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

Ang Li, *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 systems and focus on downlink transmission for PSK modulation, where the base station 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. The non-convex constraints of the 1-bit DACs are first relaxed into convex, 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 the corresponding performance-complexity tradeoff is more favorable for the proposed symbol scaling method.**

***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 [2]–[5]. It has also been shown that low-complexity linear precoding approaches such as zero-forcing (ZF) [6] and regularized ZF (RZF) [7] achieve close-to-optimal performance in the massive MIMO regime. Nevertheless, with a large number of antennas employed at the BS, a large number of radio frequency (RF) chains and high-resolution digital-to-analog converters (DACs) are needed for a fully-digital massive MIMO BS, which makes the hardware complexity and the corresponding cost prohibitively high. Moreover, the large number of hardware components will also result in a significant increase in the power consumption at the BS. All of the above make the fully-digital massive MIMO difficult to implement in practice. To achieve a compromise between the performance, hardware complexity and the consequent power consumption in practical massive MIMO systems, hybrid analog digital beamforming [8]–[13] 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 [14]–[16] that the power consumption of DACs grows exponentially with the resolution and linearly with the bandwidth in the downlink. 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 deployed at the BS, a large number of high-resolution DACs are required, which poses a significant practical challenge. Therefore, employing low-resolution DACs, especially 1-bit DACs, can greatly simplify the hardware design for the massive MIMO BSs, and further 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 [17]–[19], 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 increas-

Manuscript received October 6, 2017; revised February 28, 2018 and June 10, 2018; accepted August 23, 2018. Date of publication September 13, 2018; date of current version November 9, 2018. This work was supported in part by the Royal Academy of Engineering, U.K., in part by the Engineering and Physical Sciences Research Council under Project EP/M014150/1, in part by the China Scholarship Council in part by the Engineering and Physical Sciences Research Council under Project EP/R007934/1, and in part by the Marie Skłodowska-Curie Individual Fellowship Grant 793345. The associate editor coordinating the review of this paper and approving it for publication was I. Bergel. *(Corresponding author: Ang Li.)*

A. Li was with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K. He is now with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: ang.li.14@ucl.ac.uk).

C. Masouros is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K. (e-mail: c.masouros@ucl.ac.uk).

F. Liu was with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K. He is now with the School of Information and Electronics, Beijing Institute of Technology, Beijing 10081, 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 at Irvine, Irvine, CA 92697 USA, and also with the Institute for Advanced Study, Technical University of Munich, 80333 Munich, Germany (e-mail: swindle@uci.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2868369

ing number of studies due to the benefits mentioned above [16], [20]–[29]. Jacobsson *et al.* [16] consider both linear and non-linear precoding methods for 1-bit DAC downlink transmission, and propose several non-linear computationally-expensive precoding methods based on semi-definite relaxation (SDR), squared l_∞-norm relaxation and sphere precoding. In [20] and [21], a simple quantized ZF scheme is considered, where the transmit signal vector is obtained by a direct quantization on the ZF-precoded signals. In [22] and [23], quantized linear beamforming schemes based on minimum-mean squared error (MMSE) are proposed, whose performance is shown to be superior to the quantized ZF scheme in [20]. In [24], a non-linear symbol perturbation technique is introduced for QPSK modulation in the one-bit massive MIMO downlink. In [25] and [26], 1-bit downlink precoding schemes are proposed based on gradient methods. In [27], a branch-bound search method is proposed for MIMO downlink transmission with 1-bit quantization, which is however difficult to implement in massive MIMO due to the prohibitive complexity of the branch-bound algorithm. In [28] and [29], 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 since they ignore information regarding the finite alphabet nature of the transmit signals. As an example, for PSK signals, a more reliable decoding at the receiver can be obtained by forcing the received signal to lie deeper within each PSK decision region, farther away from the nominal constellation point, a process that would actually increase the mean-squared error (MSE). The methods in [30]–[35] take advantage of this observation by rotating the multiuser interference so that it adds constructively with the desired signal and improves the bit error rate (BER). Moreover, while there have been studies on downlink beamforming schemes with 1-bit DACs, most of these existing schemes either suffer a severe performance degradation [20]–[23] compared to the unquantized case, or require sophisticated optimizations and iterative algorithms that are computationally inefficient [28].

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 [30]–[35]. 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, which leads to a minimum uncoded error rate. By mathematically analyzing the optimization problem with the KKT conditions, it is shown that the optimal solution 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 from a significant performance loss compared to the unquantized case.

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. For this non-convex problem, as in [36] we apply a relaxation on the mathematical constraint resulting from the use of 1-bit DACs, such that the relaxed problem becomes convex, and can be efficiently solved. 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 the 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-point operations required. Numerical results show that in the case of small-scale MIMO systems, the proposed symbol scaling scheme achieves the best performance in terms of the BER. In the case of massive MIMO, the optimization-based non-linear scheme achieves an improved performance over existing schemes and better approaches the performance of the unquantized scheme, while the proposed symbol scaling algorithm can achieve a comparable performance. In terms of the computational complexity, it is demonstrated that the complexity of the 'symbol scaling' method is lower compared to that of the existing non-linear 1-bit precoding schemes and an improved performance-complexity tradeoff is observed, which favors its usefulness in practice.

We summarize the contributions of the paper below:

1) We propose downlink beamforming schemes for massive MIMO with 1-bit DACs based on the constructive interference formulation. We first consider a quantized linear beamforming scheme, where it is analytically proven that, in the massive MIMO region, optimality is achieved

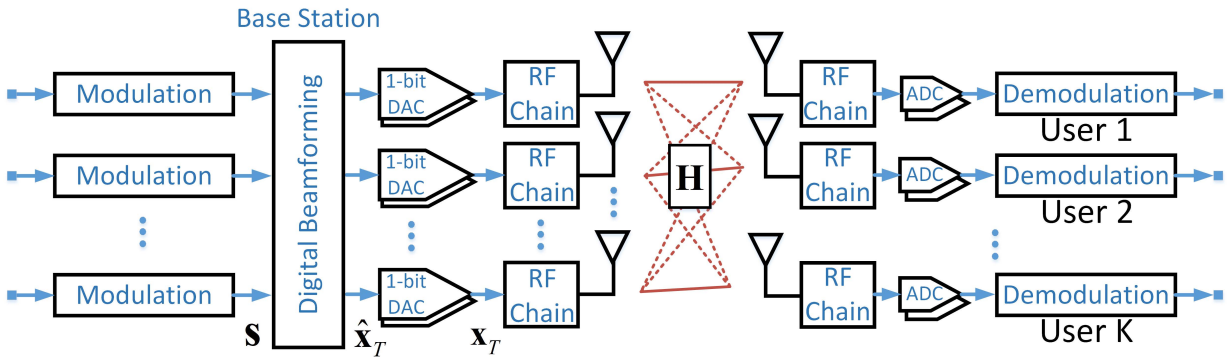

Fig. 1. Massive MIMO downlink system model with 1-bit DACs.

by employing a strict phase rotation due to the favorable propagation conditions.

2) We then consider a non-linear mapping approach where we directly optimize the transmit signal vector, which leads to a non-convex optimization due to the 1-bit quantization. As in [36], we relax the non-convex constraints such that the relaxed problem becomes convex and can be solved, 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 approach can achieve a comparable performance to the optimization-based non-linear mapping scheme.
4) We further study the computational costs of the symbol scaling schemes in terms of the required number of real-valued multiplications. Compared with existing algorithms, an improved performance-complexity trade-off is observed for the proposed symbol scaling methods.

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 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, and $\mathrm{sgn}\,[\cdot]$ is the sign function. j denotes the imaginary unit, and $\mathrm{vec}\,(\cdot)$ denotes the vectorization operation. $\mathbf{a}\,(k)$ denotes the k-th entry in vector $\mathbf{a}$, and $|\cdot|$ denotes the modulus of a complex number or the absolute value of a real number. $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathcal{C}^{n\times n}$ represents an $n \times n$ matrix in the complex set. $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary part of a complex number, respectively. $\mathbf{I}$ denotes the identity matrix, and $\mathbf{e}_l$ denotes the l-th column of $\mathbf{I}$.

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 with single-antenna receivers, 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 [20]–[26], [36], 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

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

where $\mathcal{B}$ denotes a general linear or non-linear transformation. When a linear precoding scheme is employed as in Section III-B, $\mathcal{B}$ represents the linear precoding matrix that multiplies $\mathbf{s}$ before quantization, while in the case of non-linear precoding as in Section III-C or Section IV, $\mathcal{B}$ refers to a non-linear mapping that forms the transmit signals based on $\mathbf{s}$. With 1-bit DACs employed, the output signal vector is then obtained as

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

where $\mathcal{Q}$ 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

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

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

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

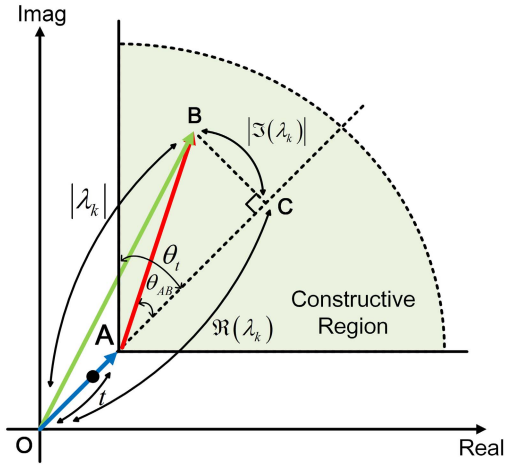

Fig. 2. Constructive interference and constructive region for QPSK.

