Chapter

# Spatial Multiplexing for MIMO/Massive MIMO

*Haonan Wang and Ang Li*

## Abstract

In this chapter, we will discuss how to achieve spatial multiplexing in multiple-input multiple-output (MIMO) communications through precoding design, for both traditional small-scale MIMO systems and massive MIMO systems. The mathematical description for MIMO communications will first be introduced, based on which we discuss both block-level precoding and the emerging symbol-level precoding techniques. We begin with simple and closed-form block-level precoders such as maximum ratio transmission (MRT), zero-forcing (ZF), and regularized ZF (RZF), followed by the classic symbol-level precoding schemes such as Tomlinson-Harashima precoder (THP) and vector perturbation (VP) precoder. Subsequently, we introduce optimization-based precoding solutions, including power minimization, SINR balancing, symbol-level interference exploitation, etc. We extend our discussion to massive MIMO systems and particularly focus on precoding designs for hardware-efficient massive MIMO systems, such as hybrid analog-digital precoding, low-bit precoding, nonlinearity-aware precoding, etc.

**Keywords:** MIMO, massive MIMO, spatial multiplexing, precoding, beamforming

## 1. Introduction

In recent years, the demand for high-speed wireless communication has grown exponentially, driven by the proliferation of smart devices, the Internet of Things (IoT), and the increasing need for reliable and efficient data transmission [1]. To meet these demands, multiple-input multiple-output (MIMO) technology has emerged as a promising solution, offering significant improvements in spectral efficiency, capacity, and reliability. In this chapter, we will explore the concept of spatial multiplexing in MIMO communications, focusing on precoding design for both traditional small-scale MIMO systems and massive MIMO systems.

MIMO communication systems employ multiple antennas at both the transmitter and receiver ends to exploit the spatial domain, enabling the simultaneous transmission of multiple data streams over the same frequency band [2]. This spatial multiplexing capability is the key factor in achieving the high data rates and improved link reliability that MIMO systems offer. Precoding is a crucial technique in MIMO communications, as it allows the transmitter to pre-process the signals before

transmission, effectively mitigating inter-stream interference and optimizing the received signal quality. We will begin our discussion with a mathematical description of MIMO communications, providing a solid foundation for understanding the principles and techniques involved in precoding design. Based on this mathematical framework, we will dive deep into both block-level precoding and the emerging symbol-level precoding technique.

Block-level precoding techniques, such as maximum ratio transmission (MRT), zero-forcing (ZF), and regularized ZF (RZF), offer simple and closed-form solutions for mitigating inter-stream interference. These methods have been widely adopted in small-scale MIMO systems due to their ease of implementation and relatively low computational complexity. We will also discuss classic symbol-level precoding schemes, including the Tomlinson-Harashima precoder (THP) and vector perturbation (VP) precoder, which offer improved performance by exploiting the inherent structure of the transmitted symbols. As we move beyond these basic precoding techniques, we will introduce optimization-based precoding solutions that aim to further enhance the performance of MIMO systems. These approaches include power minimization, SINR balancing, and symbol-level interference exploitation, among others. By optimizing various performance metrics, these advanced precoding techniques can achieve significant gains in spectral efficiency and link reliability.

In the latter part of the chapter, we will extend our discussion to massive MIMO systems, which employ a large number of antennas at the transmitter and receiver to achieve even greater spatial multiplexing gains. While the basic principles of precoding design remain applicable to massive MIMO systems, the increased scale and complexity of these systems introduce new challenges and opportunities for precoding optimization. In particular, we will focus on precoding designs for hardware-efficient massive MIMO systems, such as hybrid analog-digital precoding, low-bit precoding, and nonlinearity-aware precoding. These techniques aim to address the practical limitations of massive MIMO systems, including hardware constraints, power consumption, and implementation complexity, while still achieving desired performance gains.

In conclusion, this chapter will provide a comprehensive overview of spatial multiplexing in MIMO communications, with a focus on precoding design for both small-scale and massive MIMO systems. By exploring a wide range of precoding techniques, from simple closed-form solutions to advanced optimization-based approaches, we aim to offer the reader a deep understanding of the principles and methods involved in achieving high-performance MIMO communications.

## 2. Body of the manuscript

In Section 3, we will provide an introduction to the MIMO communication system, which will include a mathematical description of the MIMO system, performance metrics of MIMO communications, and emerging massive MIMO techniques. In Section 4, we will explain traditional precoding design, which will include preliminaries on precoding and classical precoding schemes. Subsequently, in Section 5, we will discuss optimization-based precoding to demonstrate the use of convex optimization in precoding design. Finally, in recognition of the wide application of massive MIMO, Section 6 will introduce hardware-efficient precoding as a means of achieving a favorable balance between communication performance and power consumption.

*DOI: http://dx.doi.org/10.5772/intechopen.112041*

## 3. MIMO communication systems

Due to the increasing demand for higher data rates and reliability for wireless networks, MIMO techniques have appeared and received extensive research attention. To support spatial multiplexing, parallel data streams can be transmitted simultaneously with multiple antennas deployed at the BS. To improve reliability, space-time coding techniques can be employed by sending copies of the same information across the antenna array. In this section, we present an overview of the fundamental concepts of multi-antenna technology, which serves as a foundation for the subsequent discussion on precoding. Given that spatial multiplexing is the primary focus of this chapter, our attention is primarily directed toward multi-user multi-input single-output (MU-MISO) systems.

### 3.1 Mathematical description for MIMO communications

In a wireless multi-user MISO (MU-MISO) system, as depicted in **Figure 1**, the data symbol vector is denoted as $\mathbf{s}$, and one BS with $N_t$ antennas transmits wireless signals to $K$ single-antenna receivers. Mathematically, the signal vector at the receiver can be expressed as.where $h_{i,j}$ denotes the complex channel gain between the $i$-th receiver and the $j$-th transmit antenna, $x_j$ denotes the transmit signal on the $j$-th transmit antenna, $y_i$ denotes the received signal of the $j$-th receiver, and $n_i$ denotes the additive Gaussian noise corresponding to the $i$-th receiver. Based on that, the $k$-th user's received signal can be expressed as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N_t} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N_t} \\ \vdots & \vdots & \ddots & \vdots \\ h_{K,1} & h_{K,2} & \cdots & h_{K,N_t} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N_t} \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_K \end{bmatrix}, \tag{1}$$

$$y_k = \mathbf{h}_k^{\mathrm{T}}\mathbf{x} + n_k, \tag{2}$$

where $y_k$ denotes the $k$-th user's received signal, $\mathbf{h}_k \in \mathbb{C}_{N_t \times 1}$ denotes the $k$-th user's channel vector, $\mathbf{x} \in \mathbb{C}_{N_t \times 1}$ denotes the transmit signal vector, and $n_k$ denotes the additive noise vector which follows the complex Gaussian distribution $\mathbb{CN}\left(0, \sigma_k^2\mathbf{I}\right)$ with the zero mean and $\sigma_k^2$ noise power. The combining process is eliminated at the receiver side, for the single-antenna configuration. Based on (2), the transmission process in MU-MISO can be reorganized into a matrix form, as shown below:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{3}$$

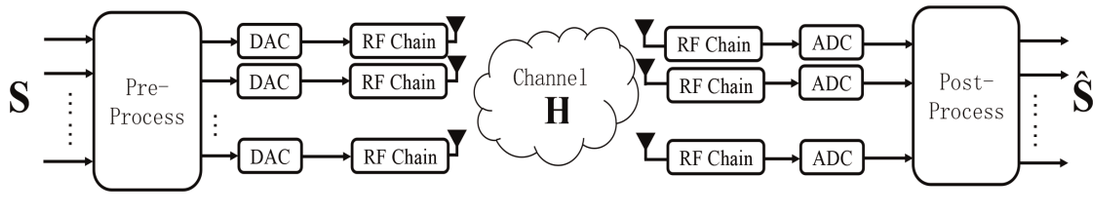


**Figure 1.**
*A block diagram of MU-MISO systems.*

with $\mathbf{y} = [y_1, y_2, \ldots, y_K]^{\mathrm{T}}$, $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]^{\mathrm{T}}$, and $\mathbf{n} = [n_1, n_2, \ldots, n_K]^{\mathrm{T}}$.

To mitigate the detrimental impact of channel fading, the transmitter performs precoding on the symbol vector to obtain the transmitted signal, expressed as $\mathbf{x} = \mathbf{W}\mathbf{s}$. Precoding is achieved using a matrix $\mathbf{W} \in \mathbb{C}_{N_t \times K}$. The design of the precoding matrix $\mathbf{W}$ is the crucial signal processing procedure in MIMO downlink transmission, as it enables each receiver to achieve a received signal $y_k$ that closely approximates the original symbol $s_k$.

