case of massive MIMO, the optimization-based non-linear scheme achieves an improved performance over existing schemes and better approaches the unquantized scheme, while the proposed symbol scaling scheme can achieve a comparable performance. In terms of the computational complexity, it is demonstrated that the complexity of the symbol scaling scheme is negligible compared to that of the non-linear mapping approach, while the performance of the symbol scaling scheme is superior to 'Pokemon' when their computational costs are similar, which favours its usefulness in practice.

For reasons of clarity, we summarize the contributions of this paper as:

1) We propose downlink beamforming schemes for massive MIMO with 1-bit DACs based on the constructive interference formulation. We firstly consider a quantized linear beamforming scheme, where it is analytically proven that, in the massive MIMO region, the optimality is achieved by employing a strict phase rotation due to the favourable propagation conditions.
2) We then consider a non-linear mapping scheme where we directly optimize the transmit signal vector. The resulting non-convex optimization is solved in two steps: we firstly relax the non-convex constraints of 1-bit DACs, followed by the normalization on the obtained signal vector to satisfy the 1-bit DAC transmission.
3) Based on a coordinate transformation of the constructive interference formulation, we further propose a low-complexity symbol scaling scheme where we directly select the quantized signal on each antenna element via a three-stage process. It is shown that the symbol scaling scheme can achieve a comparable performance to the optimization-based non-linear mapping scheme.
4) We further study and compare the computational costs of the optimization-based non-linear mapping scheme and the symbol scaling schemes in terms of the floating operations required, where it is shown mathematically and numerically that compared to the non-linear mapping approach, the complexity of the proposed symbol scaling approach is negligible.

The remainder of this paper is organized as follows. Section II introduces the system model. Both the proposed optimization-based quantized linear beamforming scheme and the non-linear mapping scheme that exploit the constructive interference are presented in Section III. The low-complexity three-stage symbol scaling method is presented in Section

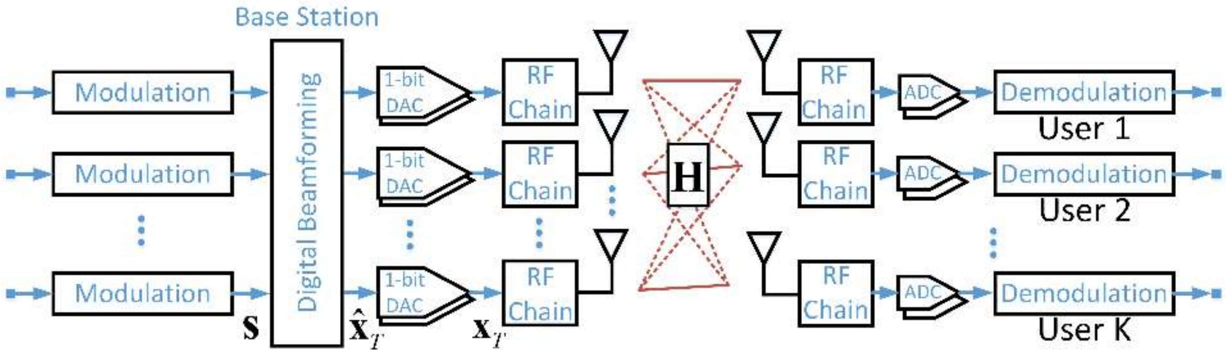

Fig. 1: Massive MIMO downlink system model with 1-bit DACs

IV. Section V includes the analysis of the computational complexity for both schemes, and the numerical results are shown in Section VI. Section VII concludes the paper.

Notations: a, $\mathbf{a}$, and $\mathbf{A}$ denote scalar, vector and matrix, respectively. $(\cdot)^T$ and $(\cdot)^H$ denote transposition and conjugate transposition of a matrix, respectively. $card(\cdot)$ denotes the cardinality of a set. j denotes the imaginary unit, and $\mathrm{vec}(\cdot)$ denotes the vectorization operation. $\mathbf{a}(k)$ denotes the k-th entry in vector $\mathbf{a}$. $|\cdot|$ denotes the modulus of a complex number or the absolute value of a real number, $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|_1$ denotes the 1-norm. $\mathcal{C}^{n\times n}$ represents an $n\times n$ matrix in the complex set, and $\mathbf{I}$ denotes the identity matrix. $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary part of a complex number, respectively.

II. System Model

We consider a multi-user massive MIMO downlink, where 1-bit DACs are employed at the BS, as depicted in Fig. 1. As we focus on the transmit-side processing, ideal ADCs with infinite precision are assumed to be employed at each receiver. The BS with N_t transmit antennas is communicating with K single-antenna users simultaneously in the same time-frequency resource, where $K \ll N_t$. We focus on the transmit beamforming designs and perfect CSI is assumed, while we also numerically study the performance of the proposed schemes with imperfect CSI in Section VI. Following the closely-related literature [10]-[13], [21], the symbol vector is assumed to be from a normalized PSK constellation. We denote the data symbol vector as $\mathbf{s}\in\mathcal{C}^{K\times 1}$, and the unquantized signal vector that is formed based on $\mathbf{s}$ as $\hat{\mathbf{x}}_T\in\mathcal{C}^{N_t\times 1}$. Then, the unquantized signal vector $\hat{\mathbf{x}}_T$ can be expressed as

$$\hat{\mathbf{x}}_T = \mathcal{B}(\mathbf{s}), \tag{1}$$

where $\mathcal{B}$ denotes a general linear or non-linear transformation. With 1-bit DACs employed, the output signal vector is then obtained as

$$\mathbf{x}_T = \mathcal{Q}(\hat{\mathbf{x}}_T). \tag{2}$$

In (2), $\mathcal{Q}$ denotes the 1-bit quantization on both the real and imaginary part of each entry in $\hat{\mathbf{x}}_T$. We denote x_n, $n\in\{1,2,\cdots,N_t\}$ as the n-th entry in $\mathbf{x}_T$, and in this paper each x_n is normalized to satisfy

$$x_n \in \left\{\pm\frac{1}{\sqrt{2N_t}}\pm\frac{1}{\sqrt{2N_t}}\cdot j\right\},\ \forall n\in\mathcal{N}, \tag{3}$$

where $\mathcal{N}=\{1,2,...,N_t\}$. The above normalization guarantees that $\|\mathbf{x}_T\|_F^2=1$, and we can then express the received signal vector as

$$\mathbf{y}=\sqrt{P}\cdot\mathbf{H}\mathbf{x}_T+\mathbf{n}, \tag{4}$$

where $\mathbf{H}\in\mathcal{C}^{K\times N_t}$ denotes the flat-fading Rayleigh channel with each entry following a standard complex Gaussian distribution. $\mathbf{n}\in\mathcal{C}^{K\times 1}$ denotes the additive Gaussian distributed noise vector with zero mean and covariance $\sigma^2\cdot\mathbf{I}$. P is the total available transmit power per antenna, and for simplicity in this paper we assume uniform power allocation for the antenna array.

III. 1-Bit Transmission Scheme based on Constructive Interference

A. Constructive Interference and Constructive Region

Constructive interference is defined as interference that pushes the received signals away from the detection thresholds of the modulation constellation [15]-[17]. The exploitation of constructive interference was firstly introduced in [15] to improve the performance of the ZF beamforming scheme, and was more recently applied to optimization-based approaches in [16], [17] and [20] based on the constructive region. To illustrate the underlying concept intuitively, in Fig. 2 we depict the constructive region for QPSK, where for simplicity and without loss of generality we focus on one quarter of the normalized QPSK constellation. As can be observed, as long as the interfered signal ($\overrightarrow{OB}$ in Fig. 2) is located in the constructive region, the distance to the detection thresholds is increased, and an improved performance can be expected. The formulation of the optimization problem based on the constructive region will be introduced in the following.

B. 1-Bit Transmission Scheme - Linear Beamforming

When a linear beamforming scheme is considered, the unquantized transmit signal vector can be expressed as

$$\hat{\mathbf{x}}_T = \mathbf{W}\mathbf{s}. \tag{5}$$

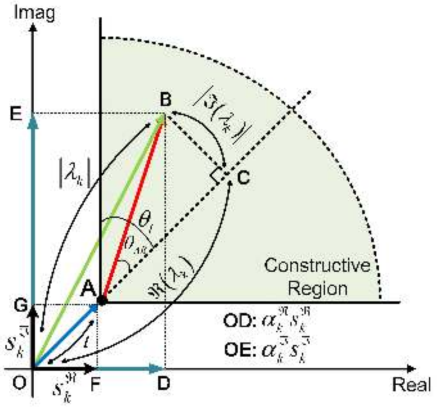

Fig. 2: Constructive interference and constructive region for QPSK

To introduce the proposed scheme, we firstly decompose the channel matrix into

$$\mathbf{H} = \left[\mathbf{h}_1^T, \mathbf{h}_2^T, \cdots, \mathbf{h}_K^T\right]^T, \tag{6}$$

where each $\mathbf{h}_k \in \mathcal{C}^{1\times N_t}$ denotes the channel vector of the k-th user. Then, the received signal for user k can be obtained as

$$\begin{aligned} y_k &= \sqrt{P}\cdot \mathbf{h}_k \mathbf{x}_T + n_k \\ &= \sqrt{P}\cdot \mathbf{h}_k \mathcal{Q}\left(\mathbf{W}\mathbf{s}\right) + n_k, \end{aligned} \tag{7}$$

where n_k is the k-th entry in $\mathbf{n}$. For the proposed quantized linear approach in this paper, the unquantized beamforming matrix $\mathbf{W}$ assuming infinite-precision DACs is firstly obtained, followed by the 1-bit quantization on the resulting transmit signal vector $\hat{\mathbf{x}}_T$.