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 additive Gaussian distributed noise 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 [30]–[32]. The exploitation of constructive interference was first introduced in [30] to improve the performance of the ZF precoding scheme, and was more recently applied to optimization-based approaches in [31], [32], [35], and [37] based on the constructive region, where it was further demonstrated that the interfering signals may not necessarily be strictly aligned with the data symbol. As long as the resulting signals plus interference are located in the constructive region, the distance to the detection thresholds is increased, and an improved performance can be expected [30]. Some applications of the constructive interference can be found in [38]–[41]. To illustrate the underlying concept intuitively, in Fig. 2 we depict the constructive region for QPSK modulation, where for simplicity and without loss of generality we focus on one quarter of the constellation. As can be observed, when the resulting interfered signal ($\vec{OB}$ in Fig. 2) is located in the constructive region, it is pushed further away from the detection thresholds. The mathematical 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

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

To introduce the proposed scheme, we first 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 first 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 first study the analytical constructive interference conditions. In Fig. 2, we denote $\vec{OA} = t \cdot s_k$ and note that $t = |\vec{OA}|$ 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 $\vec{OB} = \lambda_k s_k$, where λ_k is an introduced complex variable with $|\vec{OB}| = |\lambda_k|$. We can then obtain

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

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

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

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

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

Based on the expressions for $\vec{AC}$ and $\vec{CB}$,

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

In Fig. 2, it is observed that locating node 'B' 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

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

With the above technical details, we can formulate the optimization for the unquantized linear beamforming. The goal is to choose the linear precoder $\mathbf{W}$ in Eq. (5) to maximize the distance of the constructive region from the decision boundary, which is denoted by $t = |\vec{OA}|$ in Fig. 2. Mathematically,

we can formulate the optimization for the unquantized linear beamforming as

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

where $\mathcal{K}=\{1,2,\cdots,K\}$, and $\|\mathbf{W}\mathbf{s}\|_F\leq\sqrt{p_0}$ 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 where the users' channels experience independent Rayleigh fading, the favorable propagation property $\mathbf{H}\mathbf{H}^H\approx N_t\cdot\mathbf{I}$ holds, and the optimality conditions for each λ_k and t of the optimization problem $\mathcal{P}_1$ are obtained as

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

Proof: We prove the above proposition by analyzing the optimization problem $\mathcal{P}_1$ with the KKT conditions. We firstly transform $\mathcal{P}_1$ into a standard minimization problem, given by [32]

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

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 [42]

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

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

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

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

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

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

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

Based on (17b), 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

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

By denoting

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

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

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

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

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

In order to obtain $\mathbf{A}$, we first rewrite (8) in the compact form

$$\mathbf{H}\mathbf{W}\mathbf{s}=diag(\mathbf{\Omega})\,\mathbf{s},\tag{22}$$

where we introduce $\mathbf{\Omega}=[\lambda_1,\lambda_2,\cdots,\lambda_K]^T$. Then, by substituting (21) into (22), the matrix $\mathbf{A}$ can be obtained based on λ_k:

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

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

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

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

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

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

$$\begin{aligned}&\mathbf{s}^H diag\left(\mathbf{\Omega}^H\right)\left(\mathbf{H}\mathbf{H}^H\right)^{-1} diag\left(\mathbf{\Omega}\right)\mathbf{s} = p_0\\ &\quad\Rightarrow \mathbf{\Omega}^H diag\left(\mathbf{s}^H\right)\left(\mathbf{H}\mathbf{H}^H\right)^{-1} diag\left(\mathbf{s}\right)\mathbf{\Omega} = p_0\\ &\quad\Rightarrow \mathbf{\Omega}^H\mathbf{T}\mathbf{\Omega} = p_0,\end{aligned} \tag{26}$$

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 with uncorrelated Rayleigh fading, as $N_t \to \infty$, we have [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 as

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

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

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

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

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

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 with 1-bit quantization 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 from a performance loss compared to the ZF scheme with ideal DACs [20]. This then motivates the proposed non-linear mapping approach presented next that achieves improved performance.

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 [36], and based on the constructive interference formulation in [34]. We employ this approach to further design our low-complexity techniques in Section IV. The resulting optimization based on constructive interference can be formulated as

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

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). In the following, 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 [16], [36]

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

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

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

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 equality for 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

$$x_n = \frac{\mathrm{sgn}\left[\Re\left(\hat{x}_n\right)\right]}{\sqrt{2N_t}} + j\cdot\frac{\mathrm{sgn}\left[\Im\left(\hat{x}_n\right)\right]}{\sqrt{2N_t}},\quad \forall n\in\mathcal{N}, \tag{37}$$

where $\mathrm{sgn}\left[\cdot\right]$ is the sign function. 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

TABLE I

η WITH RESPECT TO THE NUMBER OF TRANSMIT ANTENNAS, $K = 4$, QPSK, 500 CHANNEL REALIZATIONS

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%

$\hat{\mathbf{x}}_T$ 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 constraint 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 an increase in the number of transmit antennas, and interestingly we observe that in the case of massive MIMO where each of the resulting λ_k is strictly real, the value of $(\eta_{\Re} + \eta_{\Im})$ is always equal to $(2K - 1)$, which explains why η is decreasing with an increasing number of antennas at the BS.

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. 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 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 method.

A. A New Look at the Constructive Interference Criteria

To introduce the proposed symbol scaling scheme, we first perform a coordinate transformation on the constructive interference constraint. To be specific, for $\mathcal{M}$-PSK modulations, each data symbol in the conventional real-imaginary plane can be expressed as

$$s_{(l)} = e^{j\cdot\left[\frac{2\pi}{\mathcal{M}}(l-1)+\frac{\pi}{4}\right]}, \quad l \in \{1, 2, \cdots \mathcal{M}\}, \tag{39}$$

where $s_{(l)}$ denotes the l-th constellation point. Given the constellation points, the equations that represent the two

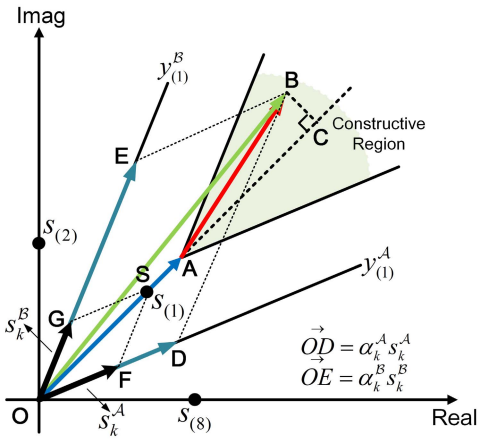

Fig. 3. Decomposition along the detection thresholds for 8-PSK.

detection thresholds for a specific constellation point $s_{(l)}$ can be expressed as

$$\begin{aligned} y^{\mathcal{A}}_{(l)} &= \tan\left[\frac{2\pi}{\mathcal{M}}(l-1) + \frac{\pi}{4} - \frac{\pi}{\mathcal{M}}\right]\cdot x \\ &= \tan\left[\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{3\pi}{\mathcal{M}}\right]\cdot x, \\ y^{\mathcal{B}}_{(l)} &= \tan\left[\frac{2\pi}{\mathcal{M}}(l-1) + \frac{\pi}{4} + \frac{\pi}{\mathcal{M}}\right]\cdot x \\ &= \tan\left[\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{\pi}{\mathcal{M}}\right]\cdot x. \end{aligned} \tag{40}$$

For the proposed symbol-scaling methods, without loss of generality we assume the data symbol for user k is $s_k = s_{(l)}$. We then propose to decompose the constellation points along their corresponding two detection thresholds, expressed as

$$s_k = s_{(l)} = s^{\mathcal{A}}_k + s^{\mathcal{B}}_k, \tag{41}$$

where $s^{\mathcal{A}}_k$ is parallel to $y^{\mathcal{A}}_{(l)}$ and $s^{\mathcal{B}}_k$ is parallel to $y^{\mathcal{B}}_{(l)}$. Accordingly, $s^{\mathcal{A}}_k$ and $s^{\mathcal{B}}_k$ can be expressed as

$$\begin{aligned} s^{\mathcal{A}}_k &= \frac{e^{j\cdot\left(\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{3\pi}{\mathcal{M}}\right)}}{\rho} = A^{\Re}_k + j\cdot A^{\Im}_k, \\ s^{\mathcal{B}}_k &= \frac{e^{j\cdot\left(\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{\pi}{\mathcal{M}}\right)}}{\rho} = B^{\Re}_k + j\cdot B^{\Im}_k, \end{aligned} \tag{42}$$

where $\left(A^{\Re}_k, A^{\Im}_k\right)$ and $\left(B^{\Re}_k, B^{\Im}_k\right)$ denote the coordinates of the bases $s^{\mathcal{A}}_k$ and $s^{\mathcal{B}}_k$ in the real-imaginary plane, respectively. The constant ρ is a scaling factor to guarantee that $s_k = s^{\mathcal{A}}_k + s^{\mathcal{B}}_k$. Note that for a normalized $\mathcal{M}$-PSK modulation, $|s_k| = 1$ and ρ is accordingly obtained as

$$\rho = \left| e^{j\cdot\left(\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{3\pi}{\mathcal{M}}\right)} + e^{j\cdot\left(\frac{2\pi}{\mathcal{M}}\cdot l + \frac{\pi}{4} - \frac{\pi}{\mathcal{M}}\right)} \right|. \tag{43}$$

The above decomposition is also shown geometrically in Fig. 3, where we employ 8-PSK modulation as an example. Specifically, for the considered constellation point in Fig. 3,

we obtain $\vec{OS} = s_{(1)}$, which further leads to

$$\vec{OF} = s_k^{\mathcal{A}} = \frac{e^{j\cdot\left(\frac{2\pi}{8}\cdot 1+\frac{\pi}{4}-\frac{3\pi}{8}\right)}}{\rho} = \frac{e^{j\cdot\frac{\pi}{8}}}{\left|e^{j\cdot\frac{\pi}{8}}+e^{j\cdot\frac{3\pi}{8}}\right|},$$
$$\vec{OG} = s_k^{\mathcal{B}} = \frac{e^{j\cdot\left(\frac{2\pi}{8}\cdot 1+\frac{\pi}{4}-\frac{\pi}{8}\right)}}{\rho} = \frac{e^{j\cdot\frac{3\pi}{8}}}{\left|e^{j\cdot\frac{\pi}{8}}+e^{j\cdot\frac{3\pi}{8}}\right|}. \quad (44)$$