## 3.2 Performance metrics for MIMO communications

In order to measure the communication performance of MIMO systems, bit error rate (BER) and channel capacity are the two performance metrics that are usually employed, as explained below.

### *3.2.1 BER*

Bit Error Rate (BER) refers to the proportion of erroneously transmitted bits to the total number of transmitted bits during the transmission process and is the most commonly used performance metric to evaluate the reliability of digital communication systems. Its mathematical definition can be given as

$$P_b = \frac{N_e}{N_b}, \tag{4}$$

where $N_e$ denotes the erroneous transmitted bits, and $N_b$ denotes the total transmitted bits.

### *3.2.2 Channel capacity*

The channel capacity represents the maximum rate of information transmission that can be sustained by a communication system when the bit error rate approaches zero. Its mathematical definition is given as the maximum mutual information between the input and output signals of the channel, which represents the extent to which the received signal preserves information about the transmitted signal after the channel. More specifically, the channel capacity is determined by identifying the input distribution that maximizes the mutual information, subject to the constraints of the channel's physical properties and the power limitations of the system. Therefore, it serves as a fundamental limit on the data transmission rate and is a crucial performance metric for evaluating the effectiveness of communication systems. The definition of channel capacity can be expressed as

$$\mathrm{C} = \max \mathrm{I}(\text{input}; \text{output}), \tag{5}$$

where C denotes the channel capacity, and I $(x; y)$ denotes the mutual information between $x$ and $y$. For SISO systems, when both the transmitter and receiver have perfect Channel State Information (CSI), the channel capacity can be obtained as

$$\mathrm{C} = B \log_2(1 + \gamma), \tag{6}$$

where $B$ denotes the system bandwidth, and $\gamma$ denotes the receive SNR. The physical interpretation of (8) has been discussed in ref. [2].

In the context of MIMO systems, it is feasible to decompose the channel into a sum of multiple SISO channels via singular value decomposition (SVD) [2]. Subsequently, utilizing "water-filling" power allocation strategy [2], it is possible to harness the full potential of the system and achieve channel capacity. In an ideal scenario where both the transmitter and receiver possess perfect CSI, the channel capacity of an $N_r \times N_t$ MIMO channel can be captured precisely using the following equation:

$$C = \log_2 \det\left(\mathbf{I}_{N_r} + \frac{\rho}{N_r}\mathbf{H}\mathbf{H}^{\mathrm{H}}\right), \tag{7}$$

where $\rho$ denotes the transmit SNR.

### 3.3 Massive MIMO

As mobile communication technologies continue to evolve, wireless network capacity and communication quality have become increasingly critical. Traditional wireless communication systems face limitations that prevent them from satisfying the modern industry's demands for high-speed, high-capacity, and high-quality communication. Massive MIMO technology has emerged as a promising solution to these challenges.

Massive MIMO is an extension of conventional MIMO technology [3, 4]. In contrast to the typical tens-of-antenna configuration in traditional MIMO systems for signal transmission and reception, Massive MIMO employs significantly more antennas, for example, hundreds or even thousands of antennas.

Massive MIMO technology enjoys wide applications in various fields of wireless communications, such as 5G and IoT [5]. It has several notable features: channel hardening, favorable propagation, power concentration, capacity enhancement, interference reduction, and spectral efficiency improvement. In particular, channel hardening refers to the property that as the antenna array size increases, the relative fluctuations of channel coefficients decrease [5]. Although randomness still exists, its impact on communication approximates that of non-fading channels. Favorable propagation is a phenomenon in which the channels of different users become nearly orthogonal in the spatial domain as the number of antennas at the base station increases significantly. This leads to a substantial reduction in inter-user interference and further improved spectral efficiency, making massive MIMO a promising technology for future wireless communication systems. Power concentration refers to Massive MIMO's ability to focus transmitted power more efficiently through finer beamforming techniques, especially for millimeter-wave communication where channel gain drops off precipitously with distance [6]. Capacity enhancement is achieved by processing more data streams than traditional MIMO systems, leading to improved network capacity. Interference reduction is accomplished through spatial multiplexing and beamforming, which minimize inter-signal interference and enhance signal quality and reliability. Last, spectral efficiency improvement results from more efficient utilization of bandwidth resources, which enhances data transmission speeds.

However, Massive MIMO technology still faces certain challenges in engineering applications, such as high power consumption [7] and hardware costs. To be more

specific, traditional MIMO systems equip each antenna with radio frequency (RF) chains and high-resolution digital-to-analog converters (DACs), causing significant power loss when the antenna array is large. In such a scenario, the advanced signal processing mechanisms required to handle a large number of antennas for signal transmission and reception are generally more complex, necessitating much more energy consumption than traditional wireless communication systems. From this perspective, hardware-efficient precoding techniques hold significant research value and promising application prospects.

## 4. Traditional precoding

In this section, we will introduce traditional precoding to discuss its working mechanism and design principle. Preliminaries will be first introduced, as the basis of further discussion. Based on that, we mainly introduce the linear block-level precoding schemes with closed-form solutions, including MRT, ZF, and RZF. After that, the traditional non-linear symbol-level precoding will be discussed, including THP and VP.

### 4.1 Preliminaries on precoding

First, we will introduce the preliminaries of the precoding process in the downlink MIMO system, as the basis of further discussion.

Without loss of generality, we mainly consider a downlink MU-MISO system, where $K$ single-antenna users are served by a common base station with $N_t$ transmit antennas at the same time. Considering that users are generally separated spatially, based on CSI, the BS needs to employ signal processing techniques before transmission such that the destructive effect of channel fading and inter-user interference can be eliminated as much as possible. This is the initial motivation for precoding. Mathematically, the precoding process can be expressed as

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k = \mathbf{W}\mathbf{s}, \tag{8}$$

where $\mathbf{w}_k \in \mathbb{C}_{N_t \times 1}$ denotes the $k$-th user's precoding vector and $s_k$ is the $k$-th user's data symbol, which is drawn from a specific modulation constellation. Based on that, with the general precoding matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \in \mathbb{C}_{N_t \times K}$ and date symbol vector $\mathbf{s} = [s_1, s_2, \ldots, s_K]^{\mathrm{T}} \in \mathbb{C}_{K \times 1}$, the received signal for the $k$-th user can be expressed as

$$y_k = \mathbf{h}_k^{\mathrm{T}}\mathbf{x} + n_k = \mathbf{h}_k^{\mathrm{T}}\mathbf{W}\mathbf{s} + n_k, \tag{9}$$

where $y_k$ is the received signal for the $k$-th user, $\mathbf{h}_k \in \mathbb{C}_{N_t \times 1}$ is the complex channel vector between the BS and the $k$-th user, and $n_k \sim \mathbb{CN}(0, \sigma^2)$ is the additive Gaussian noise with zero mean and $\sigma^2$ noise power. Based on that, the transmission process can be given as

$$\mathbf{y} = \mathbf{H}\mathbf{W}\mathbf{s} + \mathbf{n}, \tag{10}$$

where $\mathbf{y} \in \mathbb{C}_{K\times 1}$ denotes the received signal vector, $\mathbf{H} \in \mathbb{C}_{K\times N_t}$ denotes the channel matrix, and $\mathbf{n} \in \mathbb{C}_{K\times 1}$ denotes the additive noise vector.

In traditional communication systems, the presence of interference can significantly degrade the quality of the received signal. This is particularly true in multi-user systems, where signals for different users are superimposed over the spatial channel. In such scenarios, the transmitted signals from different users can interfere with each other, leading to reduced signal quality at the receiver.

The insight of precoding is to design the precoding matrix **W** such that the received signal **y** can approach the data symbol vector **s** as much as possible. In the following subsections, we will introduce linear closed-form block-level precoding, which is a classical type of precoding.

### 4.2 Linear closed-form precoding

The classical linear block-level precoding schemes have been widely used in practical engineering systems since they can ensure satisfactory communication performance with low computational complexity. In this subsection, we will mainly discuss the specific linear closed-form precoding, including MRT, ZF, and RZF, to show the principle of precoding design and the physical mechanism of the precoding effect.