To formulate the desired optimization problem, let us firstly study the analytical constructive interference conditions. In Fig. 2, without loss of generality we denote $\overrightarrow{OA} = t\cdot s_k$ and $t = |\overrightarrow{OA}|$ is the objective to be maximized. We assume the node 'B' denotes the noiseless received signal ($\mathbf{h}_k\mathbf{W}\mathbf{s}$) that is located in the constructive region, and we further denote $\overrightarrow{OB} = \lambda_k s_k$, where λ_k is an introduced complex variable with $|\overrightarrow{OB}| = |\lambda_k|$. We can then obtain that

$$\overrightarrow{OB} = \mathbf{h}_k\mathbf{W}\mathbf{s} = \lambda_k s_k. \tag{8}$$

Based on the fact that $\overrightarrow{OC}$ and $\overrightarrow{CB}$ are perpendicular, we can further obtain $\overrightarrow{OC}$ and $\overrightarrow{CB}$, expressed as

$$\overrightarrow{OC} = \Re\left(\lambda_k\right) s_k,\ \ \overrightarrow{CB} = j\cdot\Im\left(\lambda_k\right) s_k, \tag{9}$$

where geometrically the imaginary unit 'j' denotes a phase rotation of 90° along the anti-clockwise direction. As the nodes 'O', 'A', and 'C' are co-linear, we can then express $\overrightarrow{AC}$ as

$$\overrightarrow{AC} = \left[\Re\left(\lambda_k\right) - t\right] s_k. \tag{10}$$

Based on the expression of $\overrightarrow{AC}$ and $\overrightarrow{CB}$, $\tan\theta_{AB}$ is obtained as

$$\tan\theta_{AB} = \frac{|\overrightarrow{CB}|}{|\overrightarrow{AC}|} = \frac{|\Im\left(\lambda_k\right) s_k|}{|\left[\Re\left(\lambda_k\right) - t\right] s_k|} = \frac{|\Im\left(\lambda_k\right)|}{\Re\left(\lambda_k\right) - t}. \tag{11}$$

In Fig. 2, it is geometrically observed that to have node 'B' located in the constructive region is equivalent to the following condition:

$$\begin{aligned} &\theta_{AB} \le \theta_t \\ &\Rightarrow \tan\theta_{AB} \le \tan\theta_t \\ &\Rightarrow \frac{|\Im\left(\lambda_k\right)|}{\Re\left(\lambda_k\right) - t} \le \tan\theta_t \\ &\Rightarrow \left[\Re\left(\lambda_k\right) - t\right]\tan\theta_t \ge |\Im\left(\lambda_k\right)|. \end{aligned} \tag{12}$$

For $\mathcal{M}$-PSK modulation, based on the geometry of the modulation constellation it is easy to obtain the threshold angle θ_t, given by

$$\theta_t = \frac{\pi}{\mathcal{M}}. \tag{13}$$

We can then formulate the optimization for the unquantized linear beamforming as

$$\begin{aligned} \mathcal{P}_1:\ &\max_{\mathbf{W}}\ t \\ s.t.\ \ &\mathbf{h}_k\mathbf{W}\mathbf{s} = \lambda_k s_k,\ \forall k\in\mathcal{K} \\ &\left[\Re\left(\lambda_k\right) - t\right]\tan\theta_t \ge |\Im\left(\lambda_k\right)|,\ \forall k\in\mathcal{K} \\ &\|\mathbf{W}\mathbf{s}\|_F \le \sqrt{p_0} \\ &t \ge 0 \end{aligned} \tag{14}$$

where $\mathcal{K} = \{1, 2, \cdots, K\}$, and $\|\mathbf{W}\mathbf{s}\|_F \le \sqrt{p_0}$ is the instantaneous power constraint on the beamformer as the beamforming is dependent on the data symbols. Due to the existence of the subsequent 1-bit quantization operation, p_0 in $\mathcal{P}_1$ can be any positive value, and this will not have an impact on the final obtained quantized signal vector $\mathbf{x}_T$. $\mathcal{P}_1$ is a second-order cone programming (SOCP) optimization, and we can further obtain the following proposition in the case of massive MIMO.

Proposition: In the case of massive MIMO, the optimality conditions for each λ_k and t of the optimization problem $\mathcal{P}_1$ are obtained as

1) $\Im\left(\lambda_k^*\right) = 0$, $\forall k\in\mathcal{K}$;
2) $t^* = \lambda_1^* = \lambda_2^* = \cdots = \lambda_K^* = \sqrt{\frac{N_t\cdot p_0}{K}}$.

Proof: We prove the above proposition by analyzing the optimization problem $\mathcal{P}_1$ with the Lagrangian approach. We firstly transform $\mathcal{P}_1$ into a standard minimization problem, given by

$$\begin{aligned} \mathcal{P}_2:\ &\min_{\mathbf{w}_i}\ -t \\ s.t.\ \ &\mathbf{h}_k\sum_{i=1}^{K}\mathbf{w}_i s_i - \lambda_k s_k = 0,\ \forall k\in\mathcal{K} \\ &|\Im\left(\lambda_k\right)| - \left[\Re\left(\lambda_k\right) - t\right]\tan\theta_t \le 0,\ \forall k\in\mathcal{K} \\ &\sum_{i=1}^{K} s_i^H\mathbf{w}_i^H\mathbf{w}_i s_i - p_0 \le 0 \end{aligned} \tag{15}$$

where we note that the constraint on t in $\mathcal{P}_1$ can be omitted in the above formulation, and we decompose $\mathbf{W} =$

$[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K]$. We can then express the Lagrangian of $\mathcal{P}_2$ as [22]

$$\begin{aligned}\mathcal{L}\left(\mathbf{w}_i, t, \delta_k, \mu_k, \mu_0\right) = -t + \sum_{k=1}^{K} \delta_k \left(\mathbf{h}_k \sum_{i=1}^{K} \mathbf{w}_i s_i - \lambda_k s_k \right) \\ + \mu_0 \left(\sum_{i=1}^{K} s_i^H \mathbf{w}_i^H \mathbf{w}_i s_i - p_0 \right) \\ + \sum_{k=1}^{K} \mu_k \left[\left| \Im\left(\lambda_k\right) \right| - \Re\left(\lambda_k\right) \tan\theta_t + t \cdot \tan\theta_t \right],\end{aligned} \tag{16}$$

where μ_0, δ_k and μ_k are the dual variables, and $\mu_0 \geq 0$, $\mu_k \geq 0$, $\forall k \in \mathcal{K}$. Based on the Lagrangian in (16), the KKT conditions for optimality are then obtained as

$$\frac{\partial \mathcal{L}}{\partial t} = -1 + \sum_{k=1}^{K} \mu_k = 0 \tag{17a}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \left(\sum_{k=1}^{K} \delta_k \cdot \mathbf{h}_k \right) s_i + \mu_0 \cdot \mathbf{w}_i^H = \mathbf{0} \tag{17b}$$

$$\mu_0 \left(\sum_{i=1}^{K} s_i^H \mathbf{w}_i^H \mathbf{w}_i s_i - p_0 \right) = 0 \tag{17c}$$

$$\delta_k \left(\mathbf{h}_k \sum_{i=1}^{K} \mathbf{w}_i s_i - \lambda_k s_k \right) = 0, \ \forall k \in \mathcal{K} \tag{17d}$$

$$\mu_k \left[\left| \Im\left(\lambda_k\right) \right| - \Re\left(\lambda_k\right) \tan\theta_t + t \cdot \tan\theta_t \right] = 0, \ \forall k \in \mathcal{K} \tag{17e}$$

Based on (17b), firstly it is easily obtained that $\mu_0 \neq 0$ which with the fact that $\mu_0 \geq 0$ further leads to $\mu_0 > 0$. Then, we can obtain $\mathbf{w}_i^H$ as

$$\mathbf{w}_i^H = -\frac{1}{\mu_0} \cdot \left(\sum_{k=1}^{K} \delta_k \mathbf{h}_k \right) s_i, \ \forall i \in \mathcal{K}. \tag{18}$$

By denoting

$$a_k = -\frac{\delta_k^H}{\mu_0}, \ \forall k \in \mathcal{K}, \tag{19}$$

$\mathbf{w}_i$ can be obtained from (18) and expressed as

$$\mathbf{w}_i = \left(\sum_{k=1}^{K} a_k \mathbf{h}_k^H \right) s_i^H, \ \forall i \in \mathcal{K}. \tag{20}$$

Then, with the expression of each $\mathbf{w}_i$, the beamforming matrix $\mathbf{W}$ is obtained in a compact form as

$$\begin{aligned}\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K] = \left(\sum_{k=1}^{K} a_k \mathbf{h}_k^H \right) \cdot \left[s_1^H, s_2^H, \cdots, s_K^H \right] \\ = \left[\mathbf{h}_1^H, \mathbf{h}_2^H, \cdots, \mathbf{h}_K^H \right] \left[a_1, a_2, \cdots, a_K \right]^T \left[s_1^H, s_2^H, \cdots, s_K^H \right] \\ = \mathbf{H}^H \mathbf{A} \mathbf{s}^H.\end{aligned} \tag{21}$$

In order to obtain $\mathbf{A}$, we firstly rewrite (8) in a compact form, which is expressed as

$$\mathbf{H}\mathbf{W}\mathbf{s} = diag\left(\lambda_k\right) \mathbf{s}. \tag{22}$$

Then, by substituting (21) into (22), the matrix $\mathbf{A}$ can be obtained based on λ_k, given by

$$\begin{aligned}\mathbf{H}\mathbf{H}^H \mathbf{A} \mathbf{s}^H \mathbf{s} = diag\left(\lambda_k\right) \mathbf{s} \\ \Rightarrow \mathbf{A} = \frac{1}{K} \cdot \left(\mathbf{H}\mathbf{H}^H \right)^{-1} diag\left(\lambda_k\right) \mathbf{s}.\end{aligned} \tag{23}$$

The beamforming matrix $\mathbf{W}$ is then obtained as

$$\mathbf{W} = \frac{1}{K} \cdot \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H \right)^{-1} diag\left(\lambda_k\right) \mathbf{s}\mathbf{s}^H. \tag{24}$$

Based on the fact that $\mu_0 \neq 0$, it is obtained from (17c) that the power constraint of the optimization problem $\mathcal{P}_1$ is strictly active, which further leads to

$$\begin{aligned}\left\| \mathbf{W}\mathbf{s} \right\|_F = \sqrt{p_0} \\ \Rightarrow \ tr\left\{ \mathbf{W}\mathbf{s}\mathbf{s}^H \mathbf{W}^H \right\} = p_0 \\ \Rightarrow \ \mathbf{s}^H \mathbf{W}^H \mathbf{W}\mathbf{s} = p_0.\end{aligned} \tag{25}$$

Then, by substituting (24) into (25), we obtain that

$$\begin{aligned}\mathbf{s}^H diag\left(\lambda_k^H\right) \left(\mathbf{H}\mathbf{H}^H \right)^{-1} diag\left(\lambda_k\right) \mathbf{s} = p_0 \\ \Rightarrow \mathrm{vec}^T\left(\lambda_k^H\right) diag\left(\mathbf{s}^H\right) \left(\mathbf{H}\mathbf{H}^H \right)^{-1} diag\left(\mathbf{s}\right) \mathrm{vec}\left(\lambda_k\right) = p_0 \\ \Rightarrow \left[\lambda_1^H, \lambda_2^H, ..., \lambda_K^H \right] \cdot \mathbf{T} \cdot \left[\lambda_1, \lambda_2, ..., \lambda_K \right]^T = p_0,\end{aligned} \tag{26}$$

where $\mathbf{T}$ is defined as

$$\mathbf{T} = diag\left(\mathbf{s}^H\right) \left(\mathbf{H}\mathbf{H}^H \right)^{-1} diag\left(\mathbf{s}\right). \tag{27}$$