Then for each k, instead of employing a complex scaling value λ_k that is multiplied by s_k, with the above formulation (41)-(43) we introduce a symbol scaling approach where we decompose (8) along the two corresponding detection thresholds of s_k, given by

$$\mathbf{h}_k\mathbf{x}_T = \alpha_k^{\mathcal{A}} s_k^{\mathcal{A}} + \alpha_k^{\mathcal{B}} s_k^{\mathcal{B}}, \quad (45)$$

where

$$\alpha_k^{\mathcal{A}} \geq 0, \quad \alpha_k^{\mathcal{B}} \geq 0, \quad \forall k \in \mathcal{K} \quad (46)$$

are scaling factors. We observe that a larger value of $\alpha_k^{\mathcal{A}}$ or $\alpha_k^{\mathcal{B}}$ represents a larger distance to the detection threshold, and by expanding (45) using the coordinate transformation, we can obtain a generic expression of $\alpha_k^{\mathcal{A}}$ and $\alpha_k^{\mathcal{B}}$ as a function of the transmit signal vector, given by (see Appendix)

$$\alpha_k^{\mathcal{A}} = \frac{B_k^{\Im}\mathbf{h}_k^{\Re} - B_k^{\Re}\mathbf{h}_k^{\Im}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}\mathbf{x}_T^{\Re} - \frac{B_k^{\Im}\mathbf{h}_k^{\Im} + B_k^{\Re}\mathbf{h}_k^{\Re}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}\mathbf{x}_T^{\Im},$$
$$\alpha_k^{\mathcal{B}} = \frac{A_k^{\Re}\mathbf{h}_k^{\Im} - A_k^{\Im}\mathbf{h}_k^{\Re}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}\mathbf{x}_T^{\Re} + \frac{A_k^{\Re}\mathbf{h}_k^{\Re} + A_k^{\Im}\mathbf{h}_k^{\Im}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}\mathbf{x}_T^{\Im}. \quad (47)$$

In (47), for simplicity we have employed the following notation

$$\mathbf{x}_T^{\Re} = \Re(\mathbf{x}_T), \quad \mathbf{x}_T^{\Im} = \Im(\mathbf{x}_T), \quad \mathbf{h}_k^{\Re} = \Re(\mathbf{h}_k),$$
$$\mathbf{h}_k^{\Im} = \Im(\mathbf{h}_k). \quad (48)$$

By further denoting

$$\mathbf{A}_k = \frac{B_k^{\Im}\mathbf{h}_k^{\Re} - B_k^{\Re}\mathbf{h}_k^{\Im}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}, \quad \mathbf{B}_k = -\frac{B_k^{\Im}\mathbf{h}_k^{\Im} + B_k^{\Re}\mathbf{h}_k^{\Re}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}},$$
$$\mathbf{C}_k = \frac{A_k^{\Re}\mathbf{h}_k^{\Im} - A_k^{\Im}\mathbf{h}_k^{\Re}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}, \quad \mathbf{D}_k = \frac{A_k^{\Re}\mathbf{h}_k^{\Re} + A_k^{\Im}\mathbf{h}_k^{\Im}}{A_k^{\Re}B_k^{\Im} - A_k^{\Im}B_k^{\Re}}, \quad (49)$$

the formulation of (47) is simplified to

$$\alpha_k^{\mathcal{A}} = \mathbf{A}_k\mathbf{x}_T^{\Re} + \mathbf{B}_k\mathbf{x}_T^{\Im},$$
$$\alpha_k^{\mathcal{B}} = \mathbf{C}_k\mathbf{x}_T^{\Re} + \mathbf{D}_k\mathbf{x}_T^{\Im}. \quad (50)$$

By defining

$$\mathbf{R}_k = \begin{bmatrix}\mathbf{A}_k & \mathbf{B}_k\end{bmatrix}, \quad \mathbf{I}_k = \begin{bmatrix}\mathbf{C}_k & \mathbf{D}_k\end{bmatrix}, \quad (51)$$

and

$$\mathbf{x} = \begin{bmatrix}\left(\mathbf{x}_T^{\Re}\right)^T & \left(\mathbf{x}_T^{\Im}\right)^T\end{bmatrix}^T,$$
$$\mathbf{\Lambda} = \left[\alpha_1^{\mathcal{A}}, \ldots, \alpha_K^{\mathcal{A}}, \alpha_1^{\mathcal{B}}, \ldots, \alpha_K^{\mathcal{B}}\right]^T, \quad (52)$$

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

$$\mathbf{\Lambda} = \mathbf{M}\mathbf{x}, \quad (53)$$

where $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{bmatrix}\mathbf{R}_1^T & \cdots & \mathbf{R}_K^T & \mathbf{I}_1^T & \cdots & \mathbf{I}_K^T\end{bmatrix}^T. \quad (54)$$

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

$$\mathcal{P}_5: \max_{\mathbf{x},\mathbf{\Lambda}} \min_{l} \alpha_l$$
$$s.t.\ \mathbf{\Lambda} = \mathbf{M}\mathbf{x}$$
$$\alpha_l \geq 0, \quad \forall l \in \mathcal{L}$$
$$x_i^E \in \left\{\frac{1}{\sqrt{2N_t}}, -\frac{1}{\sqrt{2N_t}}\right\}, \quad \forall i \in \mathcal{I} \quad (55)$$

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, each presented in detail below.

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 (53) into

$$\mathbf{\Lambda} = \sum_{i=1}^{2N_t} \mathbf{m}_i x_i^E, \quad (56)$$

where

$$\mathbf{M} = \begin{bmatrix}\mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_{2N_t}\end{bmatrix}, \quad (57)$$

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

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

$$x_i^E = \frac{\mathrm{SGN}[\mathbf{m}_i]}{\sqrt{2N_t}}, \quad \forall i \in \mathcal{S}, \quad (58)$$

where $\mathrm{SGN}[\mathbf{a}] = 1$ only when each entry in the vector has the same sign, and is equal to 0 otherwise. $\mathcal{S}$ 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

$$\mathbf{t} = \sum_{i\in\mathcal{V}} \mathbf{m}_i x_i^E, \quad (59)$$

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 $card(\mathcal{V}) = 2N_t$, 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. $card(\mathcal{S}) = 1$, with the following criterion:

$$i = \arg\max_{i\in\mathcal{I}} \left|\sum_{n=1}^{2K} \mathbf{m}_i(n)\right|, \quad (60)$$

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

$$x_i^E = \frac{\operatorname{sgn}\left[\sum_{n=1}^{2K} \mathbf{m}_i(n)\right]}{\sqrt{2N_t}}. \tag{61}$$

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

Algorithm 1 Initialization Stage

input : $\mathbf{s}$, $\mathbf{H}$
output : $\mathbf{t}$, $\mathcal{V}$
Decompose each $s_k = s_k^{\mathcal{A}} + s_k^{\mathcal{B}}$ based on modulation type;
Obtain $\mathbf{M}$ based on (41)-(53);
Find $\mathbf{m}_i$ that satisfies the sign-identity condition;
Obtain $\mathcal{S}$;
if $\mathcal{S} \neq \emptyset$ **then**
  $x_i^E = \frac{\operatorname{sgn}(\mathbf{m}_i)}{\sqrt{2N_t}}$, $\forall i \in \mathcal{S}$;
  $\mathcal{V} = \mathcal{S}$;
else
  Obtain i based on (60), $x_i^E = \frac{\operatorname{sgn}(\|\mathbf{m}_i\|_1)}{\sqrt{2N_t}}$;
  $\mathcal{V} = \{i\}$;
end if
Calculate $\mathbf{t}$ based on (59).

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

$$\mathcal{W} = \{i | i \in \mathcal{I} \text{ and } i \notin \mathcal{V}\}. \tag{62}$$

$\mathcal{W}$ 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

$$\mathcal{F}(\mathbf{x}) = \operatorname{sum}(\mathbf{\Lambda}), \tag{63}$$

where $\operatorname{sum}(\mathbf{\Lambda})$ returns the sum of the entries in the column vector $\mathbf{\Lambda}$. Then, based on (53) the objective can be further transformed as

$$\mathcal{F}(\mathbf{x}) = \mathbf{u}\mathbf{x} = \sum_{l=1}^{2K}\sum_{i=1}^{2N_t} \mathbf{m}_i(l)\, x_i^E, \tag{64}$$

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

$$\mathbf{u}(i) = \sum_{l=1}^{2K} \mathbf{m}_i(l). \tag{65}$$

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

$$x_i^E = \frac{\operatorname{sgn}[\mathbf{u}(i)]}{\sqrt{2N_t}}, \quad \forall i \in \mathcal{W}. \tag{66}$$

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 (62);
Calculate $\mathbf{u}$ and each $\mathbf{u}(i)$ based on (64), (65);
Allocate $x_i^E = \frac{\operatorname{sgn}[\mathbf{u}(i)]}{\sqrt{2N_t}}$, $\forall i \in \mathcal{W}$;
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

$$\mathbf{t}(q) = \min(\mathbf{t}), \tag{67}$$

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

$$i = \underset{i\in\mathcal{W}}{\arg\max} |\mathbf{m}_i(q)|, \tag{68}$$

and the corresponding x_i^E is then obtained as

$$x_i^E = \frac{\operatorname{sgn}[\mathbf{m}_i(q)]}{\sqrt{2N_t}}. \tag{69}$$

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.

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

Algorithm 3 Allocation Stage - 'Max-Min'

input : $\mathcal{V}$, $\mathbf{M}$, $\mathbf{t}$
output : $\mathbf{x}_{\max-\min}$
while $\mathcal{V} \neq \mathcal{I}$ **do**
  Calculate $\mathcal{W}$ based on (62);
  Obtain q that satisfies $\mathbf{t}(q) = \min(\mathbf{t})$;
  Find $i = \underset{i\in\mathcal{W}}{\arg\max} |\mathbf{m}_i(q)|$;
  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 $\mathbf{x}_{\max-\min}$.

Algorithm 4 Refinement Stage

input : $\mathbf{x}_{\mathrm{sum}-\max}$ (or $\mathbf{x}_{\max-\min}$)
output : $\mathbf{x}_T$
Denote $\mathbf{x}_{(0)} = \mathbf{x}_{\mathrm{sum}-\max}$ (or $\mathbf{x}_{\max-\min}$);
for $i = 1 : 2N_t$ **do**
  Calculate $\mathbf{\Lambda}_{(0)} = \mathbf{M}\mathbf{x}_{(0)}$;
  Obtain $\mathbf{x}_{(i)} = \left[x_1^E, \ldots, x_{i-1}^E, -x_i^E, x_{i+1}^E, \ldots, x_{2N_t}^E\right]^T$;
  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 $\mathbf{x}_{(0)}$.

for each entry in $\mathbf{x}_{(0)}$. The algorithm for the refinement stage is then shown in Algorithm 4.