Specifically, the precoding matrix of **MRT** can be given as [4].

$$\mathbf{W}_{\mathrm{MRT}} = \frac{1}{f_{\mathrm{MRT}}} \cdot \mathbf{H}^{\mathrm{H}} = \sqrt{\frac{P_0}{\mathrm{tr}\{\mathbf{H}\mathbf{H}^{\mathrm{H}}\}}}\mathbf{H}^{\mathrm{H}}, \tag{11}$$

where $f_{\mathrm{MRT}} = \sqrt{\frac{\mathrm{tr}\{\mathbf{H}\mathbf{H}^{\mathrm{H}}\}}{P_0}}$ denotes the normalization factor to ensure the satisfaction of the transmit power constraint, and $P_0$ denotes the total transmit power. Considering that MRT can maximize the signal gain at the intended user, its performance is promising in noise-limited scenarios (low SNR regimes or large-scale MIMO scenarios), while its performance is limited in interference-limited scenarios.

**Zero-Forcing (ZF)** precoding is another classical precoding method that has been extensively used in practical applications [8]. By employing a Moore-Penrose inverse of the channel matrix **H** as the precoding matrix, ZF precoding can create an ideal environment where each user's effective channel is orthogonal with each other. Based on that, inter-user interference can be eliminated as much as possible. The ZF precoding matrix can be expressed as

$$\mathbf{W}_{\mathrm{ZF}} = \frac{1}{f_{\mathrm{ZF}}} \cdot \mathbf{H}^{\mathrm{H}}\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1} = \sqrt{\frac{P_0}{\mathrm{tr}\left\{\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}\right\}}}\mathbf{H}^{\mathrm{H}}\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}, N_t \geq K, \tag{12}$$

where $f_{\mathrm{ZF}} = \sqrt{\frac{\mathrm{tr}\left\{\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}\right\}}{P_0}}$ denotes the normalization factor for ZF precoding. ZF precoding is shown to achieve improved performance over MRT in the high SNR regime. The main idea of ZF precoding is to create orthogonal effective channels among all the users to fully eliminate inter-user interference. For its low computational complexity, ZF precoding has been widely used in practical engineering systems. However, the noise amplification effect limits its performance, especially in low SNR regions, which has been improved by RZF precoding.

By introducing a regularization factor to handle the noise amplification effect, the **RZF** precoding can further improve the performance of ZF precoding [9]. The RZF precoding matrix can be given by

$$\begin{aligned}\mathbf{W}_{\mathrm{RZF}} &= \frac{1}{f_{\mathrm{RZF}}}\cdot\mathbf{H}^{H}\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1}\\ &= \sqrt{\frac{P_0}{\mathrm{tr}\left\{\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1}\mathbf{H}\mathbf{H}^{H}\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1}\right\}}}\mathbf{H}^{H}\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1},\end{aligned} \tag{13}$$

where $f_{\mathrm{RZF}}=\sqrt{\frac{\mathrm{tr}\left\{\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1}\mathbf{H}\mathbf{H}^{H}\left(\mathbf{H}\mathbf{H}^{H}+\alpha\cdot\mathbf{I}\right)^{-1}\right\}}{P_0}}$ denotes the normalization factor for RZF precoding, and $\alpha$ denotes the regularization factor whose optimal value is $\alpha^{*}=K\sigma^{2}$.

### 4.3 Non-linear symbol-level precoding

Compared with linear precoding, non-linear precoding can achieve better performance by employing more sophisticated precoding techniques, at the cost of relatively high computational complexity. Generally speaking, based on CSI and the data symbol, non-linear precoding manipulates signal at the symbol level, which leads to a better communication performance but higher processing complexity. The transmitted signal of non-linear precoding is no longer a linearly weighted combination of symbol vectors. In this subsection, we will introduce classical non-linear precoding schemes to show their working mechanism.

**Dirty Paper Coding (DPC)** is able to reduce the destructive effect of inter-user interference and further achieve channel capacity in MIMO systems [10]. However, assuming perfect CSI and that interference information can be obtained at the transmitter, the capacity-achieving DPC requires an infinite-length coding and a high-complexity searching algorithm, which limits its application in practical systems.

Considering the high complexity of DPC, **Tomlinson-Harashima Precoding (THP)** has been proposed as an alternating near-capacity scheme whose computational complexity is relatively acceptable in practice. The basic idea of THP is to predistort the symbols before they are transmitted over the communication channel [11]. This pre-distortion is achieved by adding a feedback loop to the transmitting system, which modifies the symbols based on the previous symbols that have been transmitted. The feedback loop effectively cancels out the distortion introduced by the communication channel, leading to a higher quality and more reliable signal at the receiver. **Figure 2** shows the architecture of the THP precoding system.

Specifically, THP first decomposes the channel matrix into

$$\mathbf{H}=\mathbf{L}\mathbf{F}^{\mathrm{H}}, \tag{14}$$

with a lower-triangle matrix **L** and a unitary matrix **F**. Based on that, the transmitted signal vector **x** for THP can be further expressed as

$$\mathbf{x}_{\mathrm{THP}}=\mathbf{F}\tilde{\mathbf{x}}_{\mathrm{THP}}, \tag{15}$$

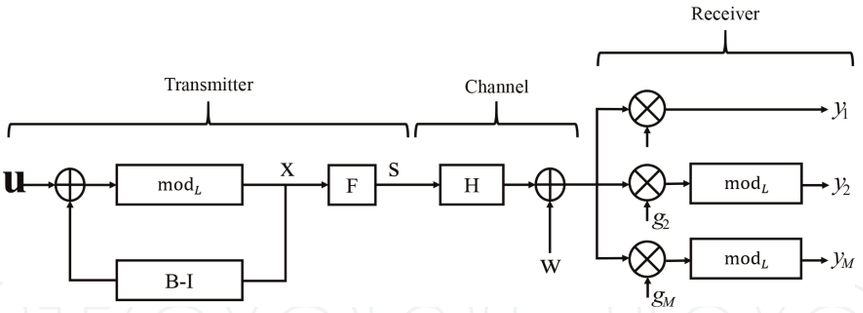


**Figure 2.**
*The geometrical representation of THP.*

where $\tilde{\mathbf{x}}$ can be obtained by

$$[\tilde{\mathbf{x}}_{\mathrm{THP}}]_k = \mathrm{mod}_\tau\left\{s_k - \sum_{l=1}^{k-1}[\mathbf{B}]_{k,l}\,[\tilde{\mathbf{x}}_{\mathrm{THP}}]_l\right\}, \forall k \in \{1,2,\cdots,K\}. \tag{16}$$

$\mathrm{mod}_\tau\{x\}$ denotes a complex modulo function, given by

$$\mathrm{mod}_\tau\{x\} = \left(\Re(x) - \tau\cdot\lfloor\frac{\Re(x)+\tau/2}{\tau}\rfloor\right) + j\left(\Im(x) - \tau\cdot\lfloor\frac{\Im(x)+\tau/2}{\tau}\rfloor\right), \tag{17}$$

where $\tau$ denotes the modulo basis and $\lfloor\cdot\rfloor$ denotes the floor approximating function. Based on the analysis above, the effective THP channel can be expressed as

$$\mathbf{B} = \mathbf{GHF}, \tag{18}$$

where **G** is a diagonal matrix that contains the complex scaling gain corresponding to each user, which is actually the inverse of the corresponding diagonal entry in **L**, i.e.,

$$g_k = [\mathbf{G}]_{k,k} = \frac{1}{[\mathbf{L}]_{k,k}}. \tag{19}$$

At the receiver side, the scaling compensation operation and the modulo operation are also required prior to the demodulation.

Considering that the performance of ZF precoding is mainly limited by its noise amplification effect, the **Vector- Perturbation (VP)** precoding [12] has been proposed as an improvement [12]. Based on the ZF precoding, VP precoding introduces a perturbation vector to the symbol vector, resulting in a transmitted signal that aligns better with the main eigenvector direction of the channel inverse matrix. This reduces the noise amplification factor and further lowers the noise amplification effect of ZF. Therefore, compared to ZF, VP can achieve significant performance gains. To be more specific, the VP precoding process can be expressed as

$$\mathbf{x}_{\mathrm{VP}} = \frac{1}{f_{\mathrm{VP}}}\cdot\mathbf{H}^{\mathrm{H}}\left(\mathbf{HH}^{\mathrm{H}}\right)^{-1}(\mathbf{s}+\tau\cdot\mathbf{l}), \tag{20}$$

where $\tau = 2|c|_{\max} + \Delta$ denotes the modulo basis corresponding to the modulation level, $|c|_{\max}$ denotes the modulus value of the maximum amplitude modulation

constellation point, and $\Delta$ is the minimum distance among the constellation points. $\mathbf{l} \in \mathbb{C}\mathbb{Z}^{K\times 1}$ denotes the complex integer perturbation vector, given as

$$\mathbf{l} = \underset{\mathbf{l}\in\mathbb{C}\mathbb{Z}^{K\times 1}}{\arg\min} \left\| \mathbf{H}^{\mathrm{H}}\left(\mathbf{H}\mathbf{H}^{\mathrm{H}}\right)^{-1}(\mathbf{s}+\tau\cdot\mathbf{l})\right\|_2^2, \tag{21}$$

which can be obtained by the sphere decoder. Based on that, the normalization factor of VP precoding can be obtained by

$$y_k = \frac{1}{f_{\mathrm{VP}}}\cdot\mathbf{h}_k\mathbf{x}_{\mathrm{VP}} + n_k = \frac{1}{f_{\mathrm{VP}}}(s_k+\tau l_k) + n_k, \tag{22}$$

where $l_k$ denotes the $k$-th element of the perturbation vector $\mathbf{l}$. In order to eliminate the perturbation component $\tau l_k$ at the receiver side, the receiver needs to accomplish the module operation after the power compensation, as shown below:

$$\begin{aligned} r_k &= \mathrm{mod}_\tau\left\{f_{\mathrm{VP}}y_k\right\} \\ &= \mathrm{mod}\tau\left\{s_k+\tau l_k+f_{\mathrm{VP}n_k}\right\} \\ &= s_k + f_{\mathrm{VP}}\hat{n}_k, \end{aligned} \tag{23}$$

where $\hat{n}_k$ denotes the effective noise of the $k$-th user.