In the case of massive MIMO, as $N_t \to \infty$, the favourable propagation property gives us that [1]

$$\mathbf{H}\mathbf{H}^H \approx N_t \cdot \mathbf{I} \Rightarrow \left(\mathbf{H}\mathbf{H}^H \right)^{-1} \approx \frac{1}{N_t} \cdot \mathbf{I}, \tag{28}$$

based on which $\mathbf{T}$ is further transformed into

$$\mathbf{T} \approx \frac{1}{N_t} \cdot diag\left(\mathbf{s}^H\right) diag\left(\mathbf{s}\right) = \frac{1}{N_t} \cdot \mathbf{I}. \tag{29}$$

From the result in (29), (26) can be expanded and further transformed into

$$\frac{1}{N_t} \cdot \left(\left|\lambda_1\right|^2 + \left|\lambda_2\right|^2 + \cdots + \left|\lambda_K\right|^2 \right) = p_0. \tag{30}$$

To maximize t, as per (12) and (30) it is then easily obtained that the optimality is achieved when each λ_k^* is real and identical, given by

$$t^* = \lambda_1^* = \cdots = \lambda_K^* = \sqrt{\frac{N_t \cdot p_0}{K}}, \tag{31}$$

which completes the proof. ■

By substituting (31) into (24), the optimal beamforming matrix $\mathbf{W}^*$ can be expressed as

$$\mathbf{W}^* = \sqrt{\frac{N_t \cdot p_0}{K^3}} \cdot \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H \right)^{-1} \mathbf{s}\mathbf{s}^H. \tag{32}$$

Then, with $\mathbf{W}^*$ obtained, the output signal vector that satisfies 1-bit DAC transmission is given as

$$\begin{aligned}\mathbf{x}_T = \mathcal{Q}\left(\mathbf{W}^* \mathbf{s}\right) \\ = \mathcal{Q}\left(\sqrt{\frac{N_t \cdot p_0}{K^3}} \cdot \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H \right)^{-1} \mathbf{s}\mathbf{s}^H \mathbf{s} \right) \\ = \mathcal{Q}\left(\sqrt{\frac{N_t \cdot p_0}{K}} \cdot \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H \right)^{-1} \mathbf{s} \right).\end{aligned} \tag{33}$$

The intuition from the above proposition and (33) is that the quantized linear scheme based on the constructive interference is equivalent to the conventional quantized ZF scheme in the case of massive MIMO with 1-bit quantization, which suffers an error floor at high SNR [10]. This then motivates the proposed non-linear mapping scheme that achieves an improved performance in the following.

C. 1-Bit Transmission Scheme - Non-linear Mapping

We proceed to introduce the optimization-based non-linear mapping scheme for massive MIMO with 1-bit DACs. This approach was first described in [21], and based on the constructive interference formulation in [19]. We employ this approach, to further design our low-complexity techniques in Section IV. The resulting optimization based on the constructive interference can be formulated as

$$\begin{aligned} \mathcal{P}_3: \; & \max_{\mathbf{x}_T} \; t \\ s.t. \; & \mathbf{h}_k \mathbf{x}_T = \lambda_k s_k, \; \forall k \in \mathcal{K} \\ & \left[\Re\left(\lambda_k\right) - t\right] \tan\theta_t \geq \left|\Im\left(\lambda_k\right)\right|, \; \forall k \in \mathcal{K} \\ & x_n \in \left\{ \pm \frac{1}{\sqrt{2N_t}} \pm \frac{1}{\sqrt{2N_t}} j \right\}, \; \forall n \in \mathcal{N} \\ & t \geq 0 \end{aligned} \tag{34}$$

It is observed that the optimization problem $\mathcal{P}_3$ is non-convex due to the output signal constraint for the 1-bit DACs in (34). To solve the above non-convex optimization, we adopt a two-step approach.

1) Relaxation: In the first step, we relax the strict modulus constraint on each x_n for both the real and imaginary part, and the resulting relaxed constraint can be expressed as

$$\left|\Re\left(x_n\right)\right| \leq \frac{1}{\sqrt{2N_t}}, \; \left|\Im\left(x_n\right)\right| \leq \frac{1}{\sqrt{2N_t}}, \; \forall n \in \mathcal{N}. \tag{35}$$

The optimization problem $\mathcal{P}_3$ is then reformulated into a relaxed version $\mathcal{P}_4$, given by

$$\begin{aligned} \mathcal{P}_4: \; & \max_{\hat{\mathbf{x}}_T} \; t \\ s.t. \; & \mathbf{h}_k \hat{\mathbf{x}}_T = \lambda_k s_k, \; \forall k \in \mathcal{K} \\ & \left[\Re\left(\lambda_k\right) - t\right] \tan\theta_t \geq \left|\Im\left(\lambda_k\right)\right|, \; \forall k \in \mathcal{K} \\ & \left|\Re\left(\hat{x}_n\right)\right| \leq \frac{1}{\sqrt{2N_t}}, \; \forall n \in \mathcal{N} \\ & \left|\Im\left(\hat{x}_n\right)\right| \leq \frac{1}{\sqrt{2N_t}}, \; \forall n \in \mathcal{N} \\ & t \geq 0 \end{aligned} \tag{36}$$

where we denote $\hat{x}_n$ as the n-th entry in the relaxed transmit signal vector $\hat{\mathbf{x}}_T$. The resulting $\mathcal{P}_4$ is convex and can be solved with convex optimization tools.

2) Normalization: The solution obtained from the relaxed optimization $\mathcal{P}_4$ cannot always guarantee the equality on both the real and imaginary part of $\hat{x}_n$. To force the constraint of 1-bit transmission, the elements of the 1-bit DAC output $\mathbf{x}_T$ are obtained as

$$x_n = \frac{\Re\left(\hat{x}_n\right)}{\sqrt{2N_t} \cdot \left|\Re\left(\hat{x}_n\right)\right|} + \frac{\Im\left(\hat{x}_n\right)}{\sqrt{2N_t} \cdot \left|\Im\left(\hat{x}_n\right)\right|} \cdot j, \; \forall n \in \mathcal{N}. \tag{37}$$

Antenna number N_t	16	32	48	64
Ratio η	20.52%	10.8%	7.28%	5.46%
Antenna number N_t	80	96	112	128
Ratio η	4.37%	3.65%	3.13%	2.73%

TABLE I: η with respect to the number of transmit antennas, $K = 4$, 500 channel realizations

We further note that, while we perform a relaxation on the 1-bit DAC constraint on each x_n in $\mathcal{P}_3$, it turns out that most entries of the obtained $\hat{\mathbf{x}}_T$ from the relaxed problem $\mathcal{P}_4$ already meet the strict-equality requirement for 1-bit quantization, i.e. only a few entries of $\hat{x}_n$ need to be normalized. To evaluate the deviation of the relaxed optimization $\mathcal{P}_4$ from the original problem $\mathcal{P}_3$, we define $n_{\Re}$ and $n_{\Im}$ as the number of entries in the obtained $\hat{\mathbf{x}}_T$ whose absolute values are smaller than $\frac{1}{\sqrt{2N_t}}$ for the real and imaginary part, respectively. We further introduce

$$\eta = \frac{n_{\Re} + n_{\Im}}{2N_t} \tag{38}$$

as the ratio of the number of entries that do not satisfy the 1-bit transmission to the total number of entries in $\hat{\mathbf{x}}_T$, and this ratio therefore represents the deviation of the solution obtained by the relaxed problem from the original problem. We have $0 \leq \eta \leq 1$, and $\mathcal{P}_4$ is equivalent to $\mathcal{P}_3$ if $\eta = 0$. It is also observed that a smaller value of η means that the relaxed optimization is closer to the original optimization.

To study this numerically, we present the value of η with respect to the number of antennas in Table I, where we have assumed a total number of $K = 4$ users in the downlink system, and the result is based on 500 channel realizations. It is observed that the ratio η decreases with the increase in the number of transmit antennas, which means that the solution obtained via the relaxed optimization problem $\mathcal{P}_4$ can be regarded as asymptotically optimal with an increasing number of transmit antennas in the case of massive MIMO.

IV. Proposed Low-Complexity Symbol Scaling Approach

While the above non-linear mapping scheme can be relaxed into a convex optimization problem, the corresponding computational complexity is still prohibitively high as the variable dimension is equal to the number of transmit antennas. We study this mathematically and numerically in Section V and VI, respectively. Therefore in this section, we propose a three-stage symbol scaling scheme, which requires much reduced complexity for a comparable performance. It will be shown in the numerical results that for the small-scale MIMO systems, the low-complexity scheme even outperforms the optimization-based non-linear mapping scheme in Section III, since no relaxation or normalization is required for this scheme.

A. A New Look at the Constructive Interference Criteria

To introduce the proposed symbol scaling scheme, we firstly perform a coordinate transformation on the formulation of the constructive interference constraint. To be specific, we firstly

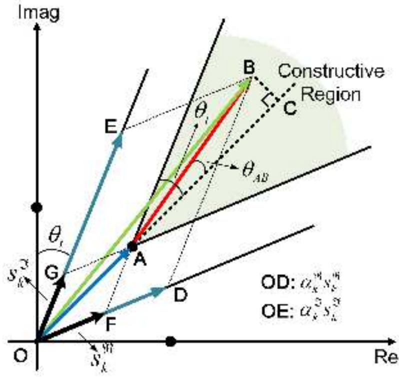

Fig. 3: Decomposition along the detection thresholds for 8-PSK

decompose each data symbol s_k along its two corresponding detection thresholds of the modulation constellation, given by

$$s_k = \overrightarrow{s_k^{\Re}} + \overrightarrow{s_k^{\Im}}, \tag{39}$$

where $\overrightarrow{s_k^{\Re}}$ and $\overrightarrow{s_k^{\Im}}$ are both complex values, and denoted as the two bases that are parallel to the two detection thresholds that correspond to the constellation point s_k. In the following, for simplicity we shall use $s_k^{\Re}$ and $s_k^{\Im}$ to denote the two bases. This is also shown geometrically in both Fig. 2 and Fig. 3 where we employ QPSK and 8-PSK modulation as examples, respectively. As observed in both figures, we decompose 'OA' that represents the data symbol s_k along its detection thresholds into 'OF' and 'OG'. For QPSK, based on Fig. 2 it is easy to observe that the real and imaginary axes are the detection thresholds, which leads to

$$\overrightarrow{OF} = s_k^{\Re} = \frac{1}{\sqrt{2}}, \ \overrightarrow{OG} = s_k^{\Im} = \frac{1}{\sqrt{2}} \cdot j \tag{40}$$

for the corresponding constellation point 'A'. For 8-PSK, 'OD' and 'OE' in Fig. 3 are the detection thresholds for the constellation point 'A'. Then, with $\theta_t = \pi/8$ for 8-PSK we can obtain the bases $s_k^{\Re}$ and $s_k^{\Im}$ that correspond to the constellation point 'A' as

$$\begin{aligned} \overrightarrow{OF} = s_k^{\Re} &= \frac{e^{j\cdot\frac{\pi}{8}}}{\left|e^{j\cdot\frac{\pi}{8}} + e^{j\cdot\frac{3\pi}{8}}\right|} = a_k + b_k \cdot j, \\ \overrightarrow{OG} = s_k^{\Im} &= \frac{e^{j\cdot\frac{3\pi}{8}}}{\left|e^{j\cdot\frac{\pi}{8}} + e^{j\cdot\frac{3\pi}{8}}\right|} = c_k + d_k \cdot j. \end{aligned} \tag{41}$$

where (a_k, b_k) and (c_k, d_k) denote the coordinates of $s_k^{\Re}$ and $s_k^{\Im}$ in the conventional real-imaginary complex plane, respectively. The extension to other constellation points and higher order PSK modulations can be easily obtained in a similar way.