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. We note that while it is observed from $\mathcal{P}_5$ that the max-min criterion is a better metric compared to the sum-max criterion, for the proposed symbol-scaling methods the refinement process seems to give an advantage to the sum-max criterion for small-scale MIMO systems, while the opposite is observed in the case of massive MIMO systems, as will be illustrated in the numerical results.

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 from the signal vectors obtained by the 'Sum-Max' and 'Max-Min' criteria.

V. Computational Complexity Analysis

In this section, following [43] and [44], we study the computational costs of the proposed schemes in terms of the required number of real-valued multiplications including both the pre-processing complexity and the per-iteration complexity. As a reference, we also study the complexity of the exhaustive search scheme, the non-linear 'SQUID' algorithm in [16], 'C1PO' and 'C2PO' schemes in [29]. For the optimization-based approach, the complexity is evaluated based on the time complexity bound introduced in [45].

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
1.'$\mathbf{Sum-Max}$' :
Obtain $\mathbf{x}_{\mathrm{sum}-\max}$ with Algorithm 2;
2.'$\mathbf{Max-Min}$' :
Obtain $\mathbf{x}_{\max-\min}$ with Algorithm 3;
Refinement Stage
Update both $\mathbf{x}_{\mathrm{sum}-\max}$ and $\mathbf{x}_{\max-\min}$ with Algorithm 4;
Calculate $\mathbf{\Lambda}_{\mathrm{s}} = \mathbf{M}\mathbf{x}_{\mathrm{sum}-\max}$ and $\mathbf{\Lambda}_{\mathrm{m}} = \mathbf{M}\mathbf{x}_{\max-\min}$;
if $\min(\mathbf{\Lambda}_{\mathrm{s}}) > \min(\mathbf{\Lambda}_{\mathrm{m}})$ **then**
  $\mathbf{x} = \mathbf{x}_{\mathrm{sum}-\max}$;
else
  $\mathbf{x} = \mathbf{x}_{\max-\min}$;
end if
Decompose $\mathbf{x} = \left[\left(\mathbf{x}_T^{\Re}\right)^T \left(\mathbf{x}_T^{\Im}\right)^T\right]^T$;
Output $\mathbf{x}_T = \mathbf{x}_T^{\Re} + \mathbf{x}_T^{\Im} \cdot j$.

A. Exhaustive Search

For massive MIMO transmission with 1-bit quantization, the output signal on each antenna element has 4 potential values, i.e., each $x_n \in \left\{\frac{1}{\sqrt{2}} + j\cdot\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} - j\cdot\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} + j\cdot\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} - j\cdot\frac{1}{\sqrt{2}}\right\}$. The exhaustive search method first searches all the possible signal combinations and then selects the best one, which means that there are a total number of 4^{N_t} signal combinations. For each signal combination, it takes $4KN_t$ real multiplications to compute $\mathbf{\Lambda}$ based on (53), as $\mathbf{M} \in \mathcal{C}^{2K\times 2N_t}$ [43]. Therefore, the total number of required real-valued multiplications for the exhaustive search scheme considering all the possible combinations is obtained as

$$\mathrm{C_E} = 4KN_t \cdot 4^{N_t} = KN_t \cdot 2^{2N_t+2}. \tag{70}$$

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. SQUID

Based on the description in [16], the major complexity for 'SQUID' lies in the calculation of

$$\mathbf{a}_{\mathbb{R}}^{(t+1)} = \left(2\mathbf{b}_{\mathbb{R}}^{(t)} - \mathbf{c}_{\mathbb{R}}^{(t)}\right) - \mathbf{Q}_{\mathbb{R}}\mathbf{H}_{\mathbb{R}}\left(2\mathbf{b}_{\mathbb{R}}^{(t)} - \mathbf{c}_{\mathbb{R}}^{(t)}\right) + \mathbf{d}_{\mathbb{R}} \tag{71}$$

within each iteration. Accordingly, the pre-processing step for 'SQUID' includes the calculation of

$$\mathbf{Q}_{\mathbb{R}} = \mathbf{H}_{\mathbb{R}}^H \left(\mathbf{H}_{\mathbb{R}} \mathbf{H}_{\mathbb{R}}^H + \frac{1}{2} \cdot \mathbf{I} \right)^{-1} \in \mathcal{R}^{2N_t \times 2K} \tag{72}$$

and

$$\mathbf{d}_{\mathbb{R}} = 2 \left(\mathbf{H}_{\mathbb{R}}^H \mathbf{s}_{\mathbb{R}} - \mathbf{Q}_{\mathbb{R}} \mathbf{H}_{\mathbb{R}} \mathbf{H}_{\mathbb{R}}^H \mathbf{s}_{\mathbb{R}} \right) \in \mathcal{R}^{2N_t \times 1}. \tag{73}$$

Following [46] where it is demonstrated that the computational cost of a complex matrix inverse in the form of $\left(\mathbf{H}\mathbf{H}^H + \frac{1}{2} \cdot \mathbf{I} \right)^{-1}$ with $\mathbf{H} \in \mathcal{C}^{K \times N_t}$ (or its equivalent real representation with $\mathbf{H}_{\mathbb{R}} \in \mathcal{R}^{2K \times 2N_t}$) is $\left(\frac{2}{3}K^3 + 2N_tK^2 + 4K^2 - \frac{2}{3}K \right)$, we obtain the pre-processing complexity for 'SQUID' as

$$\begin{aligned} \mathrm{C}_{\mathrm{SQUID}}^{\mathrm{Pre}} = \mathrm{C}_{\mathbf{Q}_{\mathbb{R}}} + \mathrm{C}_{\mathbf{d}_{\mathbb{R}}} = \frac{2}{3}K^3 + 2N_tK^2 \\ + 4K^2 - \frac{2}{3}K + 16KN_t. \end{aligned} \tag{74}$$

The per-iteration complexity for calculating $\mathbf{a}_{\mathbb{R}}^{(t+1)}$ is dominated by the calculation of $\mathbf{Q}_{\mathbb{R}} \mathbf{H}_{\mathbb{R}} \left(2\mathbf{b}_{\mathbb{R}}^{(t)} - \mathbf{c}_{\mathbb{R}}^{(t)} \right)$, which consumes

$$\mathrm{C}_{\mathrm{SQUID}}^{\mathrm{Iter}} = 4KN_t + 4KN_t = 8KN_t \tag{75}$$

real-valued multiplications. Finally, we obtain the total computational cost of the 'SQUID' algorithm as

$$\begin{aligned} \mathrm{C}_{\mathrm{SQUID}} &= \mathrm{C}_{\mathrm{SQUID}}^{\mathrm{Pre}} + \mathrm{N}_{\mathrm{SQUID}}^{\mathrm{Iter}} \cdot \mathrm{C}_{\mathrm{SQUID}}^{\mathrm{Iter}} \\ &= \frac{2}{3}K^3 + 2N_tK^2 + 4K^2 - \frac{2}{3}K + 16KN_t \\ &\quad + \mathrm{N}_{\mathrm{SQUID}}^{\mathrm{Iter}} \cdot 8KN_t. \end{aligned} \tag{76}$$

C. C1PO

Subsequently, we discuss the complexity of 'C1PO' proposed in [29] based on biconvex relaxation. According to the description in [29], the pre-processing step for 'C1PO' involves the calculation of

$$\begin{aligned} \mathbf{A} &= \left(\mathbf{I} - \frac{\mathbf{s}\mathbf{s}^H}{\|\mathbf{s}\|_2^2} \right) \mathbf{H} \in \mathcal{C}^{K \times N_t}, \\ \mathbf{Q} &= \left(\mathbf{I} + \frac{1}{\gamma} \mathbf{A}^H \mathbf{A} \right)^{-1} \in \mathcal{C}^{N_t \times N_t}. \end{aligned} \tag{77}$$

The calculation of $\mathbf{A}$ involves $\left(K^2 + KN_t \right)$ complex-valued multiplications, which is equivalent to $\left(4K^2 + 4KN_t \right)$ real-valued multiplications. Accordingly, the corresponding computational cost is obtained as

$$\begin{aligned} \mathrm{C}_{\mathrm{C1PO}}^{\mathrm{Pre}} &= \mathrm{C}_{\mathbf{A}} + \mathrm{C}_{\mathbf{Q}} \\ &= \frac{2}{3}N_t^3 + 2KN_t^2 + 4N_t^2 - \frac{2}{3}N_t + 4K^2 + 4KN_t. \end{aligned} \tag{78}$$

Within each iteration, the major complexity is from the calculation of

$$\mathbf{z}^{(t+1)} = \mathbf{Q}\mathbf{x}^{(t)}, \tag{79}$$

which leads to a per-iteration computational cost

$$\mathrm{C}_{\mathrm{C1PO}}^{\mathrm{Iter}} = 4N_t^2. \tag{80}$$

The final complexity expression for the 'C1PO' algorithm in [28] is obtained as

$$\begin{aligned} \mathrm{C}_{\mathrm{C1PO}} &= \mathrm{C}_{\mathrm{C1PO}}^{\mathrm{Pre}} + \mathrm{N}_{\mathrm{C1PO}}^{\mathrm{Iter}} \cdot \mathrm{C}_{\mathrm{C1PO}}^{\mathrm{Iter}} \\ &= \frac{2}{3}N_t^3 + 2KN_t^2 + 4N_t^2 - \frac{2}{3}N_t + 4K^2 + 4KN_t \\ &\quad + \mathrm{N}_{\mathrm{C1PO}}^{\mathrm{Iter}} \cdot 4N_t^2. \end{aligned} \tag{81}$$

Compared to the 'SQUID' method introduced in [16], generally we observe that the 'C1PO' method is more computationally expensive, mainly due to the inclusion of a $N_t \times N_t$ matrix inverse in calculating $\mathbf{Q}$ in the pre-processing step.