## 5. Optimization-based precoding

With the deepening of research on precoding technology, an increasing number of mathematical tools, such as convex optimization, have been introduced into the precoding design process to improve precoding performance as much as possible. In addition, optimization-based precoding can flexibly serve various communication targets, and therefore has a wide range of applications in practical engineering systems.

### 5.1 Block-level precoding

#### *5.1.1 Preliminary*

Based on the analysis above, due to the linear relationship between the transmitted signal vector $\mathbf{x}$, the symbol vector $\mathbf{s}$, and the precoding matrix $\mathbf{W}$, the transmitted signal $\mathbf{x}$ can be regarded as a linear weighted combination of the precoding matrix $\mathbf{W}$, where the weighting coefficients are given by the symbol vector $\mathbf{s}$. Therefore, the wireless transmission process of (7) and (8) can be reformulated in the following form:

$$y_k = \mathbf{h}_k\sum_{i=1}^{K}\mathbf{w}_i s_i + n_k = \mathbf{h}_k\mathbf{w}_k s_k + \mathbf{h}_k\sum_{i\neq k}^{K}\mathbf{w}_i s_i + n_k, \tag{24}$$

where the first component denotes the expected received signal of the $k$-th user, the second component denotes the interference, and the third component denotes the additive noise. Based on that, the received SINR of the $k$-th user can be given as

$$\gamma_k = \frac{|\mathbf{h}_k \mathbf{w}_k|^2}{\sum_{i \neq k}^{K} |\mathbf{h}_k \mathbf{w}_i|^2 + \sigma^2}. \tag{25}$$

Based on the analysis above, there are two main schemes for optimization-based block-level precoding, as discussed in the following.

### *5.1.2 Power minimization (PM) scheme*

Power minimization precoding, also known as minimum power beamforming[1], is a technique used to minimize the total transmitted power subject to a set of quality of service (QoS) constraints. The goal of this technique is to transmit the signal with the minimum possible power while ensuring that the received signal quality meets the desired level. This technique is particularly useful in situations where power consumption is a critical issue or in large-scale MIMO systems where the number of antennas is much larger than the number of users.

The PM design problem can be formulated as below [13]:

$$\begin{aligned} \mathcal{P}_1 : \ & \min_{\mathbf{w}_i} \sum_{i=1}^{K} \|\mathbf{w}_i\|_F^2 \\ \text{s.t.} \quad & \frac{|\mathbf{h}_k \mathbf{w}_k|^2}{\sum_{i \neq k}^{K} |\mathbf{h}_k \mathbf{w}_i|^2 + \sigma^2} \geq \Gamma_k, \forall k \in \{1, 2, \cdots, K\} \end{aligned} \tag{26}$$

where $\Gamma_k$ denotes the SINR threshold for the $k$-th user. It is proved that $\mathcal{P}_1$ is convex which can be solved via convex optimization algorithms efficiently. In addition to conventional convex optimization algorithms, literature has revealed an uplink-downlink duality in ref. [14], which has led to the development of an efficient iterative algorithm for solving downlink precoding optimization. Meanwhile, after transforming PM optimization into a semi-definite programming (SDP) problem, the semi-definite relaxation (SDR) approach [15–17] can be used to design the precoding matrix efficiently.

### *5.1.3 SINR balancing (SB) scheme*

SINR balancing precoding is a technique used to balance the signal-to--interference-plus-noise ratio (SINR) across all users in a multi-user system. The goal of this technique is to allocate the transmit power among the users such that each user experiences an equal SINR. This technique is particularly useful in situations where there are multiple users with different channel conditions, as it ensures that each user receives an equal quality of service. To be more specific, the SB design problem can be formulated as below [18]:

[1] It is noted that in this chapter the term 'beamforming' and 'precoding' are interchangeable.

$$\begin{aligned} \mathcal{P}_2: \quad & \max_{\mathbf{w}_i} \min_k \gamma_k \\ \text{s.t.} \quad & \gamma_k = \frac{|\mathbf{h}_k \mathbf{w}_k|^2}{\sum_{i \neq k}^{K} |\mathbf{h}_k \mathbf{w}_i|^2 + \sigma^2}, \forall k \in \{1, 2, \cdots, K\} \\ & \sum_{i=1}^{K} \|\mathbf{w}_i\|_F^2 \leq P_0 \end{aligned} \tag{27}$$

where $P_0$ is the maximum transmit power. Unlike the PM design problem, $\mathcal{P}_2$ is non-convex, which brings difficulties to the optimal precoding design. However, SB precoding can be efficiently designed through the bisection search method in ref. [16], or via an iterative algorithm in [14].

## 5.2 Symbol-level precoding

Block-level precoding is a precoding design based on CSI and is generally independent of the transmitted symbols. These algorithms tend to eliminate inter-user interference. In recent years, symbol-level precoding has received increasing attention [19]. Compared with block-level precoding, symbol-level precoding accomplishes precoding design based on both CSI and transmitted symbols, which gives it the ability to manipulate interference vectors more wisely compared with block-level precoding. With symbol-level precoding, the system can manage and utilize inter-user interference, which offers an additional power gain to improve system performance. In this subsection, we first introduce the concept of constructive interference (CI) to reveal the main idea of interference exploitation and then discuss the design problem of symbol-level precoding in different scenarios.

### *5.2.1 Concept for interference exploitation*

Interference is commonly considered a factor that limits performance in wireless communication systems. It arises due to the superimposition of transmit signals for different users in the wireless channel during multi-user transmission. Precoding strategies capitalize on the availability of CSI at the base station, along with data symbol information, to predict interference before transmission. Information theory analysis reveals that known interference will not affect the broadcast channel's capacity when CSI is available at the transmitter. However, most existing linear precoding schemes aim to eliminate, avoid or limit interference, and operate on a block level. Recent studies suggest that constructive interference (CI) precoding via Symbol-Level Precoding (SLP) can control both the power and direction of interfering signals, allowing interference to contribute to error-less signal detection and improve system performance [20]. Interference exploitation techniques are most useful in systems where interference can be predicted. In this subsection, we will give an illustrative example to demonstrate the division of instantaneous interference into CI and destructive interference (DI) [20].

Let us consider a scenario where the desired symbol $u$ is from a nominal BPSK constellation, with the assumption that $u = 1$. We use $i$ to denote the interfering signal and discuss two cases: (i) $i > 0$ and (ii) $i < 0$.

In the first case, when $i > 0$, as shown in **Figure 3(a)**, the received signal can be expressed as $\tilde{y} = h_u u + \tilde{h}_i i + n = \tilde{r} + n$, where $\tilde{r}$ represents the received signal excluding noise, and $n$ denotes the additive noise at the receiver side. **Figure 3(a)** shows that $\text{Proj}_{\text{O}\tilde{\text{E}}}(\tilde{r}) > \text{Proj}_{\text{O}\tilde{\text{E}}}(h_u u)$, which means that the interference has pushed $r$ further away

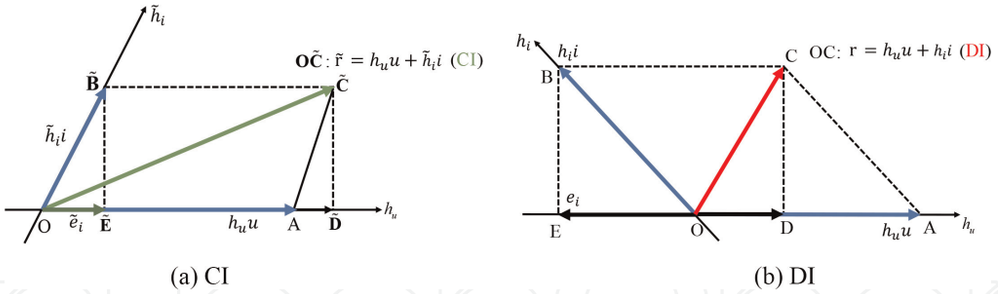


**Figure 3.**
*The geometrical representation of CI and DI.*

from the detection threshold of BPSK when compared to the original data symbol $u$. Here $\text{Proj}_{\mathbf{d}}(\mathbf{x})$ denotes the projection of vector $\mathbf{x}$ on the direction of $\mathbf{d}$. In this situation, the interfering signal is actually constructive and contributes to the useful signal power. Given a fixed noise power, $\tilde{y} = \tilde{r} + n$ is more likely to be detected correctly than the interference-free case $y' = h_u u + n$. Thus, we can expect improved performance.