Then for each k, instead of employing a complex scaling value λ_k that is multiplied by s_k, with the above formulation (39)-(41) we introduce a symbol scaling approach where we decompose (8) along the two corresponding detection thresholds of s_k, given by

$$\mathbf{h}_k \mathbf{x}_T = \alpha_k^{\Re} s_k^{\Re} + \alpha_k^{\Im} s_k^{\Im}, \tag{42}$$

where

$$\alpha_k^{\Re} \geq 0, \ \alpha_k^{\Im} \geq 0, \ \forall k \in \mathcal{K}, \tag{43}$$

are two introduced scaling factors that are multiplied to the bases $s_k^{\Re}$ and $s_k^{\Im}$, respectively. We can then observe that a larger value of $\alpha_k^{\Re}$ or $\alpha_k^{\Im}$ therefore represents a larger distance to the other detection threshold, and we further denote $\left(\alpha_k^{\Re}, \alpha_k^{\Im}\right)$ as the coordinate of the node 'B' in the complex plane expanded by the bases $s_k^{\Re}$ and $s_k^{\Im}$. By expanding (42) using the coordinate transformation, we can obtain the generic expression of $\alpha_k^{\Re}$ and $\alpha_k^{\Im}$ as a function of the transmit signal vector, given by (see Appendix)

$$\begin{aligned} \alpha_k^{\Re} &= \frac{d_k \Re(\mathbf{h}_k) - c_k \Im(\mathbf{h}_k)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Re} - \frac{d_k \Im(\mathbf{h}_k) + c_k \Re(\mathbf{h}_k)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Im}, \\ \alpha_k^{\Im} &= \frac{a_k \Im(\mathbf{h}_k) - b_k \Re(\mathbf{h}_k)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Re} + \frac{a_k \Re(\mathbf{h}_k) + b_k \Im(\mathbf{h}_k)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Im}. \end{aligned} \tag{44}$$

In (44), for simplicity we have employed the following denotations

$$\mathbf{x}_T^{\Re} = \Re(\mathbf{x}_T), \ \mathbf{x}_T^{\Im} = \Im(\mathbf{x}_T). \tag{45}$$

By further denoting

$$\begin{aligned} \mathbf{A}_k &= \frac{d_k \Re(\mathbf{h}_k) - c_k \Im(\mathbf{h}_k)}{a_k d_k - b_k c_k}, \ \mathbf{B}_k = -\frac{d_k \Im(\mathbf{h}_k) + c_k \Re(\mathbf{h}_k)}{a_k d_k - b_k c_k}, \\ \mathbf{C}_k &= \frac{a_k \Im(\mathbf{h}_k) - b_k \Re(\mathbf{h}_k)}{a_k d_k - b_k c_k}, \ \mathbf{D}_k = \frac{a_k \Re(\mathbf{h}_k) + b_k \Im(\mathbf{h}_k)}{a_k d_k - b_k c_k}, \end{aligned} \tag{46}$$

the formulation of (44) is simplified into

$$\begin{aligned} \alpha_k^{\Re} &= \mathbf{A}_k \mathbf{x}_T^{\Re} + \mathbf{B}_k \mathbf{x}_T^{\Im}, \\ \alpha_k^{\Im} &= \mathbf{C}_k \mathbf{x}_T^{\Re} + \mathbf{D}_k \mathbf{x}_T^{\Im}. \end{aligned} \tag{47}$$

By defining

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{A}_k & \mathbf{B}_k \end{bmatrix}, \ \mathbf{I}_k = \begin{bmatrix} \mathbf{C}_k & \mathbf{D}_k \end{bmatrix}, \tag{48}$$

and

$$\mathbf{x} = \begin{bmatrix} \left(\mathbf{x}_T^{\Re}\right)^T & \left(\mathbf{x}_T^{\Im}\right)^T \end{bmatrix}^T, \ \mathbf{\Lambda} = \left[\alpha_1^{\Re}, ..., \alpha_K^{\Re}, \alpha_1^{\Im}, ..., \alpha_K^{\Im}\right]^T, \tag{49}$$

(47) can be further expressed in a compact form as

$$\mathbf{\Lambda} = \mathbf{M}\mathbf{x}, \tag{50}$$

where $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{bmatrix} \mathbf{R}_1^T & \cdots & \mathbf{R}_K^T & \mathbf{I}_1^T & \cdots & \mathbf{I}_K^T \end{bmatrix}^T. \tag{51}$$

With the above formulation, we can then construct the optimization problem as

$$\begin{aligned} \mathcal{P}_5: \ & \max_{\mathbf{x}} \min_{l} \ \alpha_l \\ s.t. \ & \mathbf{\Lambda} = \mathbf{M}\mathbf{x} \\ & \alpha_l \geq 0, \ \forall l \in \mathcal{L} \\ & x_i^E \in \left\{ \frac{1}{\sqrt{2N_t}}, -\frac{1}{\sqrt{2N_t}} \right\}, \ \forall i \in \mathcal{I} \end{aligned} \tag{52}$$

where we have omitted $\Re$ and $\Im$ in the expression of the entries of $\mathbf{\Lambda}$, and simply denote α_l as its l-th entry. In $\mathcal{P}_5$, $\mathcal{L} = \{1, 2, \cdots, 2K\}$, x_i^E denotes the i-th entry in $\mathbf{x}$ and $\mathcal{I} = \{1, 2, \cdots, 2N_t\}$. The above optimization problem $\mathcal{P}_5$ is interpreted as follows: we aim to maximize the minimum value of α_l by selecting each x_i^E as either $\frac{1}{\sqrt{2N_t}}$ or $-\frac{1}{\sqrt{2N_t}}$. With the above problem formulation, the relaxation-normalization process on the transmit signals is no longer needed. The above formulation motivates us to propose the following low-complexity scheme, which consists of three stages: an initialization stage, an allocation stage, and a refinement stage, all presented in the following in detail.

B. Initialization Stage

In the initialization stage, we directly select the value of x_i^E for some i by simple observation. To achieve this, we firstly decompose (50) into

$$\mathbf{\Lambda} = \sum_{i=1}^{2N_t} \mathbf{M}_i x_i^E, \tag{53}$$

where we decompose $\mathbf{M}$ into

$$\mathbf{M} = \left[\begin{array}{cccc} \mathbf{M}_1 & \mathbf{M}_2 & \cdots & \mathbf{M}_{2N_t} \end{array} \right], \tag{54}$$

with each $\mathbf{M}_i \in \mathcal{C}^{2K \times 1}$. Then, we have the following observation.

$\mathbf{Observation}$: As long as all the entries of $\mathbf{M}_i$ share the same sign, then it is optimal to set the sign of the corresponding x_i^E equal to that of $\mathbf{M}_i$, as in this case the values of each entry in $\mathbf{\Lambda}$ are guaranteed to increase.

Then, the corresponding x_i^E is obtained as

$$x_i^E = \frac{\mathrm{sgn}\left(\mathbf{M}_i\right)}{\sqrt{2N_t}}, \; \forall i \in \mathcal{S}, \tag{55}$$

where $\mathrm{sgn}\left(\mathbf{a}\right)$ defines a vector sign function and is only valid when each entry in the vector $\mathbf{a}$ has the same sign. $\mathcal{S}$ denotes the set that consists of the column indices of $\mathbf{M}$ that satisfy the sign-identity condition. We further introduce a column vector $\mathbf{t}$ that represents a temporary value of $\mathbf{\Lambda}$, given by

$$\mathbf{t} = \sum_{i \in \mathcal{V}} \mathbf{M}_i x_i^E, \tag{56}$$

where the set $\mathcal{V}$ consists of the column indices of $\mathbf{M}$ whose corresponding x_i^E have been allocated a value. We note that when $card\left(\mathcal{V}\right) = 2N_t$, we have $\mathbf{t} = \mathbf{\Lambda}$.

In the case that no column in $\mathbf{M}$ satisfies the sign-identity condition, in the initialization stage we select only one column, i.e. $card\left(\mathcal{S}\right) = 1$, with the following criterion:

$$i = \arg\max_{i \in \mathcal{I}} \left\| \mathbf{M}_i \right\|_1, \tag{57}$$

which selects the column that has the maximum effect on the value of $\mathbf{\Lambda}$. Then, the value of the corresponding x_i^E is set as

$$x_i^E = \frac{\mathrm{sgn}\left(\left\| \mathbf{M}_i \right\|_1\right)}{\sqrt{2N_t}}. \tag{58}$$

In the initialization stage, we have $\mathcal{V} = \mathcal{S}$ or $card\left(\mathcal{V}\right) = 1$. We summarize the algorithm for the initialization stage in Algorithm 1.

Algorithm 1 Initialization Stage

```
input : s, H
output : t, $\mathcal{V}$
Decompose each $s_k = s_k^{\Re} + s_k^{\Im}$ based on modulation type;
Obtain $\mathbf{M}$ based on (42)-(51);
Find $\mathbf{M}_i$ that satisfies the sign-identity condition;
Obtain $\mathcal{S}$;
if $\mathcal{S} \neq \emptyset$ then
    $x_i^E = \frac{\mathrm{sgn}(\mathbf{M}_i)}{\sqrt{2N_t}}$, $\forall i \in \mathcal{S}$;
    $\mathcal{V} = \mathcal{S}$;
else
    Obtain $i$ based on (57), $x_i^E = \frac{\mathrm{sgn}(\|\mathbf{M}_i\|_1)}{\sqrt{2N_t}}$;
    $\mathcal{V} = \{i\}$;
end if
Calculate t based on (56).
```

C. Allocation Stage

At this stage we allocate the value of each x_i^E for the residual i that belongs to $\mathcal{W}$, where we define the set $\mathcal{W}$ as

$$\mathcal{W} = \left\{ i \,|\, i \in \mathcal{I} \,\mathrm{and}\, i \notin \mathcal{V} \right\}. \tag{59}$$

$\mathcal{W}$ consists of those x_i^E whose values have not been allocated in the initialization stage. In the following allocation stage, we consider both a ‘Sum-Max’ and a ‘Max-Min’ criteria for the allocation scheme.