D. C2PO

We move on to evaluate the complexity of the 'C2PO' algorithm proposed in [29], which requires a significantly lower computational cost than 'SQUID' and 'C1PO' by removing the matrix inverse operation in the pre-processing step. To be more specific, the pre-processing step for 'C2PO' only involves the calculation of

$$\mathbf{v} = \frac{\mathbf{H}^H \mathbf{s}}{\|\mathbf{s}\|_2} \in \mathcal{C}^{N_t \times 1}, \tag{82}$$

which requires KN_t complex-valued multiplications, and is equivalent to

$$\mathrm{C}_{\mathrm{C2PO}}^{\mathrm{Pre}} = 4KN_t \tag{83}$$

real-valued multiplications. Within each iteration, the dominant complexity is from the calculation of

$$\mathbf{z}^{(t+1)} = \mathbf{x}^{(t)} - \tau \cdot \bar{\mathbf{H}}^{\gamma} \bar{\mathbf{H}} \mathbf{x}^{(t)}, \tag{84}$$

where $\bar{\mathbf{H}}^{\gamma} \in \mathcal{C}^{N_t \times (K+1)}$ and $\bar{\mathbf{H}} \in \mathcal{C}^{(K+1) \times N_t}$. Accordingly, we obtain the per-iteration complexity as

$$\mathrm{C}_{\mathrm{C2PO}}^{\mathrm{Iter}} = 8 \left(K + 1 \right) N_t, \tag{85}$$

and the total complexity for 'C2PO' as

$$\begin{aligned} \mathrm{C}_{\mathrm{C2PO}} &= \mathrm{C}_{\mathrm{C2PO}}^{\mathrm{Pre}} + \mathrm{N}_{\mathrm{C2PO}}^{\mathrm{Iter}} \cdot \mathrm{C}_{\mathrm{C2PO}}^{\mathrm{Iter}} \\ &= 4KN_t + \mathrm{N}_{\mathrm{C2PO}}^{\mathrm{Iter}} \cdot 8 \left(K + 1 \right) N_t. \end{aligned} \tag{86}$$

By removing the matrix inverse operation, we observe that the computational cost of 'C2PO' is significantly lower compared to the 'SQUID' and 'C1PO' methods. However, it will be shown in the numerical results that the performance of 'C2PO' is inferior to 'C1PO', 'SQUID' and the proposed 'Symbol Scaling' schemes.

E. Symbol Scaling Scheme

In the following, we calculate the computational cost for each stage of the proposed symbol scaling approach. Similar to the 'C2PO' method, our proposed 'Symbol Scaling' scheme does not require the matrix inverse operation, and the pre-processing step only involves the construction of $\mathbf{M}$ in (54). Accordingly, the pre-processing complexity is obtained as

$$\mathrm{C}_{\mathrm{SS}}^{\mathrm{Pre}} = 8KN_t. \tag{87}$$

Within the algorithm, the major complexity for the 'sum-max' criterion is from the calculation of $\mathcal{F}(\mathbf{x})$ in (64) and $\mathbf{\Lambda}_{(0)} = \mathbf{M}\mathbf{x}_{(0)}$, and the corresponding complexity is

$$\mathrm{C}_{\mathrm{SS}}^{\mathrm{S-M}} = 2N_t + 4KN_t. \tag{88}$$

For the 'max-min' criterion, the major computational cost is from the calculation of q and i. Since $card(\mathcal{V})$ is difficult to obtain analytically in the initialization stage, we consider a worst-case complexity where $card(\mathcal{V}) = 1$, and the corresponding complexity is obtained as

$$\mathrm{C}_{\mathrm{SS}}^{\mathrm{M-M}} = (2N_t - 1)(2K + 2N_t) = 4N_t^2 + 4KN_t - 2K - 2N_t. \tag{89}$$

In the refinement stage, within each iteration we only need to re-calculate the corresponding $\mathbf{m}_i \cdot \left(-x_i^E\right)$ for both criteria, and the corresponding complexity for the refinement stage is obtained as

$$\mathrm{C}_{\mathrm{SS}}^{\mathrm{Ref}} = 4KN_t. \tag{90}$$

Based on the above, we can express the total computational cost of the proposed 'Symbol Scaling' method in terms of the real-valued multiplications as

$$\begin{aligned}\mathrm{C}_{\mathrm{SS}} &= \mathrm{C}_{\mathrm{SS}}^{\mathrm{Pre}} + \mathrm{C}_{\mathrm{SS}}^{\mathrm{S-M}} + \mathrm{C}_{\mathrm{SS}}^{\mathrm{M-M}} + 2\mathrm{C}_{\mathrm{SS}}^{\mathrm{Ref}} \\ &= 4N_t^2 + 24KN_t - 2K.\end{aligned} \tag{91}$$

Based on the above analysis, we observe that the required complexity of our proposed 'Symbol Scaling' is comparable to 'C2PO' algorithm, and both are much more computationally efficient than the 'C1PO' and 'SQUID' algorithms, since the matrix inverse operation is avoided.

F. Optimization-Based Non-Linear Mapping $\mathcal{P}_3$

For the proposed non-linear mapping scheme, it is difficult to calculate the required number of multiplications and additions. Therefore, we resort to [45] and evaluate its complexity based on the arithmetic complexity.

For this non-convex optimization problem, the complexity is dominated by solving the relaxed convex problem $\mathcal{P}_4$ via the interior-point method [42]. Based on our reformulated $\mathcal{P}_5$ in Section IV, we first express the equivalent real representation of $\mathcal{P}_4$ in a standard form as

$$\begin{aligned}\mathcal{P}_6: \ &\max_{\mathbf{v}} \mathbf{c}^T\mathbf{v} \\ &s.t.\ \mathbf{q}_l\mathbf{v} \le 0, \quad \forall l \in \mathcal{L} \\ &\quad \mathbf{e}_{i+1}^T\mathbf{v} \le \frac{1}{\sqrt{2N_t}}, \quad \forall i \in \mathcal{I} \\ &\quad -\mathbf{e}_{i+1}^T\mathbf{v} \le \frac{1}{\sqrt{2N_t}}, \quad \forall i \in \mathcal{I} \\ &\quad \mathbf{v} = \left[t, x_1^E, x_2^E, \cdots, x_{2N_t}^E\right]^T \\ &\quad \mathbf{c} = [1, 0, 0, \cdots, 0]^T\end{aligned} \tag{92}$$

In $\mathcal{P}_6$, $\mathbf{q}_l = \left[1 \ -\hat{\mathbf{m}}_l\right]$, where $\hat{\mathbf{m}}_l$ denotes the l-th row of $\mathbf{M}$. Based on [45], the arithmetic complexity bound of the above optimization via the interior-point method is given by

$$\mathrm{C}_{\mathrm{N}} = (M+N)^{1.5}N^2 \cdot \mathrm{D}(\mathbf{p}, \varepsilon), \tag{93}$$

where ε is the accuracy of the solution, N denotes the dimension of the variable $\mathbf{v}$, and M is the total number of the constraints in the optimization. Based on the construction of $\mathcal{P}_6$, we obtain

$$M = 4N_t + 2K, \quad N = 2N_t + 1, \tag{94}$$

which further leads to the expression of C_{N} as

$$\mathrm{C}_{\mathrm{N}} = (6N_t + 2K + 1)^{1.5}(2N_t + 1)^2 \cdot \mathrm{D}(\mathbf{p}, \varepsilon). \tag{95}$$

$\mathrm{D}(\mathbf{p}, \varepsilon)$ is the number of digits of accuracy for a solution with the accuracy ε, and is given by

$$\mathrm{D}(\mathbf{p}, \varepsilon) = \ln\left(\frac{\mathrm{Dim}(\mathbf{p}) + \|\mathbf{p}\|_1 + \varepsilon^2}{\varepsilon}\right), \tag{96}$$

where the column vector $\mathbf{p}$ represents a permutation vector that contains the parameters in both the objective function and the constraints [45]. For our considered problem $\mathcal{P}_6$, $\mathbf{p}$ is given in (97), shown at the bottom of next page, which further leads to

$$\|\mathbf{p}\|_1 = 10N_t + 4K + 2\sqrt{2N_t} + \|\mathbf{M}\|_1 + 2. \tag{98}$$

In (96), $\mathrm{Dim}(\mathbf{p})$ denotes the dimension of the permutation vector $\mathbf{p}$, and is accordingly obtained as

$$\begin{aligned}\mathrm{Dim}(\mathbf{p}) &= (M+1)(N+2) + 2 \\ &= (4N_t + 2K + 1)(2N_t + 2) + 2 \\ &= 8N_t^2 + 10N_t + 4KN_t + 4K + 4.\end{aligned} \tag{99}$$

Given the expressions for $\mathrm{Dim}(\mathbf{p})$ and $\|\mathbf{p}\|_1$, we arrive at the final expression of the complexity for $\mathcal{P}_6$, which is shown in (100), as shown at the bottom of the next page.

Nevertheless, we note that the obtained complexity expression of the optimization-based method in (100) may not be directly comparable to that of the algorithm-based methods in (76), (81), (86) and (91), as the complexity of the optimization problem is obtained based on the time complexity bound, while the complexity of the algorithm-based schemes is obtained based on the exact required number of real operations.

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, and we compare our proposed methods with both the quantized linear approaches and the non-linear mapping algorithms. For iterative algorithms, the number of iterations is chosen to be the smallest value beyond which the performance of the algorithms does not improve significantly, as illustrated later in Fig. 8. For clarity, the following abbreviations are used throughout this section:

1) 'ZF Unquantized': Unquantized ZF precoding with infinite-precision DACs;
2) 'ZF 1-Bit': Quantized ZF approach with 1-bit DACs introduced in [20];
3) 'MMSE 1-Bit': MMSE-based quantized linear scheme in [22];

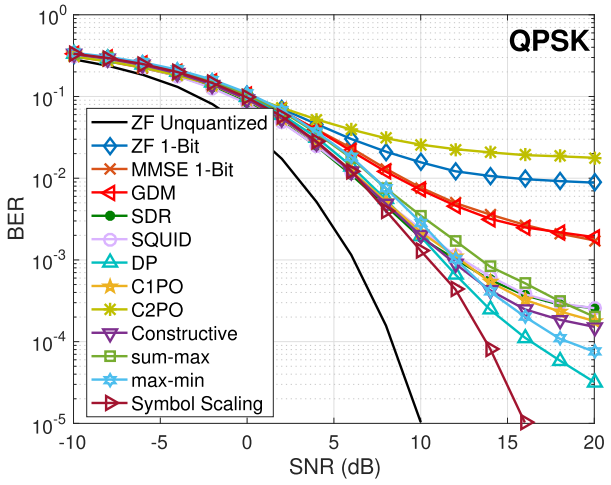

Fig. 4. BER vs. transmit SNR, $N_t = 8$, $K = 2$, $\mathrm{N}_{\mathrm{SQUID}}^{\mathrm{Iter}} = 50$, $\mathrm{N}_{\mathrm{C1PO}}^{\mathrm{Iter}} = 20$, $\mathrm{N}_{\mathrm{C2PO}}^{\mathrm{Iter}} = 20$, QPSK.