On the other hand, in the second case, when $i < 0$, as shown in **Figure 3(b)**, the interfering signal causes the received signal $r$ to move closer to the detection threshold. In this case, the interfering signal reduces the useful signal power and is therefore destructive. The noiseless received signal $r = h_u u + h_i i$ is more susceptible to noise than $r' = u$ in this scenario.

In summary, symbol-level precoding offers more precise interference management and control, with the added benefit of improved performance through beneficial interference. This makes it a better communication performance option compared to traditional block-level precoding. Next, we will introduce the design principles of symbol-level precoding by discussing classical CI-SLP precoding methods.

### *5.2.2 Phase rotation metric*

As depicted in **Figure 4**, CI-SLP is a technique that manipulates inter-user interference to ensure that the noise-free receive signal falls within the constructive region.

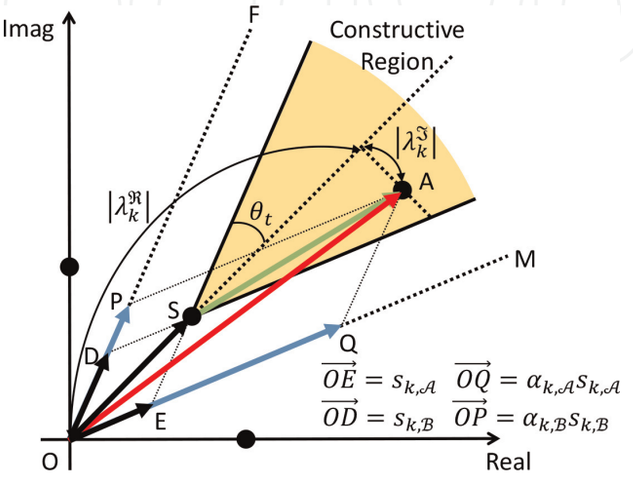


**Figure 4.**
*CI-SLP, 'phase-rotation' metric, 8-PSK.*

The SLP matrix **W** is designed to maximize the distance between the worst user's constructive region and the detection threshold, thereby improving the transmission performance. Masouros [21] first proposed the "phase rotation" metric for PSK modulated systems. Based on this metric, the noise-free receive signal can be expressed as follows [22]:

$$\vec{OA} = \mathbf{h}_k^{\mathrm{T}}\mathbf{W}\mathbf{s} = \lambda_k s_k. \tag{28}$$

The constructive factor $\lambda_k$ quantifies the constructive effect of interference exploitation for that user. Based on this factor, the constructive region can be described as follows:

$$\begin{aligned} &\theta_{AB} \le \theta_t \Rightarrow \tan\theta_{AB} \le \tan\theta_t \\ &\Rightarrow \frac{\left|j \cdot \lambda_k^{\mathcal{I}} s_k\right|}{\left|\left[\lambda_k^{\mathcal{R}} - \sqrt{\Gamma_k \sigma^2}\right] s_k\right|} \le \tan\theta_t \\ &\Rightarrow \left[\lambda_k^{\mathcal{R}} - \sqrt{\Gamma_k \sigma^2}\right] \tan\theta_t \ge \left|\lambda_k^{\mathcal{I}}\right| \end{aligned} \tag{29}$$

According to the transmit power minimization criterion, the CI-SLP design problem is shown below

$$\begin{aligned} \mathcal{P}_3:\ &\min_{\mathbf{w}} \|\mathbf{W}\mathbf{s}\|_{\mathrm{F}}^2 \\ &\text{s.t.} \mathbf{h}_k \mathbf{W}\mathbf{s} = \lambda_k s_k, \forall k \in \{1, 2, \cdots, K\} \\ &\left[\lambda_k^{\mathcal{R}} - \sqrt{\Gamma_k \sigma^2}\right] \tan\theta_t \ge \left|\Gamma_k^{\mathcal{I}}\right|, \forall k \in \{1, 2, \cdots, K\}, \end{aligned} \tag{30}$$

where $\Gamma_k^{\mathcal{I}}$ denotes the Quality of Serves (QoS) threshold of the $k$-th user.

The convexity of $\mathcal{P}_3$ can be proven, similar to the traditional PM problem, enabling the use of several convex optimization algorithms to solve this problem conveniently. Similarly, the CI-SLP design problem based on the SB criterion can be formulated as

$$\begin{aligned} \mathcal{P}_4:\quad &\max_{\mathbf{W},t} t \\ \text{s.t.}\quad &\mathbf{h}_k \mathbf{W}\mathbf{s} = \lambda_k s_k, \forall k \in \{1, 2, \cdots, K\} \\ &\left[\lambda_k^{\mathcal{R}} - t\right] \tan\theta_t \ge \left|\lambda_k^{\mathcal{I}}\right|, \forall k \in \{1, 2, \cdots, K\} \\ &\|\mathbf{W}\mathbf{s}\|_{\mathrm{F}}^2 \le P_0. \end{aligned} \tag{31}$$

It is worth noting that the convexity of the equation shown above can also be proven, which distinguishes it from the traditional SB problem and renders it more mathematically tractable.

### *5.2.3 Symbol scaling metric*

In QAM modulation, the interference exploitation is conditional, unlike PSK modulation. The constellation signal points of QAM modulation can be classified into four groups based on their interference exploitation characteristics, as shown in **Figure 5**. Group A' represents signal points that do not exploit any interference, while Group B′ and Group C′ represent signal points that exploit interference in the real and

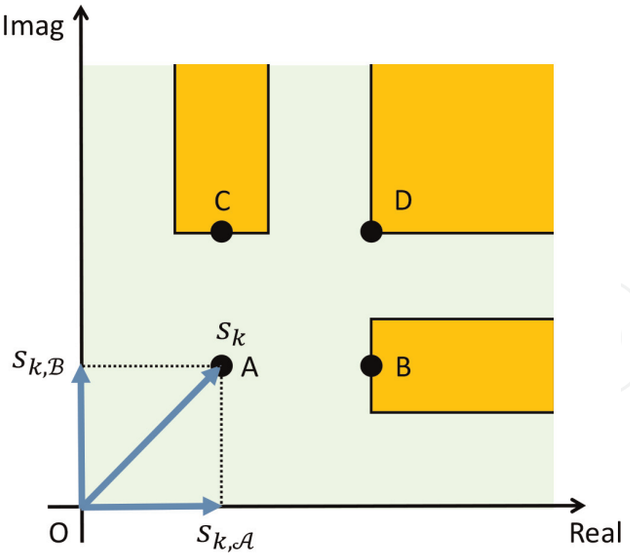


**Figure 5.**
*CI-SLP, 'symbol-scaling' metric, 16-QAM.*

imaginary parts, respectively. Group D′ represents signal points that exploit interference in both the real and imaginary parts, resulting in full interference exploitation.