1) Sum-Max: For the allocation scheme based on the ‘Sum-Max’ criterion, instead of considering a max-min optimization as in $\mathcal{P}_5$, we consider a sum-max optimization where the objective function is constructed as

$$\mathcal{F}\left(\mathbf{x}\right) = \mathrm{sum}\left(\mathbf{\Lambda}\right), \tag{60}$$

where $\mathrm{sum}\left(\mathbf{a}\right)$ returns the sum of the entries in a column vector $\mathbf{a}$. Then, based on (50) the objective can be further transformed into

$$\mathcal{F}\left(\mathbf{x}\right) = \mathbf{m}\mathbf{x} = \sum_{i=1}^{2N_t} \mathbf{m}\left(i\right) x_i^E, \tag{61}$$

where $\mathbf{m} \in \mathcal{C}^{1 \times 2N_t}$ is the sum of the entries in each row of M. Each $\mathbf{m}\left(i\right)$ denotes the i-th entry in $\mathbf{m}$, given by

$$\mathbf{m}\left(i\right) = \sum_{l=1}^{2K} \mathbf{M}_i\left(l\right). \tag{62}$$

It is then easy to observe that $\mathcal{F}\left(\mathbf{x}\right)$ is maximized when the sign of each x_i^E is the same as that of $\mathbf{m}\left(i\right)$, and therefore the optimal x_i^E for the ‘Sum-Max’ criterion is given by

$$x_i^E = \frac{\mathrm{sgn}\left[\mathbf{m}\left(i\right)\right]}{\sqrt{2N_t}}, \; \forall i \in \mathcal{W}. \tag{63}$$

While the above solution guarantees that the sum of α_l is maximized, it does not specifically consider each value of α_l, which may lead to performance loss. Indeed, it is possible that the value of one α_l can be very small or even negative. This is the reason why the refinement in Section IV-D is further introduced. The algorithm for the allocation stage based on ‘Sum-Max’ is summarized in Algorithm 2.

Algorithm 2 Allocation Stage - ‘Sum-Max’

```
input : $\mathcal{V}$, $\mathbf{M}$
output : $\mathbf{x}_{\mathrm{sum-max}}$
Calculate $\mathcal{W}$ based on (59);
Calculate $\mathbf{m}$ and each $\mathbf{m}(i)$ based on (61), (62);
Allocate $x_i^E = \frac{\mathrm{sgn}[\mathbf{m}(i)]}{\sqrt{2N_t}}$, $\forall i \in \mathcal{W}$;
Obtain $\mathbf{x}$, denoted as $\mathbf{x}_{\mathrm{sum-max}}$.
```

2) Max-Min: For the ‘Max-Min’ allocation criterion, in each step we aim to improve the minimum value in $\mathbf{\Lambda}$ as much as possible. Denoting q as the row index of the minimum entry in $\mathbf{t}$ obtained in the initialization stage, we have

$$\mathbf{t}(q) = \min(\mathbf{t}), \tag{64}$$

where $\min(\mathbf{t})$ returns the minimum value in $\mathbf{t}$. Subsequently, we iteratively select $\mathbf{M}_i$ with the largest absolute value in the q-th row, given by

$$i = \underset{i \in \mathcal{W}}{\arg\max} \left|\mathbf{M}_i(q)\right|, \tag{65}$$

and the corresponding x_i^E is then obtained as

$$x_i^E = \frac{\mathrm{sgn}\left[\mathbf{M}_i(q)\right]}{\sqrt{2N_t}}. \tag{66}$$

Then, we update $\mathcal{V}$ and $\mathbf{t}$, and based on the updated $\mathbf{t}$ we repeat the above procedure until $\mathcal{V} = \mathcal{I}$. This means that each entry in $\mathbf{x}$ has been allocated, and the algorithm for the allocation stage based on ‘Max-Min’ is summarized in Algorithm 3.

Algorithm 3 Allocation Stage - ‘Max-Min’

```
input : $\mathcal{V}$, $\mathbf{M}$, $\mathbf{t}$
output : $\mathbf{x}_{\mathrm{max-min}}$
while $\mathcal{V} \neq \mathcal{I}$ do
    Calculate $\mathcal{W}$ based on (59);
    Obtain $q$ that satisfies $\mathbf{t}(q) = \min(\mathbf{t})$;
    Find $i = \underset{i \in \mathcal{W}}{\arg\max} \left|\mathbf{M}_i(q)\right|$;
    Allocate $x_i^E = \frac{\mathrm{sgn}[\mathbf{M}_i(q)]}{\sqrt{2N_t}}$;
    Update $\mathcal{V}$ and $\mathbf{t}$;
end while
Obtain $\mathbf{x}$, denoted as $\mathbf{x}_{\mathrm{max-min}}$.
```

D. Refinement Stage

In the refinement stage, we check whether the performance based on the obtained signal vector in the allocation stage can be further improved based on a greedy algorithm. To introduce the refinement process, we denote the obtained expanded 1-bit signal vector after the allocation stage as $\mathbf{x}$ (obtained based on either the ‘Sum-Max’ or the ‘Max-Min’ criterion). First, we sequentially change the sign of one entry (for example x_i^E) in $\mathbf{x}$ at a time while fixing the signs of other entries in $\mathbf{x}$, and denote the modified signal vector as $\mathbf{x}_{(i)}$. We then compare the minimum value in $\mathbf{\Lambda}$ obtained by the modified $\mathbf{x}_{(i)}$ with the minimum value in the original $\mathbf{\Lambda}$ obtained by $\mathbf{x}_{(0)}$. The sign of x_i^E is selected as the one that returns a larger minimum value in $\mathbf{\Lambda}$. The refinement process is sequentially performed for each entry in $\mathbf{x}_{(0)}$. The algorithm for the refinement stage is then shown in Algorithm 4.

Algorithm 4 Refinement Stage

```
input : $\mathbf{x}_{\mathrm{sum-max}}$ (or $\mathbf{x}_{\mathrm{max-min}}$)
output : $\mathbf{x}_T$
Denote $\mathbf{x}_{(0)} = \mathbf{x}_{\mathrm{sum-max}}$ (or $\mathbf{x}_{\mathrm{max-min}}$);
for $i = 1 : 2N_t$ do
    Calculate $\mathbf{\Lambda}_{(0)} = \mathbf{M}\mathbf{x}_{(0)}$;
    Obtain $\mathbf{x}_{(i)} = \left[x_1^E, ..., x_{i-1}^E, -x_i^E, x_{i+1}^E, ..., x_{2N_t}^E\right]^T$;
    Calculate $\mathbf{\Lambda}_{(i)} = \mathbf{M}\mathbf{x}_{(i)}$;
    if $\min\left(\mathbf{\Lambda}_{(i)}\right) > \min\left(\mathbf{\Lambda}_{(0)}\right)$ then
        $x_i^E \leftarrow -x_i^E$;
        Update $\mathbf{x}_{(0)}$;
    end if
end for
Obtain $\mathbf{x}_T$ based on the updated $\mathbf{x}_{(0)}$.
```

The refinement stage is performed for the signal vectors obtained by both the ‘Sum-Max’ and ‘Max-Min’ criteria independently. The final output signal vector of the proposed symbol scaling scheme that generates the best performance is then selected between the signal vectors obtained with these two criteria.

E. Algorithm

Based on the above description, the algorithm for the three-stage symbol scaling scheme is summarized in Algorithm 5, where the final output signal vector of the proposed symbol scaling scheme that generates the best performance is selected within the signal vectors obtained by the ‘Sum-Max’ and ‘Max-Min’ criteria.

Algorithm 5 The Proposed Symbol Scaling Scheme

```
input : $\mathbf{s}$, $\mathbf{H}$
output : $\mathbf{x}_T$
Initialization Stage
Obtain $\mathcal{V}$, $\mathbf{M}$, and $\mathbf{t}$ with Algorithm 1;
Allocation Stage
$\mathbf{1}$.‘$\mathbf{Sum-Max}'$ :
Obtain $\mathbf{x}_{\mathrm{sum-max}}$ with Algorithm 2;
$\mathbf{2}$.‘$\mathbf{Max-Min}'$ :
Obtain $\mathbf{x}_{\mathrm{max-min}}$ with Algorithm 3;
Refinement Stage
Update both $\mathbf{x}_{\mathrm{sum-max}}$ and $\mathbf{x}_{\mathrm{max-min}}$ with Algorithm 4;
Calculate $\mathbf{\Lambda}_{\mathrm{s}} = \mathbf{M}\mathbf{x}_{\mathrm{sum-max}}$ and $\mathbf{\Lambda}_{\mathrm{m}} = \mathbf{M}\mathbf{x}_{\mathrm{max-min}}$;
if $\min(\mathbf{\Lambda}_{\mathrm{s}}) > \min(\mathbf{\Lambda}_{\mathrm{m}})$ then
    $\mathbf{x} = \mathbf{x}_{\mathrm{sum-max}}$;
else
    $\mathbf{x} = \mathbf{x}_{\mathrm{max-min}}$;
end if
Decompose $\mathbf{x} = \left[\begin{array}{cc} \left(\mathbf{x}_T^{\Re}\right)^T & \left(\mathbf{x}_T^{\Im}\right)^T \end{array}\right]^T$;
Output $\mathbf{x}_T = \mathbf{x}_T^{\Re} + \mathbf{x}_T^{\Im} \cdot j$.
```

Antenna Number	Schemes			
	Exhaustive Search	Proposed Non-linear Mapping $\mathcal{P}_4$	Proposed Symbol Scaling	Non-linear Pokemon, $n_{\max} = 20$
64	$\mathcal{O}\left\{1.39 \times 10^{42}\right\}$	$\mathcal{O}\left\{2.83 \times 10^{7}\right\}$	$\mathcal{O}\left\{7.9 \times 10^{4}\right\}$	$\mathcal{O}\left\{6.6 \times 10^{5}\right\}$
96	$\mathcal{O}\left\{3.86 \times 10^{61}\right\}$	$\mathcal{O}\left\{1.11 \times 10^{8}\right\}$	$\mathcal{O}\left\{1.74 \times 10^{5}\right\}$	$\mathcal{O}\left\{1.48 \times 10^{6}\right\}$
128	$\mathcal{O}\left\{9.49 \times 10^{80}\right\}$	$\mathcal{O}\left\{2.94 \times 10^{8}\right\}$	$\mathcal{O}\left\{3.05 \times 10^{5}\right\}$	$\mathcal{O}\left\{2.63 \times 10^{6}\right\}$
256	$\mathcal{O}\left\{2.20 \times 10^{158}\right\}$	$\mathcal{O}\left\{3.18 \times 10^{9}\right\}$	$\mathcal{O}\left\{1.2 \times 10^{6}\right\}$	$\mathcal{O}\left\{1.05 \times 10^{7}\right\}$

TABLE II: Comparison of the computational costs of different schemes, $K = 8$

V. Computational Complexity Analysis

In this section we study the computational costs of the proposed schemes in terms of the floating-point operations required. As a reference, we also study the complexity of the exhaustive search scheme and the non-linear 'Pokemon' scheme in [14]. The computational costs of all considered approaches are calculated based on real multiplications and additions.