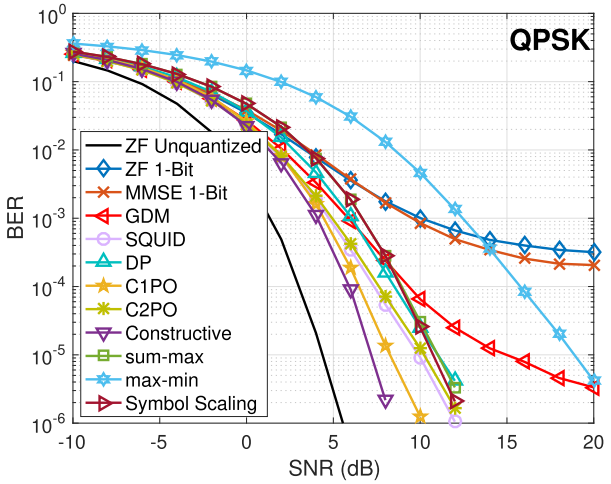

Fig. 5. BER vs. transmit SNR, $N_t = 128$, $K = 16$, $\mathrm{N}_{\mathrm{SQUID}}^{\mathrm{Iter}} = 50$, $\mathrm{N}_{\mathrm{C1PO}}^{\mathrm{Iter}} = 20$, $\mathrm{N}_{\mathrm{C2PO}}^{\mathrm{Iter}} = 20$, QPSK.

4) 'SQUID': Non-linear 'SQUID' method proposed in [16] based on the squared l_∞-norm relaxation with $\mathrm{N}_{\mathrm{SQUID}}^{\mathrm{Iter}} = 50$ iterations;
5) 'C1PO': Non-linear C1PO algorithm proposed in [29] with $\mathrm{N}_{\mathrm{C1PO}}^{\mathrm{Iter}} = 20$ iterations;
6) 'C2PO': Non-linear C2PO algorithm proposed in [29] with $\mathrm{N}_{\mathrm{C2PO}}^{\mathrm{Iter}} = 20$ iterations;
7) 'SDR': Non-linear method proposed in [16] based on the semidefinite relaxation (SDR);
8) 'DP': the direct perturbation technique proposed in [24] for QPSK modulation;
9) 'GDM': the gradient descend method proposed in [25];
10) 'Constructive': Proposed non-linear mapping scheme $\mathcal{P}_4$ in Section III-B;
11) 'sum-max': Proposed symbol scaling approach based on the 'sum-max' allocation scheme with Algorithm 1, 2 and 4;
12) 'max-min': Proposed symbol scaling approach based on the 'max-min' allocation scheme with Algorithm 1, 3 and 4;
13) '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 first consider a moderate scale MIMO system with $N_t = 8$ 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 other existing 1-bit precoding algorithms achieve an inferior performance. This is because both the non-linear mapping method and the 1-bit approaches in [16] and [29] involve the relaxation-normalization process. For small-scale MIMO systems, we can infer that η in (38) will be large, 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 $N_t = 128$ transmit antennas and $K = 16$ users in Fig. 5, where the SDR-based approach is not included due to its prohibitive complexity [16]. 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 optimizatio-based method 'Constructive' 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 favorable for massive MIMO systems.

$$\mathbf{p} = \Bigg[(2K+4N_t),(2N_t+1),1,\underbrace{1,\cdots,1}_{2K},\underbrace{(-\hat{\mathbf{m}}_1),(-\hat{\mathbf{m}}_2),\cdots,(-\hat{\mathbf{m}}_{2K})}_{2K},\underbrace{1,\cdots,1}_{2N_t},\underbrace{-1,\cdots,-1}_{2N_t},\underbrace{\frac{1}{\sqrt{2N_t}},\cdots,\frac{1}{\sqrt{2N_t}}}_{4N_t}\Bigg]^T \tag{97}$$

$$\mathrm{C_N} = (6N_t+2K+1)^{1.5}(2N_t+1)^2 \cdot \ln\left(\frac{8N_t^2+20N_t+4KN_t+2\sqrt{2N_t}+8K+\|\mathbf{M}\|_1+6+\varepsilon^2}{\varepsilon}\right) \tag{100}$$

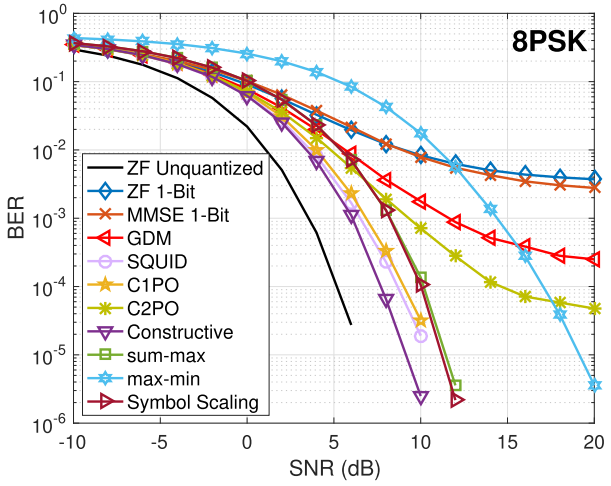

Fig. 6. BER vs. transmit SNR, $N_t = 128$, $K = 8$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{SQUID}} = 50$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C1PO}} = 20$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C2PO}} = 20$, 8PSK.

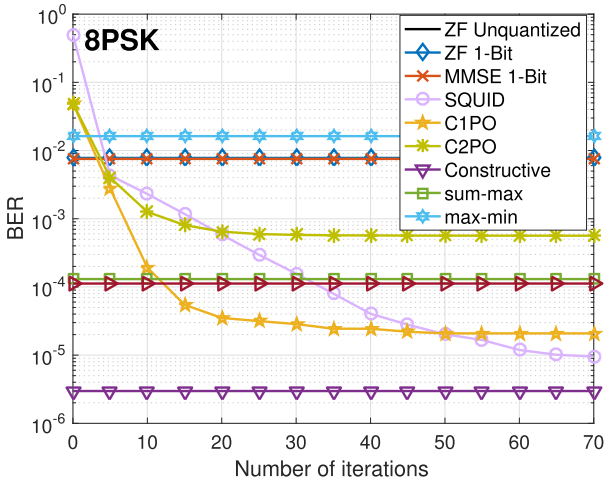

Fig. 8. BER vs. iteration number $\mathrm{N}^{\mathrm{Iter}}$, $N_t = 128$, $K = 8$, $\rho = 10$dB, 8PSK.

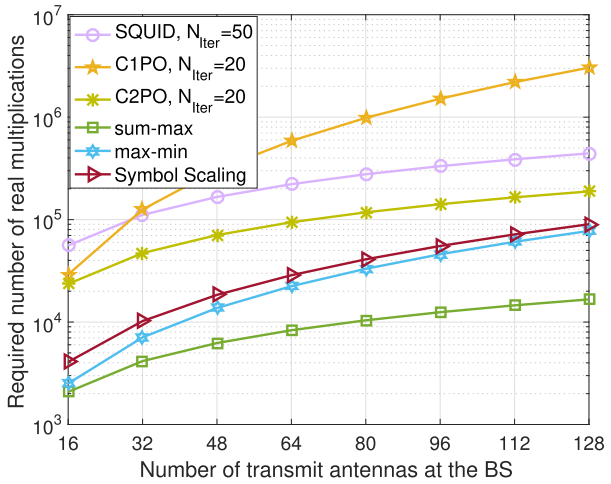

Fig. 7. Analytical computational cost comparison, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{SQUID}} = 50$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C1PO}} = 20$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C2PO}} = 20$, $K = 8$.

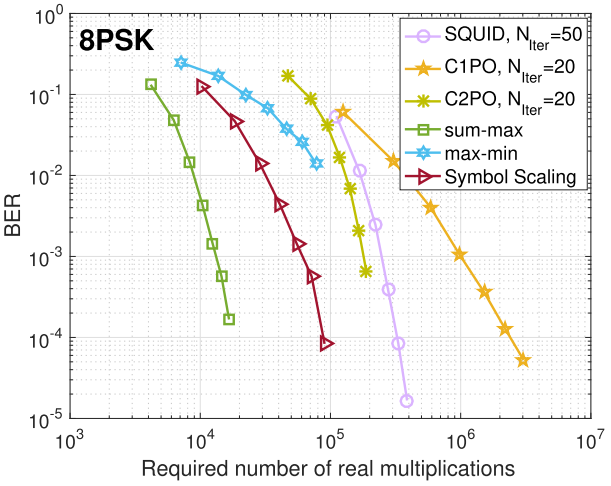

Fig. 9. BER vs. analytical computational costs, $K = 8$, $\rho = 10$dB, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{SQUID}} = 50$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C1PO}} = 20$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C2PO}} = 20$, 8PSK.

Moreover, while we observe around a 2dB SNR loss for the 'Symbol Scaling' approach compared to some existing 1-bit precoding algorithms in the case of massive MIMO, the corresponding computational cost is also greatly reduced in this scenario, which is shown in Fig. 7 in the following.

In Fig. 6, we show the performance of different schemes with $N_t = 128$ and $K = 8$ for 8PSK modulation. 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, we observe that in the case of 8PSK, only a 1dB SNR loss is observed compared to the non-linear iterative 'C1PO' algorithm. Moreover, when 8PSK is considered, our proposed 'Symbol Scaling' algorithm based on CI is more favorable in terms of BER compared to the low-complexity 'C2PO' scheme in [29].