The interference exploitation procedure via the "symbol-scaling" [23] metric and decomposition of the noiseless receive signal of the $k$-th user can be described as follows:

$$\mathbf{h}_k^{\mathrm{T}}\mathbf{W}\mathbf{s} = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{s}_k, \tag{32}$$

where

$$\boldsymbol{\alpha}_k = \left[\alpha_k^{\mathcal{A}}, \alpha_k^{\mathcal{B}}\right]^{\mathrm{T}}, \mathbf{s}_k = \left[s_k^{\mathcal{A}}, s_k^{\mathcal{B}}\right]^{\mathrm{T}} \tag{33}$$

with

$$s_k^{\mathcal{A}} = \Re(s_k), s_k^{\mathcal{B}} = (s_k), \quad k = 1, 2, \ldots, K. \tag{34}$$

Based on that, the CI-SLP design problem in QAM-modulated systems can be described as follows

$$\begin{aligned} \mathcal{P}_5: \max_{\mathbf{W},\boldsymbol{\Omega}_k,t} \quad & t \\ \text{s.t.} \quad & \mathbf{h}_k^T\mathbf{W}\mathbf{s} = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{s}_k, \forall k \in \mathcal{K} \\ & t \le \alpha_m^{\mathcal{O}}, \forall \alpha_m^{\mathcal{O}} \in \mathcal{O} \\ & t = \alpha_n^{\mathcal{I}}, \forall \alpha_n^{\mathcal{I}} \in \mathcal{I} \\ & \|\mathbf{W}\mathbf{s}\|_2^2 \le p_0. \end{aligned} \tag{35}$$

The set $\mathcal{O}$ comprises the indices of successful interference exploitation corresponding to the real part of the symbol in group B′, the imaginary part of the symbol in group C′, and both the real and imaginary parts of the symbol in group D′. Conversely, the set $\mathcal{I}$ comprises the indices of unsuccessful interference exploitation corresponding to the imaginary part of the symbol in group B′, the real part of the

symbol in group C′, and both the real and imaginary parts of the symbol in group A'. It follows that $\mathcal{O}$ and $\mathcal{I}$ satisfy the following relationship:

$$\begin{aligned} \mathcal{O} \cup \mathcal{I} = \mathcal{K}, \mathcal{O} \cap \mathcal{I} = \varnothing, \\ \text{card}\{\mathcal{O}\} + \text{card}\{\mathcal{I}\} = 2K. \end{aligned} \tag{36}$$

The definitions of the sets $\mathcal{O}$ and $\mathcal{I}$ reveal the difference between the phase rotation criterion and the symbol scaling criterion. The former exploits interference unconditionally, i.e., all constellation points participate in interference exploitation, while the latter exploits interference conditionally. For QAM modulation systems, the inner constellation points do not participate in interference exploitation, and beneficial interference only results in performance gains for the outer constellation points. This difference arises from the inherent properties of QAM and PSK modulation schemes. In PSK modulation, the amplitude of the constellation points does not carry any information, and therefore, any constellation point can be exploited for interference without adversely affecting the detection of other constellation points. However, for the inner constellation points in QAM modulation, interference vectors that push the noiseless receive signal points in any direction will adversely affect the error decision of other constellation points. It is worth noting that these two design criteria only differ in their description of the interference exploitation process and are essentially equivalent. [23] has proven that under PSK modulation, the symbol scaling criterion and the phase rotation criterion are equivalent, as depicted in **Figure 4**, where the symbol-scaling metric is also applicable. Therefore, the symbol scaling criterion is more universal in this sense.

## 6. Hardware-efficient precoding

The use of technologies such as General Artificial Intelligence (AI), has led to a surge in users' demand for mobile data traffic. One way to address this issue is to utilize massive MIMO systems, which employ a large number of antennas at the base station to improve data rate and link reliability. This approach allows signals to be dynamically adjusted in both horizontal and vertical directions, reducing interference between small areas and enabling more accurate pointing toward specific users. However, directly applying Massive MIMO technology to traditional communication system architectures can result in new problems [3]. To be more specific, traditional MIMO systems equip each antenna with RF chains and high-resolution DACs, causing significant power loss when the antenna array is large. To solve this issue, there are three general approaches: reducing the number of RF chains, lowering the resolution of the DACs, or employing power-efficient nonlinear power amplifiers. However, these hardware-efficient architectures introduce new challenges to precoding designs, which will be explained in more detail in the following.

### 6.1 Hybrid analog-digital (HAD) precoding

Fully-digital precoders can be used in traditional sub-6 GHz bands, but for millimeter wave (mmWave) communications, the cost and power consumption of hardware components make this approach impractical. To solve this issue, researchers

have developed the hybrid analog-digital structure, which provides a promising trade-off between the cost, complexity, and capacity of the mmWave network. This structure reduces hardware complexity and power consumption by reducing the total number of RF chains. Specifically, the mmWave transceivers first process data streams with a low-dimension digital precoder, followed by high-dimension analog precoding using low-cost phase shifters, switches [24], or lens [25]. While the performance of the hybrid precoder is usually inferior to that of a fully-digital precoder, it offers a cost-efficient and energy-efficient solution for mmWave communication.

In an MU-MIMO system illustrated in **Figure 6**, $N_t$ transmit antennas are utilized by the BS to serve $K$ single-antenna users simultaneously. The transmitter has $N_{\mathrm{RF}}^t$ RF chains, where $N_{\mathrm{RF}}^t \ll N_t$. In this subsection, we use phase shifter-based hybrid architecture as an illustrative example, without loss of generality.

Based on that, the transmit symbol vector $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s}, \tag{37}$$

where $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}_{N_{\mathrm{RF}}^t \times N_t}$ denotes the hybrid precoding matrix, $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}_{K \times N_{\mathrm{RF}}^t}$ denotes the digital baseband precoding matrix, and $\mathbf{s} \in \mathbb{C}_{K \times 1}$ denotes the data symbol vector with $\mathbb{E}\{\mathbf{s}\mathbf{s}^{\mathrm{H}}\} = \frac{1}{K}\mathbf{I}_K$, respectively. Considering that the hybrid precoding matrix is the mathematical description of phase shifters, we have the constant-module constraint for the hybrid precoder, as shown below:

$$|\mathbf{F}_{\mathrm{RF}}(i,j)| = 1, \quad 1 \le i \le N_{\mathrm{RF}}^t, \quad 1 \le j \le N_t. \tag{38}$$

Meanwhile, the power constraint at the transmit side can be expressed as

$$\|\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{RF}}\|_{\mathrm{F}}^2 = P_0, \tag{39}$$

where $P_0$ is the maximum transmit power.

Based on that, the $k$-th user's received signal can be expressed as

$$y_k = \mathbf{h}_k^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + n_k, \tag{40}$$

where $\mathbf{h}_k \in \mathbb{C}_{N_t \times 1}$ denotes the complex channel matrix for the $k$-th user, and $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ denotes the additive Gaussian noise vector for the $k$-th user with the zero-mean and $\sigma_k^2$ noise power.

Aimed at maximizing the spectral efficiency, a common HAD precoding design problem can be formulated as [26].

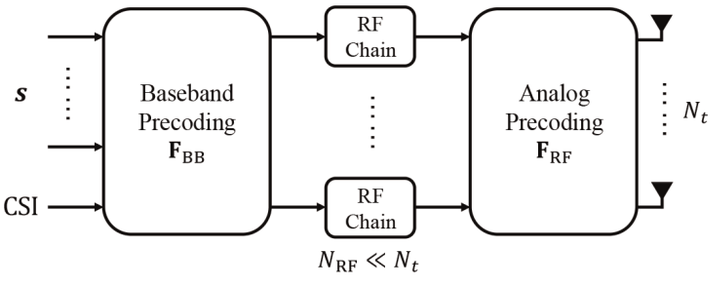


**Figure 6.**
*The HAD MIMO system.*

$$\begin{aligned}\mathcal{P}_6: \max_{\mathbf{F}_{\mathrm{RF}}, \mathbf{f}_k^{\mathrm{BB}}} & \sum_{k=1}^{K} \log_2\left(1+\frac{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\mathbf{f}_k^{\mathrm{BB}}\right|^2}{\sum_{i\neq k}\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\mathbf{f}_i^{\mathrm{BB}}\right|^2+\sigma_k^2}\right) \\ \text{s.t. } & \mathbf{F}_{\mathrm{RF}}\in\mathcal{F}, \forall 1\le k\le K, \\ & \left\|\mathbf{F}_{\mathrm{RF}}\left[\mathbf{f}_1^{\mathrm{BB}}, \mathbf{f}_2^{\mathrm{BB}}, \ldots, \mathbf{f}_K^{\mathrm{BB}}\right]\right\|_F^2 = P_0,\end{aligned} \tag{41}$$

where $\mathcal{F}$ denotes the available region of $\mathbf{F}_{\mathrm{RF}}$, as defined below:

$$\mathcal{F}=\left\{\mathbf{F}_{\mathrm{RF}}\middle|\left|\mathbf{F}_{\mathrm{RF}}(i,j)\right|=1,\ 1\le i\le N_{\mathrm{RF}}^t,\ 1\le j\le N_t\right\}. \tag{42}$$

The non-convexity of $\mathcal{P}_6$ is due to the constant-module constraint of $\mathbf{F}_{\mathrm{RF}}$, making it difficult to solve. To address this issue, a two-stage hybrid precoding algorithm was proposed in ref. [27] where the analog precoder maximizes the effective channel gain and the digital precoder mitigates multi-user interference based on the ZF principle. In ref. [28], it was demonstrated that hybrid precoding can achieve any fully-digital precoding when the number of RF chains is twice the number of data streams, and a near-optimal hybrid precoding design was proposed for single-user and multi-user transmissions with fewer RF chains. Reference [29] focused specifically on partially-connected structures in multi-user scenarios and proposed hybrid precoding designs based on successive interference cancelation (SIC). This approach decomposes the total spectral efficiency optimization problem into a series of sub-rate optimization problems that can be solved efficiently using the power iteration algorithm. Other works on hybrid precoding include low-complexity designs based on MRT [30], virtual path selection [31], and SVD [32].