A. Exhaustive Search

For massive MIMO transmission with 1-bit quantization, the output signal on each antenna element has 4 potential values, and for each signal combination it takes $4KN_t$ multiplications and $4KN_t$ additions to compute $\mathbf{\Lambda}$ based on (50) as $\mathbf{M} \in \mathcal{C}^{2K\times 2N_t}$. Therefore, the complexity of the exhaustive search scheme is obtained as

$$\mathrm{C_E} = \mathcal{O}\left\{8KN_t \cdot 4^{N_t}\right\} = \mathcal{O}\left\{8KN_t \cdot 2^{2N_t}\right\}. \tag{67}$$

It is easy to conclude that in the case of massive MIMO, the exhaustive search scheme is inapplicable due to the overwhelmingly high computational cost.

B. Optimization-based Non-linear Mapping $\mathcal{P}_4$

For the proposed non-linear mapping scheme, in the relaxation stage the complexity is dominated from solving the relaxed convex problem $\mathcal{P}_5$ via the interior-point method [22]. It has been shown in [23] that the arithmetic complexity of the interior-point method is given by

$$\mathrm{C_I} = \mathcal{O}\left\{(M+N)^{1.5}M^2\right\}, \tag{68}$$

where M is the dimension of the variable, and N is the number of constraints. Based on the real representation $\mathcal{P}_5$, we obtain $M = 2N_t$ and $N = 2K$, which leads to

$$\begin{aligned}\mathrm{C_N^1} &= \mathcal{O}\left\{(2K+2N_t)^{1.5}(2N_t)^2\right\} \\ &= \mathcal{O}\left\{8\sqrt{2}(K+N_t)^{1.5}N_t^2\right\}.\end{aligned} \tag{69}$$

In the normalization stage, the dominant complexity comes from the search for the signals that do not satisfy the output constraint for the 1-bit transmission. There are a total number of $2N_t$ entries in $\hat{\mathbf{x}}_T$ including both the real and imaginary part, and therefore a one-dimensional search of $2N_t$ entries is required. Then, the resulting complexity is obtained as

$$\mathrm{C_N^2} = O\left\{2N_t\right\}, \tag{70}$$

which leads to the total computational cost for the optimization-based non-linear mapping scheme as

$$\mathrm{C_N} = \mathrm{C_N^1} + \mathrm{C_N^2} = \mathcal{O}\left\{8\sqrt{2}(K+N_t)^{1.5}N_t^2\right\} + \mathcal{O}\left\{2N_t\right\}. \tag{71}$$

In the case of massive MIMO where N_t is large, we have the following approximation:

$$\mathrm{C_N} \approx \mathcal{O}\left\{8\sqrt{2}(K+N_t)^{1.5}N_t^2\right\}. \tag{72}$$

C. Symbol Scaling Scheme

For the proposed symbol scaling approach, in the following we calculate its computational cost for each stage. For both allocation criteria, the main computational cost in the initialization and allocation stage comes from the calculation of $\mathbf{t} \in \mathcal{C}^{2N_t\times 1}$ based on (56). While the calculation of $\mathbf{t}$ is not necessary for the 'Sum-Max' criterion, we note that $\mathbf{t}$ is required in the refinement stage. Each additional $\left(\mathbf{M}_i x_i^E\right)$ term that is added to $\mathbf{t}$ requires $2N_t$ multiplications and $2N_t$ additions, and $\mathbf{t}$ is updated $2K$ times after the allocation stage, where we note $\mathbf{M} \in \mathcal{C}^{2K\times 2N_t}$. The resulting computation cost is

$$\mathrm{C_L^1} = \mathcal{O}\left\{2K\left(2N_t + 2N_t\right)\right\} = \mathcal{O}\left\{8KN_t\right\}. \tag{73}$$

Moreover, for the 'Max-Min' allocation criterion, we need to iteratively allocate the value for the residual x_i^E, which introduces an additional computational cost for 'Max-Min' in the allocation stage. Since $card\left(\mathcal{V}\right)$ is difficult to obtain analytically in the initialization stage, we consider a worst-case complexity where $card\left(\mathcal{V}\right) = 1$, and in each iteration obtaining q and i in Algorithm 3 requires $2K$ and $2N_t$ operations, respectively. The required number of computations is thus

$$\mathrm{C_L^2} = \mathcal{O}\left\{\left(2N_t - 1\right)\left(2K + 2N_t\right)\right\} \approx \mathcal{O}\left\{4N_t^2 + 4KN_t\right\} \tag{74}$$

in the case of massive MIMO. In the refinement stage, it is easy to observe that the initial $\mathbf{\Lambda}_{(0)} = \mathbf{t}$. Then, in each iteration of Algorithm 4 we only need to calculate the corresponding $\mathbf{M}_i \cdot \left(-x_i^E\right)$ and include it in $\mathbf{\Lambda}_{(i)}$. For each x_i^E this takes $2N_t$ multiplications and $2N_t$ additions, and therefore the computational cost for the refinement stage is

$$\mathrm{C_L^3} = \mathcal{O}\left\{2N_t\left(2N_t + 2N_t\right)\right\} = \mathcal{O}\left\{8N_t^2\right\}. \tag{75}$$

Based on Algorithm 5, both $\mathbf{x}_{\mathrm{sum-max}}$ and $\mathbf{x}_{\mathrm{max-min}}$ should be refined. Accordingly, we can obtain the total computational cost for the proposed symbol scaling approach as

$$\begin{aligned} \mathrm{C_L} &= \mathrm{C_L^1} + \mathrm{C_L^2} + 2\mathrm{C_L^3} \\ &= \mathcal{O}\left\{8KN_t\right\} + \mathcal{O}\left\{4N_t^2 + 4KN_t\right\} + \mathcal{O}\left\{2\left(8N_t^2\right)\right\} \\ &= \mathcal{O}\left\{20N_t^2 + 12KN_t\right\}. \end{aligned} \tag{76}$$

D. Pokemon

As a comparison, we also include the complexity of the non-linear 'Pokemon' scheme proposed in [14]. The 'Pokemon' approach is based on biconvex relaxation, whose performance is dependent on the number of required iterations. Based on [14], in each iteration we need to first calculate a vector $\mathbf{q} \in \mathcal{C}^{2N_t \times 1}$ based on $\mathbf{q} = \mathbf{U}\mathbf{x}$ where $\mathbf{U} \in \mathcal{C}^{2N_t \times 2N_t}$, and then update the signal vector $\mathbf{x} \in \mathcal{C}^{2N_t \times 1}$ with a projection function. The calculation of $\mathbf{q}$ requires a total of $4N_t^2$ multiplications and $4N_t^2$ additions, while the update of $\mathbf{x}$ requires $4N_t$ multiplications. Assuming a maximum number of iterations $n_{\max}$, this leads to the total computational cost for 'Pokemon' as

$$\begin{aligned} \mathrm{C_P} &= \mathcal{O}\left\{n_{\max}\left(4N_t^2 + 4N_t^2 + 4N_t\right)\right\} \\ &= \mathcal{O}\left\{n_{\max}\left(8N_t^2 + 4N_t\right)\right\}. \end{aligned} \tag{77}$$

Comparing the computational cost of 'Pokemon' with the proposed symbol scaling method, we have

$$\frac{\mathrm{C_L}}{\mathrm{C_P}} = \frac{\mathcal{O}\left\{20N_t^2 + 12KN_t\right\}}{\mathcal{O}\left\{n_{\max}\left(8N_t^2 + 4N_t\right)\right\}} = \mathcal{O}\left\{\frac{5N_t + 3K}{n_{\max}\left(2N_t + 1\right)}\right\}. \tag{78}$$

In the case of massive MIMO where K is finite while the antenna number $N_t \to \infty$, (78) is further transformed into

$$\frac{\mathrm{C_L}}{\mathrm{C_P}} = \mathcal{O}\left\{\frac{5 + \frac{3K}{N_t}}{n_{\max}\left(2 + \frac{1}{N_t}\right)}\right\} \approx \mathcal{O}\left\{\frac{2.5}{n_{\max}}\right\}. \tag{79}$$

To numerically study the complexity gains of the proposed symbol scaling method, in Table II we show the number of floating-point operations required as the number of transmit antennas increases, where for 'Pokemon' we employ $n_{\max} = 20$ following [14]. As can be seen, the computational cost of the proposed non-linear mapping scheme is higher than that of the proposed symbol scaling approach and the 'Pokemon' method, while the number of operations required for the proposed symbol scaling approach is approximately 12% of the number of operations for 'Pokemon'.