In Fig. 7, we compare the computational complexity of each approach in terms of the required number of real multiplications. It is observed that the computational cost of the proposed symbol-scaling method based on sum-max requires the lowest computational cost, while the number of operations required for the proposed symbol scaling approach is much smaller than those for the existing 1-bit precoding schemes. The complexity gains of the proposed symbol scaling approach therefore favor its practical application, especially for the 'sum-max' approach.

To further compare the proposed schemes with existing 1-bit iterative precoding schemes, in Fig. 8 we present the BER performance with different number of iterations. 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 these iterative schemes improves as $\mathrm{N}^{\mathrm{Iter}}$ increases. Nevertheless, we note that the improvement becomes less significant with a larger $\mathrm{N}^{\mathrm{Iter}}$, where 'C1PO', 'C2PO' and 'SQUID' achieve their best performance at $\mathrm{N}^{\mathrm{Iter}} = 25$, $\mathrm{N}^{\mathrm{Iter}} = 25$ and $\mathrm{N}^{\mathrm{Iter}} = 55$.

To demonstrate the performance-complexity tradeoff directly, in Fig. 9 we depict the BER with respect to the

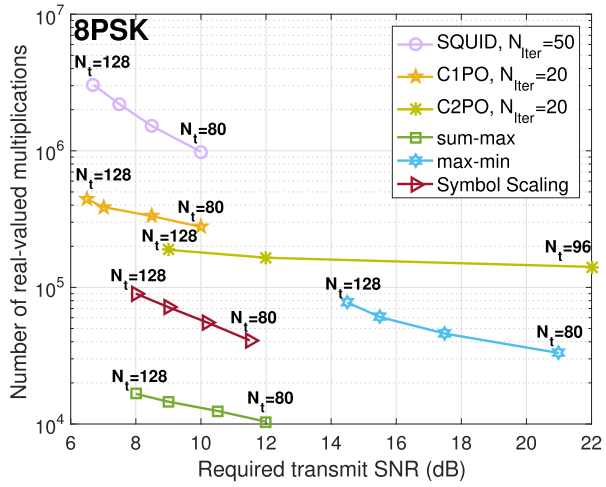

Fig. 10. Analytical computational costs vs. required transmit SNR for BER target 10^{-3}, $K = 8$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{SQUID}} = 50$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C1PO}} = 20$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C2PO}} = 20$, 8PSK.

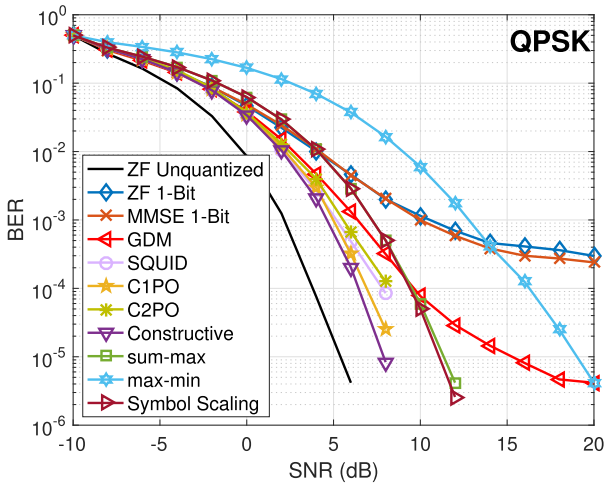

Fig. 11. BER vs. transmit SNR, $N_t = 128$, $K = 16$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{SQUID}} = 50$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C1PO}} = 20$, $\mathrm{N}^{\mathrm{Iter}}_{\mathrm{C2PO}} = 20$, QPSK, Imperfect CSI, $\beta = 0.1$.

required number of real operations for a range of transmit antennas from $N_t = 32$ to $N_t = 128$, where the number of users is fixed as $K = 8$. Fig. 10 presents the complexity with respect to the required SNR to achieve a given 10^{-3} target BER for a range of transmit antennas. From both figures, generally we observe that the proposed CI-based 1-bit precoding methods required a reduced computational cost. In terms of the performance and complexity tradeoff, while the 'SQUID', 'C1PO', 'C2PO' and the proposed 'Symbol Scaling' methods achieve different tradeoffs, we observe that 'SQUID' is superior to 'C1PO' and 'Symbol Scaling' is more favorable compared to 'C2PO'.

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. Channel estimation techniques for massive MIMO with 1-bit quantization is an ongoing topic of research [19], [47], 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

$$\hat{\mathbf{H}} = \sqrt{1-\delta} \cdot \mathbf{H} + \sqrt{\delta} \cdot \mathbf{Q}. \tag{101}$$

In (101), $\hat{\mathbf{H}}$ is the obtained CSI at the BS, and each entry in $\mathbf{Q}$ follows $\mathcal{CN}(0,1)$. The variance of the channel error is denoted by δ, and is modeled as inversely proportional to the transmit SNR: $\delta = \beta/\rho$, where β denotes the error coefficient [16], [31]. The BER result with imperfect CSI is depicted in Fig. 11, where we observe that the performance trend is similar to that seen in the earlier figures. 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.

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 quantized linear beamforming and non-linear mapping. With the analysis of the Lagrangian and KKT conditions, the quantized linear scheme is mathematically proven to be equivalent to quantized ZF beamforming. For the proposed non-linear mapping scheme, it is shown to be non-convex and solved by first 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 favors its usefulness in practical systems. Our future work is to consider precoding techniques for 1-bit massive MIMO transmission with QAM modulations based on interference exploitation.

Appendix
Coordinate Transformation

We employ 8PSK modulation in Fig. 3 as an example to demonstrate the coordinate transformation, where we focus on the constellation point $s_{(1)}$ in Fig. 3. Then, in the conventional real-imaginary complex plane, for node 'B' in Fig. 3, we obtain

$$\vec{OB} = \mathbf{h}_k \mathbf{x}_T = B_r \cdot 1 + B_i \cdot j, \tag{102}$$

where 1 and j are the bases for the real-imaginary plane, and we denote (B_r, B_i) as the corresponding coordinates. Accordingly, B_r and B_i are obtained as

$$\begin{aligned} B_r &= \Re\left(\mathbf{h}_k \mathbf{x}_T\right) = \mathbf{h}_k^{\Re}\mathbf{x}_T^{\Re} - \mathbf{h}_k^{\Im}\mathbf{x}_T^{\Im}, \\ B_i &= \Im\left(\mathbf{h}_k \mathbf{x}_T\right) = \mathbf{h}_k^{\Im}\mathbf{x}_T^{\Re} + \mathbf{h}_k^{\Re}\mathbf{x}_T^{\Im}. \end{aligned} \tag{103}$$

In the plane expanded by the two detection thresholds that correspond to the constellation point $s_{(1)}$, following (45) $\vec{OB}$ is decomposed into

$$\vec{OB} = \mathbf{h}_k \mathbf{x}_T = \alpha_k^{\mathcal{A}} s_k^{\mathcal{A}} + \alpha_k^{\mathcal{B}} s_k^{\mathcal{B}}. \tag{104}$$

Based on (41) and the fact that $\alpha_k^{\mathcal{A}}$ and $\alpha_k^{\mathcal{B}}$ are real numbers, (104) is further transformed into

$$\begin{aligned} \mathbf{h}_k \mathbf{x}_T &= \alpha_k^{\mathcal{A}} \left(A_k^{\Re} + A_k^{\Im} \cdot j \right) + \alpha_k^{\mathcal{B}} \left(B_k^{\Re} + B_k^{\Im} \cdot j \right) \\ &= \left(A_k^{\Re} \alpha_k^{\mathcal{A}} + B_k^{\Re} \alpha_k^{\mathcal{B}} \right) + \left(A_k^{\Im} \alpha_k^{\mathcal{A}} + B_k^{\Im} \alpha_k^{\mathcal{B}} \right) \cdot j. \end{aligned} \tag{105}$$

By revisiting (103), we obtain

$$\begin{aligned} B_r &= \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} - \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im} = A_k^{\Re} \alpha_k^{\mathcal{A}} + B_k^{\Re} \alpha_k^{\mathcal{B}}, \\ B_i &= \Im\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Re} + \Re\left(\mathbf{h}_k\right) \mathbf{x}_T^{\Im} = A_k^{\Im} \alpha_k^{\mathcal{A}} + B_k^{\Im} \alpha_k^{\mathcal{B}}, \end{aligned} \tag{106}$$

which leads to

$$\begin{aligned} \alpha_k^{\mathcal{A}} &= \frac{B_k^{\Im} B_r - B_k^{\Re} B_i}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \\ &= \frac{B_k^{\Im} \left[\mathbf{h}_k^{\Re} \mathbf{x}_T^{\Re} - \mathbf{h}_k^{\Im} \mathbf{x}_T^{\Im} \right] - B_k^{\Re} \left[\mathbf{h}_k^{\Im} \mathbf{x}_T^{\Re} + \mathbf{h}_k^{\Re} \mathbf{x}_T^{\Im} \right]}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \\ &= \frac{B_k^{\Im} \mathbf{h}_k^{\Re} - B_k^{\Re} \mathbf{h}_k^{\Im}}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \mathbf{x}_T^{\Re} - \frac{B_k^{\Im} \mathbf{h}_k^{\Im} + B_k^{\Re} \mathbf{h}_k^{\Re}}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \mathbf{x}_T^{\Im}, \end{aligned} \tag{107}$$

and

$$\begin{aligned} \alpha_k^{\mathcal{B}} &= \frac{A_k^{\Re} B_i - A_k^{\Im} B_r}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \\ &= \frac{A_k^{\Re} \left[\mathbf{h}_k^{\Im} \mathbf{x}_T^{\Re} + \mathbf{h}_k^{\Re} \mathbf{x}_T^{\Im} \right] - A_k^{\Im} \left[\mathbf{h}_k^{\Re} \mathbf{x}_T^{\Re} - \mathbf{h}_k^{\Im} \mathbf{x}_T^{\Im} \right]}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \\ &= \frac{A_k^{\Re} \mathbf{h}_k^{\Im} - A_k^{\Im} \mathbf{h}_k^{\Re}}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \mathbf{x}_T^{\Re} + \frac{A_k^{\Re} \mathbf{h}_k^{\Re} + A_k^{\Im} \mathbf{h}_k^{\Im}}{A_k^{\Re} B_k^{\Im} - A_k^{\Im} B_k^{\Re}} \mathbf{x}_T^{\Im}. \end{aligned} \tag{108}$$

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

REFERENCES

[1] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[2] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Large-scale MIMO transmitters in fixed physical spaces: The effect of transmit correlation and mutual coupling," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2794–2804, Jul. 2013.