### 6.2 Low-bit precoding

Using low-resolution DACs instead of high-resolution DACs in massive MIMO architecture can be an effective way to reduce the power consumption of BS. This approach reduces the power consumption per RF chain, as depicted in **Figure** 7, instead of reducing the number of RF chains like in the hybrid architecture.

High-resolution DACs are required for each transmit signal to avoid signal distortion, but they consume significant power due to their linear relationship with bandwidth and exponential relationship with resolution [33]. Large-scale antenna arrays, with hundreds of antenna elements, require a significantly large number of DACs, posing practical challenges. To address this issue, low-resolution DACs, particularly 1-bit DACs, can substantially simplify hardware and reduce the corresponding power consumption at the BS. Furthermore, 1-bit DACs generate CE signals, which facilitate

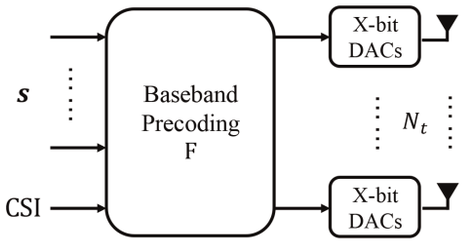


**Figure 7.**
*The architecture of low-bit MIMO system.*

the use of power-efficient amplifiers, further reducing hardware complexity. The common low-bit precoding design problem can be formulated as [34].

$$\begin{aligned} \mathcal{P}_7 : \ & \min_{\mathbf{x}} \|\mathbf{s} - \beta_{\mathrm{DAC}} \cdot \mathbf{H}\mathbf{x}\|_2^2 + K\beta_{\mathrm{DAC}}^2\sigma^2 \\ & \text{s.t.} \mathbf{x} \in \mathcal{X}_{\mathrm{DAC}} \\ & \beta_{\mathrm{DAC}} > 0. \end{aligned} \tag{43}$$

The optimization problem $\mathcal{P}_7$ seeks to minimize the MSE between transmitted and received symbols using low-resolution DACs. For 1-bit DACs, the set of output signals is denoted as $\mathcal{X}_{\mathrm{DAC}} = \left\{ \pm\sqrt{\frac{P_0}{2N_t}} \pm \sqrt{\frac{P_0}{2N_t}} \cdot j \right\}$. In ref. [35], a non-linear precoding method based on a biconvex relaxation framework achieved promising performance with a low computational cost. Its corresponding VLSI design architectures were illustrated in refs. [36]. Alternatively, Jacobsson et al. [37] proposed several 1-bit precoding schemes based on SDR, sphere encoding, and squared $l_\infty$-norm relaxation, while Landau and de Lamare [38] described a 1-bit precoding method based on the branch-and-bound framework that can theoretically achieve optimal performance. Other downlink precoding designs for low-resolution DACs include SER minimization in refs. [39, 40] and alternating minimization in ref. [34]. Nonlinear precoding designs tend to outperform linear methods when low-resolution DACs are used at the transmitter. For example, CI-based symbol-level precoding design has been discussed in low-resolution DACs systems [41–43]. Several efficient solutions [43–45] have been proposed for the NP-hard optimization problem, both for 1-bit and few-bit DACs systems.

### 6.3 Nonlinearity-aware precoding

In a massive multiple-input-multiple-output (MIMO) system, the integration of power-efficient nonlinear power amplifiers (PAs) can reduce the power consumption of each RF chain, similar to the architecture of low-bit digital-to-analog converters. Consequently, this leads to an improved energy efficiency of the system. However, in traditional multi-antenna systems, the limited linear region of nonlinear PAs causes significant signal distortions when transmitting signals with high peak-to-average power ratios (PAPRs). This consequently negatively impacts system performance.

To resolve the issue of PAPR, traditional research falls into two categories: (a) constant envelope precoding (CEP) schemes that maintain signal power at a constant value, commonly known as SLP schemes; and (b) frame-level precoding matrix optimization aimed at reducing the PAPR of the transmit signal. CEP eliminates the performance loss introduced by nonlinear PAs by limiting the amplitude of the transmit signal to a constant value, while the low-PAPR precoding relaxes the strict CE constraint by allowing the maximum PAPR to a certain value. In recent years, there has been a growing body of literature that explores the precoding design based on the knowledge of the nonlinear response characteristics of PAs. This approach represents a departure from the traditional emphasis solely on reducing the peak-to-average power ratio (PAPR) of transmitted signals. To be more specific, nonlinearity-aware precoding utilizes a clipping function to model the response characteristics of nonlinear PAs and developed a precoder that can resist both interference and PA nonlinearity by describing the modeled response characteristics [46]. The nonlinearity-aware precoding system can be shown in **Figure 8** Considering a multi-

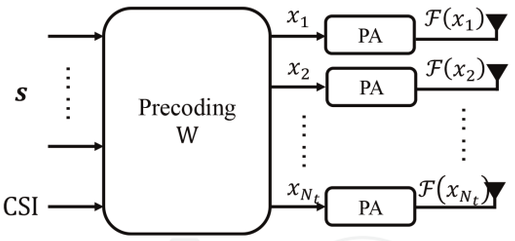


**Figure 8.**
*The nonlinearity-aware precoding system.*

user MISO system, the $k$-th user's received signal can be expressed as.where $\mathscr{F}(\cdot): \mathbb{C} \to \mathbb{C}$ is the nonlinearity function that delineates the input-output response properties of nonlinear power amplifiers [47]. Based on that, the nonlinearity-aware precoding design problem aimed at maximizing the sum rate can be expressed as

$$y_k = \mathbf{h}_k^T \mathscr{F}(\mathbf{W}\mathbf{s}) + n_k, \tag{44}$$

$$\begin{aligned} \mathcal{P}_8 : \max_{\mathbf{W} \in \mathbb{C}^{K \times N_t}} \quad & R_{\text{sum}}(\mathbf{W}) \\ \text{s.t.} \quad & \mathbb{F}\left[\|\phi(\mathbf{W}\mathbf{s})\|^2\right] = P_{\text{t}}, \end{aligned} \tag{45}$$

where $P_t$ denotes the maximum transmit power constraint. The problem has been addressed through the introduction of a distortion-aware beamforming (DAB) algorithm as proposed by [48]. This method adopts an iterative approach to optimize data rate while minimizing the effect of distortions. In addition, several other precoding strategies have been developed with a focus on accounting for nonlinearity in the system. Specifically, Aghdam et al. [49] studied a precoding scheme that incorporates power amplifier effects in massive MU-MIMO downlink systems and put forth a robust algorithm to mitigate interference and nonlinearity resulting from power amplifiers. Moreover, Zayani et al. [50] presented a power control mechanism and a precoding scheme for SU-MISO communication systems that utilize nonlinear power amplifiers at the base station. The proposed method maximizes the received SINR while utilizing an iterative precoding algorithm. Finally, Jee et al. [51] optimized both precoding and power allocation strategies jointly to maximize the achievable sum rate of MU-MIMO systems.

## 7. Conclusions

In this chapter, we have provided a comprehensive overview of precoding design for achieving spatial multiplexing in MIMO communications.

We began in Section 3 by introducing the fundamental concepts of MIMO systems, including the mathematical description of MIMO communications, performance metrics, and the increasingly important and widely used massive MIMO technology in 5G. These concepts laid a solid foundation for the subsequent discussions on the precoding design.

In Section 4, we discussed traditional precoding design methods, including closed-form linear block-level precoding techniques such as MRT, ZF, and RZF, as well as traditional nonlinear symbol-level precoding techniques such as THP and VP.

Through these algorithms, we introduced the basic principles and guidelines of precoding design.

In Section 5, we discussed more complex precoding design methods based on convex optimization, including power minimization, SINR balancing, and the emerging CI-SLP precoding. These methods provide more flexibility and adaptability in precoding design and can achieve better performance in practical communication systems.

In Section 6, we focused on the hardware-efficient precoding design for massive MIMO systems in 5G. We discussed hybrid analog-digital precoding, low-bit precoding, and nonlinearity-aware precoding, which are essential for reducing power consumption and computational complexity while maintaining high communication performance.