VI. Numerical Results

In this section we present the numerical results of the proposed approaches based on Monte Carlo simulations. In each plot, the transmit SNR is defined as $\rho = P/\sigma^2$. Both QPSK and 8-PSK modulations are considered in the numerical results. We compare our proposed methods with both the quantized linear approaches and the non-linear mapping algorithms, and for clarity the following abbreviations are used throughout this section:

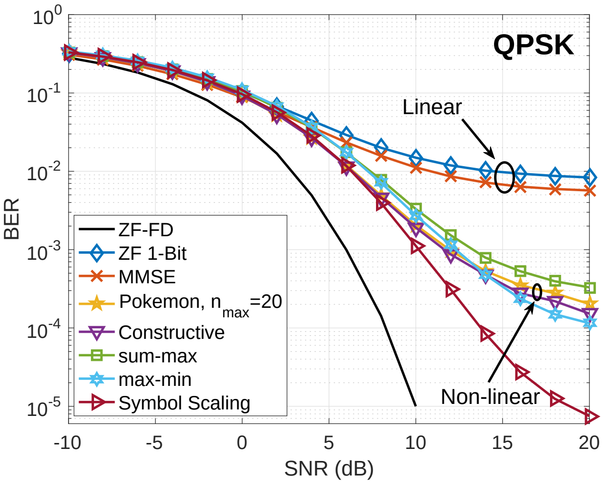

Fig. 4: BER v.s. transmit SNR, $N_t = 8$, $K = 2$, $n_{\max} = 20$, QPSK

1) 'ZF-FD': Unquantized ZF beamforming with infinite-precision DACs;
2) 'ZF 1-Bit': Quantized ZF approach with 1-bit DACs introduced in [10];
3) 'MMSE': MMSE-based quantized linear scheme in [11];
4) 'Pokemon, $n_{\max} = K$': Non-linear Pokemon algorithm proposed in [14] with K iterations;
5) 'Constructive': Proposed non-linear mapping scheme $\mathcal{P}_4$ in Section III-B;
6) 'sum-max': Proposed symbol scaling approach based on the 'sum-max' allocation scheme with Algorithm 1, 2 and 4;
7) 'max-min': Proposed symbol scaling approach based on the 'max-min' allocation scheme with Algorithm 1, 3 and 4;
8) 'Symbol Scaling': Proposed symbol scaling method obtained via Algorithm 5 where we select the best signal vector out of 'sum-max' or 'max-min' criteria.

In Fig. 4, we firstly consider a moderate scale MIMO with a total number of $N_t = 8$ transmit antennas at the BS and $K = 2$ single-antenna users in the system. For approaches with 1-bit quantization, we observe that the proposed symbol scaling scheme based on Algorithm 5 achieves the best BER performance, while both the proposed non-linear mapping scheme and 'Pokemon' achieve an inferior performance. This is because both the non-linear mapping method and the 'Pokemon' approach involve the relaxation-normalization process. For small-scale MIMO systems, based on Table I we can infer that η will be large in this case, which means that the deviation of the solution obtained by the relaxation-normalization process from the solution of the original 1-bit optimization problem is large, and the normalization process may lead to further detection errors. For the proposed symbol scaling scheme, the performance is promising since we directly select the quantized signal for each antenna element and therefore no relaxation or quantization is needed.

We then consider a massive MIMO system with $N_t = 128$

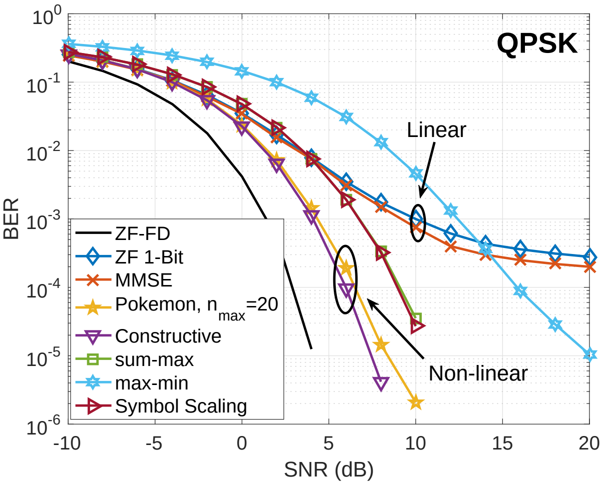

Fig. 5: BER v.s. transmit SNR, $N_t = 128$, $K = 16$, $n_{\max} = 20$, QPSK

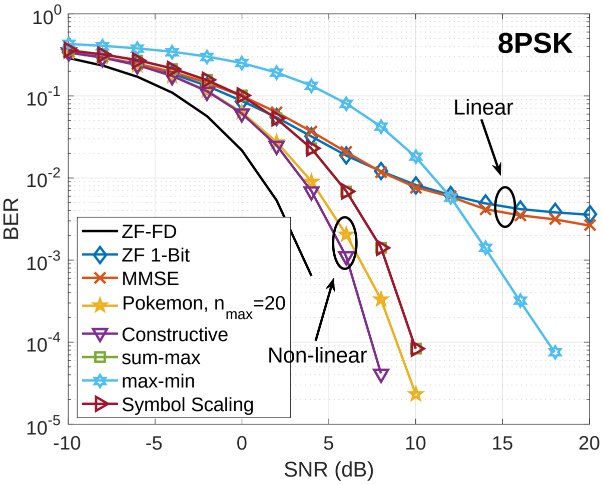

Fig. 6: BER v.s. transmit SNR, $N_t = 128$, $K = 8$, $n_{\max} = 20$, 8-PSK

transmit antennas and $K = 16$ users in Fig. 5. In the case of massive MIMO, all the schemes can achieve a lower BER thanks to the large number of antennas at the BS, and generally non-linear schemes outperform linear schemes. For approaches with 1-bit DACs, the proposed non-linear mapping method outperforms the non-linear 'Pokemon' algorithm and achieves the best BER performance. As for the proposed low-complexity symbol scaling scheme, by comparing Fig. 4 and Fig. 5, we can observe that the 'Max-Min' criterion is most suitable for small-scale MIMO systems, while the 'Sum-Max' criterion is more favourable for massive MIMO systems. Moreover, while we have observed around a 2dB SNR loss compared to the 'Pokemon' algorithm in the case of massive MIMO, its computational cost is approximately 12% of that for Pokemon in this scenario, which is shown mathematically in Table II and will be shown numerically in Fig. 7.

In Fig. 6, we show the performance of different schemes for 8-PSK modulation with $N_t = 128$ and $K = 8$. For 1-bit quantized beamforming approaches, it is observed that the proposed optimization-based non-linear scheme achieves the best BER performance. For the symbol scaling approach, it is observed that in the case of 8-PSK, only a 1dB SNR loss is observed compared to the non-linear iterative 'Pokemon' algorithm, and therefore the proposed low-complexity symbol scaling approach is more favourable in terms of the performance and complexity tradeoff.

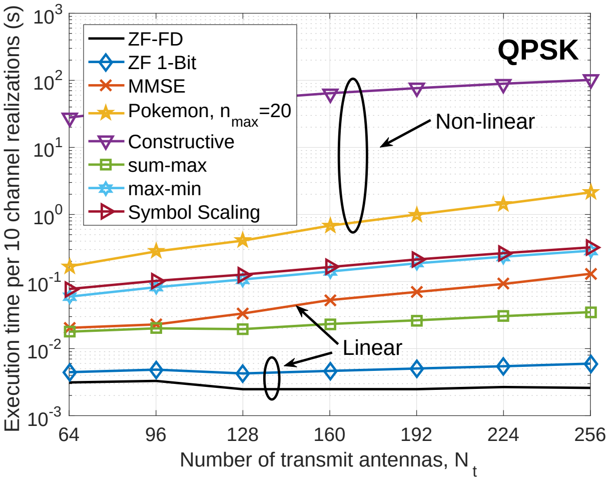

Fig. 7: Execution time for each scheme per 10 channel realizations, $K = 4$, $n_{\max} = 20$, QPSK

In Fig. 7, we compare the computational complexity of each approach in terms of the execution time required per 10 channel realizations. It is not surprising to observe that the computational cost of the proposed non-linear scheme is the highest. Compared to the non-linear 'Pokemon' algorithm, the execution time required for the proposed symbol scaling method is much less, especially for the 'sum-max' case. For 'Symbol Scaling' that returns the best performance based on Algorithm 5, the execution time required is similar to that of the 'max-min', which validates our analysis in Section V-C that most of the computational cost in the allocation stage comes from the 'Max-Min' criterion. Moreover, it is observed that the execution time of 'Symbol Scaling' is approximately 12% of that of the 'Pokemon' scheme in Fig. 7. This matches our analysis in (79) ($\frac{\mathrm{C_L}}{\mathrm{C_P}} \approx 0.12$ when $n_{\max} = 20$), and the above complexity gains of the proposed symbol scaling approach therefore favour its practical application.

To further compare the proposed schemes with 'Pokemon', in Fig. 8 we present the BER performance with different number of iterations for Pokemon. The number of iterations does not have an effect on other methods and therefore the BER for the other methods remains constant. It is observed that the performance of Pokemon improves as $n_{\max}$ increases. Nevertheless, we note that the improvement becomes less significant with a larger $n_{\max}$ and Pokemon achieves its best performance when $n_{\max}$ is around 25. An important observation is when $n_{\max} = 2, 3$, where the computational cost of Pokemon and our proposed scheme is similar, as shown by (79), and our proposed symbol scaling approach is

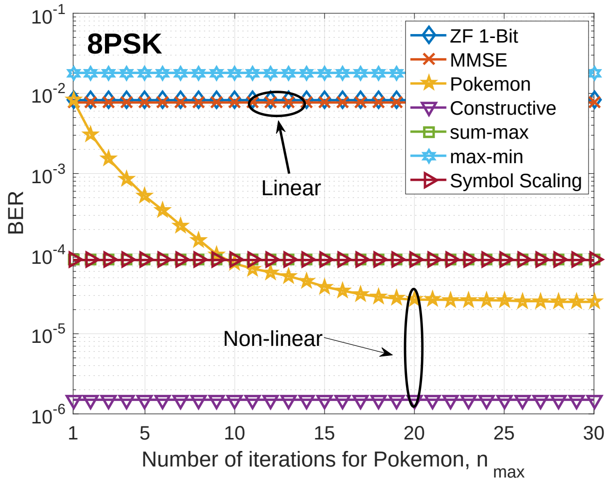

Fig. 8: BER v.s. Pokemon iteration number $n_{\max}$, $N_t = 128$, $K = 8$, $\rho = 10$dB, 8-PSK

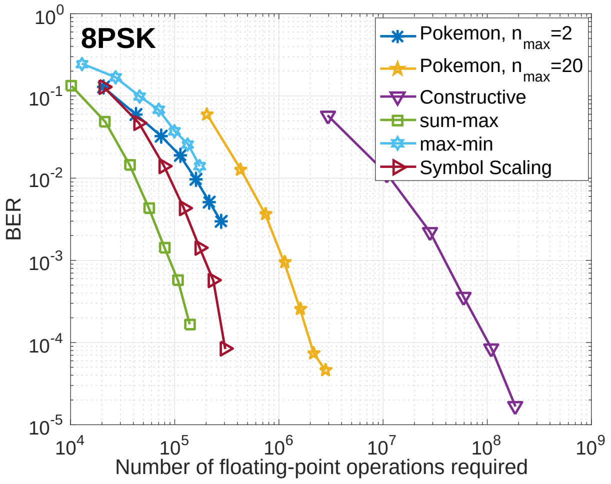

Fig. 9: BER v.s. analytical floating operations required, $K = 8$, $\rho = 10$dB, 8-PSK

shown to achieve an improved performance, which validates the superiority of the proposed approach.