[3] S. Biswas, C. Masouros, and T. Ratnarajah, "Performance analysis of large multi-user MIMO systems with space-constrained 2D antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3492–3505, May 2016.

[4] C. Masouros and M. Matthaiou, "Space-constrained massive MIMO: Hitting the wall of favorable propagation," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 771–774, May 2015.

[5] A. J. Garcia-Rodriguez and C. Masouros, "Exploiting the increasing correlation of space constrained massive MIMO for CSI relaxation," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1572–1587, Apr. 2016.

[6] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, "Performance of MIMO systems with channel inversion," in *Proc. IEEE 55th Veh. Technol. Conf. Veh. Technol. (VTC Spring)*, vol. 1, May 2002, pp. 35–39.

[7] 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.

[8] S. Han, C.-L. I, Z. Xu, 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.

[9] A. F. Molisch *et al.*, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[10] A. Li and C. Masouros, "Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 438–441, Feb. 2017.

[11] A. Li and C. Masouros, "Energy-efficient SWIPT: From fully digital to hybrid analog–digital beamforming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3390–3405, Apr. 2018, doi: 10.1109/TVT.2017.2782775.

[12] X. Xue, Y. Wang, L. Dai, and C. Masouros, "Relay hybrid precoding design in millimeter-wave massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 8, pp. 2011–2026, Apr. 2018, doi: 10.1109/TSP.2018.2799201

[13] A. Garcia-Rodriguez, V. Venkateswaran, P. Rulikowski, and C. Masouros, "Hybrid analog-digital precoding revisited under realistic RF modeling," *IEEE Wireless Commun. Lett.*, vol. 5, no. 5, pp. 528–531, Oct. 2016.

[14] B. Razavi, *Principles of Data Conversion System Design*, 1st ed. Hoboken, NJ, USA: Wiley, 1994.

[15] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, 3rd ed. London, U.K.: Oxford Univ. Press, 2016.

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

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

[18] 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, Jun. 2017.

[19] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[20] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sep. 2017.

[21] A. K. Saxena, I. Fijalkow, A. Mezghani, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, 2016, pp. 758–762.

[22] A. Mezghani, R. Ghiat, and J. A. Nossek, "Transmit processing with low resolution D/A-converters," in *Proc. 16th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Yasmine Hammamet, Tunisia, Dec. 2009, pp. 683–686.

[23] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE precoder for massive MIMO using 1-bit quantization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 3381–3385.

[24] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6483–6487.

[25] A. Noll, H. Jedda, and J. A. Nossek, "PSK precoding in multi-user MISO systems," in *Proc. 21st Int. ITG Workshop Smart Antennas (WSA)*, Berlin, Germany, 2017, pp. 1–7.

[26] H. Jedda, J. A. Nossek, and A. Mezghani, "Minimum BER precoding in 1-bit massive MIMO systems," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Rio de Janerio, Brazil, Jul. 2016, pp. 1–5.

[27] L. T. N. Landau and R. C. de Lamare, "Branch-and-bound precoding for multiuser MIMO systems with 1-bit quantization," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 770–773, Dec. 2017.

[28] O. Castañeda, T. Goldstein, and C. Studer, "POKEMON: A non-linear beamforming algorithm for 1-bit massive MIMO," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 3464–3468.

[29] O. Castañeda, S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "1-bit massive MU-MIMO precoding in VLSI," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 4, pp. 508–522, Dec. 2017.

[30] C. Masouros, "Correlation rotation linear precoding for MIMO broadcast communications," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 252–262, Jan. 2011.

[31] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Vector perturbation based on symbol scaling for limited feedback MISO downlinks," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 562–571, Feb. 2014.
[32] C. Masouros and G. Zheng, "Exploiting known interference as green signal power for downlink beamforming optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3628–3640, Jul. 2015.
[33] 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.
[34] C. Masouros, T. Ratnarajah, M. Sellathurai, C. Papadias, and A. 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.
[35] M. Alodeh, S. Chatzinotas, and B. Ottersten, "Constructive multiuser interference in symbol level precoding for the MISO downlink channel," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2239–2252, May 2015.
[36] H. Jedda, A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, "Massive MIMO downlink 1-bit precoding with linear programming for PSK signaling," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–5.
[37] A. Li and C. Masouros, "Exploiting constructive mutual coupling in P2P MIMO by analog-digital phase alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1948–1962, Mar. 2017.
[38] S. Timotheou, G. Zheng, C. Masouros, and I. Krikidis, "Exploiting constructive interference for simultaneous wireless information and power transfer in multiuser downlink systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1772–1784, May 2016.
[39] P. V. Amadori and C. Masouros, "Large scale antenna selection and precoding for interference exploitation," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4529–4542, Oct. 2017.
[40] F. Liu, C. Masouros, P. V. Amadori, and H. Sun, "An efficient manifold algorithm for constructive interference based constant envelope precoding," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1542–1546, Oct. 2017.
[41] P. V. Amadori and C. Masouros, "Interference-driven antenna selection for massive multiuser MIMO," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 5944–5958, Aug. 2016.
[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[43] R. Hunger, "Floating point operations in matrix-vector calculus," Dept. Inst. Circuit Theory Signal Process., Tech. Univ. Munich, München, Germany, Tech. Rep., 2005. [Online]. Available: https://mediatum.ub.tum.de/doc/625604
[44] S. Jacobsson, O. Castañeda, C. Jeon, G. Durisi, and C. Studer. (2018). "Nonlinear precoding for phase-quantized constant-envelope massive MU-MIMO-OFDM." [Online]. Available: https://arxiv.org/abs/1710.06825
[45] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: SIAM, 2001.
[46] M. Wu, B. Yin, K. Li, C. Dick, J. R. Cavallaro, and C. Studer, "Implicit vs. explicit approximate matrix inversion for wideband massive MU-MIMO data detection," *J. Signal Process. Syst.*, vol. 90, no. 10, pp. 1311–1328, Dec. 2017.
[47] C. Stockle, J. Munir, A. Mezghani, and J. A. Nossek, "Channel estimation in massive MIMO systems using 1-bit quantization," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–6.

Ang Li (S'14–M'18) received the Ph.D. degree from the Communications and Information Systems Research Group, Department of Electrical and Electronic Engineering, University College London, in 2018. He is currently a Post-Doctoral Research Associate with the School of Electrical and Information Engineering, The University of Sydney. His research interests lie in the fields of beamforming and signal processing techniques for multiple-input multiple-output systems.

Christos Masouros (M'06–SM'14) received the Diploma degree in electrical and computer engineering from the University of Patras, Greece, in 2004, and the M.Sc. by Research and Ph.D. degrees in electrical and electronic engineering from The University of Manchester, U.K., in 2006 and 2009, respectively. In 2008, he joined the Philips Research Labs, U.K., as a Research Intern. Between 2009 and 2010, he was a Research Associate with The University of Manchester. Between 2010 and 2012, he was a Research Fellow with Queen's University Belfast. He was a recipient of the Royal Academy of Engineering Research Fellowship during 2011–2016.

He is currently a Senior Lecturer with the Communications and Information Systems Research Group, Department of Electrical and Electronic Engineering, University College London. His research interests lie in the fields of wireless communications and signal processing with particular focus on green communications, large-scale antenna systems, cognitive radio, interference mitigation techniques for multiple-input multiple-output, and multicarrier communications. He received the Best Paper Award at the IEEE GlobeCom Conference 2015. He is recognized as an Exemplary Editor for the IEEE COMMUNICATIONS LETTERS and as an Exemplary Reviewer for the IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently an Associate Editor of the IEEE COMMUNICATIONS LETTERS, a Guest Editor of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING Issue on Exploiting Interference Towards Energy Efficient and Secure Wireless Communications, and an Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS.

Fan Liu (S'16) received the bachelor's degree in information engineering and the Ph.D. degree in electronic science and technology from the Beijing Institute of Technology, Beijing, China, in 2013 and 2018, respectively. From 2016 to 2018, he was a visiting student with the Communications and Information Systems Research Group, Department of Electrical and Electronic Engineering, University College London, London, U.K. His research interests include precoding designs for multiple-input multiple-output systems, signal detection and estimation, and convex optimization. He was a recipient of the Marie Curie Individual Fellowship in 2018. He is recognized as an Exemplary Reviewer for the IEEE TRANSACTIONS ON COMMUNICATIONS.

A. Lee Swindlehurst (S'83–M'84–SM'99–F'04) received the B.S. and M.S. degrees in electrical engineering from Brigham Young University (BYU), Provo, UT, USA, in 1985 and 1986, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1991. From 1990 to 2007, he was with the Department of Electrical and Computer Engineering, BYU, where he was the Department Chair from 2003 to 2006. During 1996–1997, he held a joint appointment as a Visiting Scholar with Uppsala University and also with the Royal Institute of Technology, Sweden. From 2006 to 2007, he was on leave as a Vice President of Research with ArrayComm LLC, San Jose, CA, USA. From 2013 to 2016, he was an Associate Dean for Research and Graduate Studies with the Samueli School of Engineering, University of California at Irvine, Irvine, CA, USA. During 2014–2017, he was a Hans Fischer Senior Fellow with the Institute for Advanced Studies, Technical University of Munich. Since 2007, he has been a Professor with the Electrical Engineering and Computer Science Department, University of California at Irvine. His research focuses on array signal processing for radar, wireless communications, and biomedical applications. He has authored over 300 publications in these areas. He was a recipient of the 2000 IEEE W. R. G. Baker Prize Paper Award, the 2006 IEEE Communications Society Stephen O. Rice Prize in the Field of Communication Theory, the 2006 and 2010 IEEE Signal Processing Society's Best Paper Awards, and the 2017 IEEE Signal Processing Society Donald G. Fink Overview Paper Award. He was the inaugural Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.