To demonstrate the performance-complexity tradeoff directly, in Fig. 9 we depict the BER with respect to the number of floating-point operations required for a range of transmit antennas from $N_t = 32$ to $N_t = 128$, where the number of users is fixed as $K = 8$. It can be observed that the proposed optimization-based method achieves the best performance at the cost of the highest complexity. An important comparison is between the proposed 'Symbol Scaling' approach and the 'Pokemon' scheme with $n_{\max} = 2$, where we observe a significant performance gain of our proposed algorithm for the same computational complexity, especially when the number of antennas is large. Moreover, while the performance of the proposed low-complexity method based on 'sum-max' achieves an inferior performance to the 'Symbol Scaling' approach when N_t is large, it indeed achieves a better BER performance with a lower computational cost compared to Pokemon with $n_{\max} = 2$. Both of the above observations reveal the superiority of the proposed scheme based on symbol scaling.

All the above results are based on the assumption of perfect CSI. In the following, we numerically investigate the performance of the proposed approaches with imperfect CSI. The channel estimation techniques for massive MIMO with 1-bit quantization is an ongoing topic of research [9], [24], and an exact model for the imperfect CSI for this scenario is still not known. Therefore, in the following we employ a generic CSI model, where the BS only has knowledge of a noisy version of $\mathbf{H}$, given by

$$\hat{\mathbf{H}} = \mathbf{H} + \mathbf{Q}. \tag{80}$$

In (80), $\hat{\mathbf{H}}$ is the obtained CSI at the BS. $\mathbf{Q}$ denotes an error matrix with $\mathbf{Q} \sim \mathcal{CN}(\mathbf{0}, \delta \cdot \mathbf{I})$, where δ denotes the variance of the channel error. δ is modelled as inversely proportional to the transmit SNR and is expressed as $\delta = \beta/\rho$, where β denotes the error coefficient [16]. The BER result with imperfect CSI is depicted in Fig. 10, where a similar trend can be observed. We can further observe that the proposed non-linear mapping method still achieves the best performance among the schemes with 1-bit quantization in the case of imperfect CSI, while the proposed low-complexity symbol scaling approach can achieve a comparable performance with a greatly reduced computational cost.

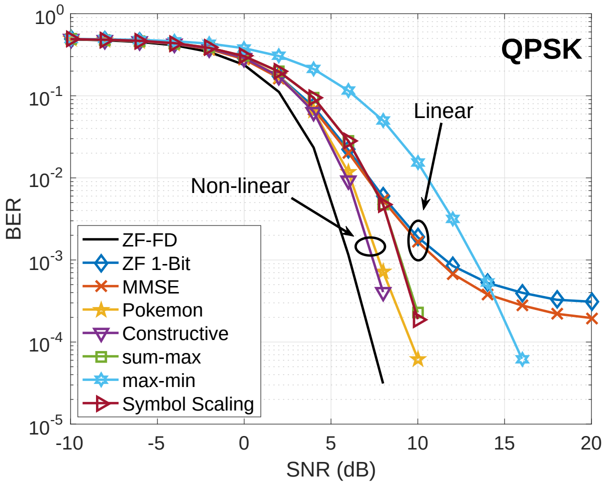

Fig. 10: BER v.s. transmit SNR, $N_t = 128$, $K = 16$, $n_{\max} = 20$, QPSK, Imperfect CSI, $\beta = 2.5$

VII. Conclusion

In this paper, we propose several transmit beamforming schemes for the massive MIMO downlink with 1-bit DACs based on the formulation of constructive interference, and we consider both a quantized linear method and a non-linear mapping approach. With the analysis of the Lagrangian and KKT conditions, the quantized linear scheme is mathematically proven to be equivalent to the quantized ZF beamforming. For the proposed non-linear mapping scheme, it is shown to be non-convex and solved by firstly relaxing the 1-bit quantization constraint, followed by a normalization. We further propose a low-complexity symbol scaling approach, where the quantized transmit signals are directly obtained. Numerical results reveal the superiority of the proposed symbol scaling scheme in small-scale MIMO systems. In the case of massive MIMO, the performance advantage of the proposed non-linear mapping method is validated, while the proposed symbol scaling scheme achieves a better performance-complexity tradeoff, which favours its usefulness in practical systems.

Appendix
Coordinate Transformation

We employ 8-PSK modulation in Fig. 3 as the example to demonstrate the coordinate transformation, where we focus on the constellation point 'A' in Fig. 3. Then, in the conventional real-imaginary complex plane, for node 'B' in Fig. 3, we have

$$\overrightarrow{OB} = \mathbf{h}_k \mathbf{x}_T = B_r \cdot 1 + B_i \cdot j, \tag{81}$$

where 1 and j are the bases, and we denote (B_r, B_i) as the corresponding coordinates. Based on (8), B_r and B_i are obtained as

$$\begin{aligned} B_r &= \Re\left(\mathbf{h}_k \mathbf{x}_T\right) = \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} - \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}, \\ B_i &= \Im\left(\mathbf{h}_k \mathbf{x}_T\right) = \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} + \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}. \end{aligned} \tag{82}$$

In the plane expanded by the two detection thresholds that correspond to the constellation point 'A', following (42) $\overrightarrow{OB}$ is decomposed into

$$\overrightarrow{OB} = \mathbf{h}_k \mathbf{x}_T = \alpha_k^{\Re} s_k^{\Re} + \alpha_k^{\Im} s_k^{\Im}. \tag{83}$$

Based on (41) and the fact that $\alpha_k^{\Re}$ and $\alpha_k^{\Im}$ are real numbers, (83) is further transformed into

$$\begin{aligned} \mathbf{h}_k \mathbf{x}_T &= \alpha_k^{\Re}\left(a_k + b_k \cdot j\right) + \alpha_k^{\Im}\left(c_k + d_k \cdot j\right) \\ &= \left(a_k \alpha_k^{\Re} + c_k \alpha_k^{\Im}\right) + \left(b_k \alpha_k^{\Re} + d_k \alpha_k^{\Im}\right) \cdot j. \end{aligned} \tag{84}$$

By substituting (82) into (84), we obtain

$$\begin{aligned} B_r &= \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} - \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im} = a_k \alpha_k^{\Re} + c_k \alpha_k^{\Im}, \\ B_i &= \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} + \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im} = b_k \alpha_k^{\Re} + d_k \alpha_k^{\Im}, \end{aligned} \tag{85}$$

which leads to the expression of $\alpha_k^{\Re}$ and $\alpha_k^{\Im}$, given by

$$\begin{aligned} \alpha_k^{\Re} &= \frac{d_k B_r - c_k B_i}{a_k d_k - b_k c_k} \\ &= \frac{d_k\left[\Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} - \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}\right] - c_k\left[\Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} + \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}\right]}{a_k d_k - b_k c_k} \\ &= \frac{d_k \Re\left(\mathbf{h}_k\right) - c_k \Im\left(\mathbf{h}_k\right)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Re} - \frac{d_k \Im\left(\mathbf{h}_k\right) + c_k \Re\left(\mathbf{h}_k\right)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Im}, \end{aligned} \tag{86}$$

and

$$\begin{aligned} \alpha_k^{\Im} &= \frac{a_k B_i - b_k B_r}{a_k d_k - b_k c_k} \\ &= \frac{a_k\left[\Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} + \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}\right] - b_k\left[\Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} - \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im}\right]}{a_k d_k - b_k c_k} \\ &= \frac{a_k \Im\left(\mathbf{h}_k\right) - b_k \Re\left(\mathbf{h}_k\right)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Re} + \frac{a_k \Re\left(\mathbf{h}_k\right) + b_k \Im\left(\mathbf{h}_k\right)}{a_k d_k - b_k c_k} \mathbf{x}_T^{\Im}. \end{aligned} \tag{87}$$

The extension to the constellation points of other PSK modulations can be similarly obtained and is omitted for brevity.

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Sig. Process. Mag.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[2] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, "Performance of MIMO Systems with Channel Inversion," in *Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No.02CH37367)*, vol. 1, 2002, pp. 35–39.

[3] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication-part I: Channel Inversion and Regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[4] S. Han, C. I. I, and C. Rowell, "Large-Scale Antenna Systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[5] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid Beamforming for Massive MIMO: A Survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, 2017.

[6] R. H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, April 1999.

[7] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink Performance of Wideband Massive MIMO with One-Bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Oct. 2016.

[8] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput Analysis of Massive MIMO Uplink with Low-Resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, June 2017.

[9] 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. Sig. Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

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

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

[12] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE Precoder for Massive MIMO Using 1-Bit Quantization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 3381–3385.

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

[14] O. Castaneda, T. Goldstein, and C. Studer, "POKEMON: A Non-Linear Beamforming Algorithm for 1-Bit Massive MIMO," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 3464–3468.

[15] C. Masouros, "Correlation Rotation Linear Precoding for MIMO Broadcast Communications," *IEEE Trans. Sig. Process.*, vol. 59, no. 1, pp. 252–262, Jan. 2011.

[16] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Vector Perturbation based on Symbol Scaling for Limited Feedback MISO Downlinks," *IEEE Trans. Sig. Process.*, vol. 62, no. 3, pp. 562–571, Feb. 2014.

[17] C. Masouros and G. Zheng, "Exploiting Known Interference as Green Signal Power for Downlink Beamforming Optimization," *IEEE Trans. Sig. Process.*, vol. 63, no. 14, pp. 3628–3640, July 2015.

[18] G. Zheng, I. Krikidis, C. Masouros, S. Timotheou, D. A. Toumpakaris, and Z. Ding, "Rethinking the Role of Interference in Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 152–158, Nov. 2014.

[19] C. Masouros, T. Ratnarajah, M. Sellathurai, C. B. Papadias, and A. K. Shukla, "Known Interference in the Cellular Downlink: A Performance Limiting Factor or a Source of Green Signal Power?" *IEEE Commun. Mag.*, vol. 51, no. 10, pp. 162–171, Oct. 2013.

[20] M. Alodeh, S. Chatzinotas, and B. Ottersten, "Constructive Multiuser Interference in Symbol Level Precoding for the MISO Downlink Channel," *IEEE Trans. Sig. Process.*, vol. 63, no. 9, pp. 2239–2252, May 2015.

[21] 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*, Available online: https://arxiv.org/abs/1704.06426, 2017.

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

[23] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001.

[24] C. Stockle, J. Munir, A. Mezghani, and J. A. Nossek, "Channel Estimation in Massive MIMO Systems Using 1-Bit Quantization," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Edinburgh, 2016, pp. 1–6.