Overall, this chapter highlights the importance of efficient precoding design for achieving efficient and reliable wireless transmission. Precoding design is a critical component of MIMO technology, and it requires a careful balance between communication performance, power consumption, and computational complexity. The discussions in this chapter provide a comprehensive understanding of the various precoding techniques that can be employed to achieve spatial multiplexing in MIMO communications and underscore the significance of efficient precoding design for realizing the full potential of MIMO technology in wireless communication systems.

## Nomenclature

| | |
|---|---|
| SISO | single-input single-output |
| MISO | multi-input single-output |
| MIMO | multi-input multi-output |
| MRT | maximum ratio transmission |
| ZF | zero-forcing |
| RZF | regularized zero-forcing |
| THP | Tomlinson-Harashima precoding |
| VP | vector perturbation |
| IoT | internet of things |
| PM | power minimization precoding |
| SNR | signal-to-noise ratio |
| SINR | signal-to-interference-and-noise ratio |
| SB | SINR balancing precoding |
| IE | interference exploitation |
| CI | constructive interference |
| DI | destructive interference |
| BLP | block-level precoding |
| SLP | symbol-level precoding |
| HAD | hybrid analog-digital precoding |
| CEP | constant envelope precoding |

## Author details

Haonan Wang* and Ang Li*
Xi'an Jiaotong University, Xi'an, Shaanxi, China

*Address all correspondence to: whn8215858@stu.xjtu.edu.cn and ang.li.2020@xjtu.edu.cn

## References

[1] Andrews JG, Buzzi S, Choi W, et al. What will 5G Be? IEEE Journal on Selected Areas in Communications. 2014;**32**(6):1065-1082. DOI: 10.1109/JSAC.2014.2328098

[2] Tse D, Viswanath P. Fundamentals of Wireless Communication. Cambridge, UK: Cambridge University Press; 2005

[3] Swindlehurst AL, Ayanoglu E, Heydari P, et al. Millimeter-wave massive MIMO: The next wireless revolution? IEEE Communications Magazine. 2014;**52**(9):56-62. DOI: 10.1109/MCOM.2014.6894453

[4] Joham M, Utschick W, Nossek JA. Linear transmit processing in MIMO communications systems. IEEE Transactions on Signal Processing. 2005; **53**(8):2700-2712. DOI: 10.1109/TSP.2005.850331

[5] Lu L, Li GY, Swindlehurst AL, et al. An overview of massive MIMO: Benefits and challenges. IEEE Journal of Selected Topics in Signal Processing. 2014;**8**(5): 742-758. DOI: 10.1109/JSTSP.2014.2317671

[6] Hochwald BM, Marzetta TL, Tarokh V. Multiple-Antenna Channel hardening and its implications for rate feedback and scheduling. IEEE Transactions on Information Theory. 2004;**50**(9):1893-1909. DOI: 10.1109/TIT.2004.833345

[7] Heath RW, Gonzalez-Prelcic N, Rangan S, et al. An overview of signal processing techniques for Millimeter wave MIMO systems. IEEE Journal of Selected Topics in Signal Processing. 2016;**10**(3):436-453. DOI: 10.1109/JSTSP.2016.2523924

[8] Jiang Y, Varanasi MK, Li J. Performance analysis of ZF and MMSE equalizers for MIMO systems: An In-depth study of the high SNR regime. IEEE Transactions on Information Theory. 2011;**57**(4):2008-2026. DOI: 10.1109/TIT.2011.2112070

[9] Peel CB, Hochwald BM, Swindlehurst AL. A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: Channel inversion and regularization. IEEE Transactions on Communications. 2005;**53**(1):195-202. DOI: 10.1109/TCOMM.2004.840638

[10] Costa M. Writing on dirty paper. IEEE Transactions on Information Theory. 1983;**29**(3):439-441. DOI: 10.1109/TIT.1983.1056659

[11] Garcia-Rodriguez A, Masouros C. Power-efficient Tomlinson-Harashima precoding for the downlink of multi-user MISO systems. IEEE Transactions on Communications. 2014;**62**(6): 1884-1896. DOI: 10.1109/TCOMM.2014.2317189

[12] Hochwald BM, Peel CB, Swindlehurst AL. A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: Perturbation. IEEE Transactions on Communications. 2005;**53**(3): 537-544. DOI: 10.1109/TCOMM.2005.843995

[13] Bengtsson M, Ottersten B. Optimum and suboptimum transmit beamforming. In: Handbook of Antennas in Wireless Communications. Boca Raton, Florida, US: CRC Press; 2018. pp. 18-1-18-33. DOI: 10.1201/9781315220031-18

[14] Schubert M, Boche H. Solution of the multiuser downlink beamforming problem with individual SINR

constraints. IEEE Transactions on Vehicular Technology. 2004;**53**(1):18-28. DOI: 10.1109/TVT.2003.819629

[15] Huang Y, Palomar DP. Rank-constrained separable semidefinite programming with applications to optimal beamforming. IEEE Transactions on Signal Processing. 2009; **58**(2):664-678. DOI: 10.1109/TSP.2009.2031732

[16] Huang Y, Palomar DP. A dual perspective on separable semidefinite programming with applications to optimal downlink beamforming. IEEE Transactions on Signal Processing. 2010; **58**(8):4254-4271. DOI: 10.1109/TSP.2010.2049570

[17] Law KL, Wen X, Vu MT, et al. General rank multiuser downlink beamforming with shaping constraints using real-valued OSTBC. IEEE Transactions on Signal Processing. 2015; **63**(21):5758-5771. DOI: 10.1109/TSP.2015.2455516

[18] Wiesel A, Eldar YC, Shamai S. Linear precoding via conic optimization for fixed MIMO receivers. IEEE Transactions on Signal Processing. 2005; **54**(1):161-176. DOI: 10.1109/TSP.2005.861073

[19] Li A, Spano D, Krivochiza J, et al. A tutorial on interference exploitation via symbol-level precoding: Overview, state-of-the-art and future directions. IEEE Communications Surveys & Tutorials. 2020;**22**(2):796-839. DOI: 10.1109/COMST.2020.2980570

[20] Masouros C, Ratnarajah T, Sellathurai M, et al. Known interference in the cellular downlink: A performance limiting factor or a source of green signal power? IEEE Communications Magazine. 2013;**51**(10):162-171. DOI: 10.1109/MCOM.2013.6619580

[21] Masouros C. Correlation rotation linear precoding for MIMO broadcast communications. IEEE Transactions on Signal Processing. 2010;**59**(1):252-262. DOI: 10.1109/TSP.2010.2088395

[22] Li A, Masouros C. Interference exploitation precoding made practical: Optimal closed-form solutions for PSK modulations. IEEE Transactions on Wireless Communications. 2018;**17**(11): 7661-7676. DOI: 10.1109/TWC.2018.2869382

[23] Li A, Masouros C, Vucetic B, et al. Interference exploitation precoding for multi-level modulations: Closed-form solutions. IEEE Transactions on Communications. 2020;**69**(1):291-308. DOI: 10.1109/TCOMM.2020.3031616

[24] Méndez-Rial R, Rusu C, González-Prelcic N, et al. Hybrid MIMO architectures for Millimeter wave communications: Phase shifters or switches? IEEE Access. 2016;**4**:247-267. DOI: 10.1109/ACCESS.2015.2514261

[25] Brady J, Behdad N, Sayeed AM. Beamspace MIMO for Millimeter-wave communications: System architecture, Modeling, analysis, and measurements. IEEE Transactions on Antennas and Propagation. 2013;**61**(7):3814-3827. DOI: 10.1109/TAP.2013.2254442

[26] El Ayach O, Rajagopal S, Abu-Surra S, et al. Spatially sparse precoding in Millimeter wave MIMO systems. IEEE Transactions on Wireless Communications. 2014;**13**(3):1499-1513. DOI: 10.1109/TWC.2014.011714.130846

[27] Alkhateeb A, Leus G, Heath RW. Limited feedback hybrid precoding for multi-user Millimeter wave systems. IEEE Transactions on Wireless Communications. 2015;**14**(11): 6481-6494. DOI: 10.1109/TWC.2015.2455980

[28] Sohrabi F, Yu W. Hybrid digital and Analog beamforming Design for Large-Scale Antenna Arrays. IEEE Journal of Selected Topics in Signal Processing. 2016;**10**(3):501-513. DOI: 10.1109/JSTSP.2016.2520912

[29] Gao X, Dai L, Han S, et al. Energy-efficient hybrid Analog and digital precoding for mmWave MIMO systems with large antenna arrays. IEEE Journal on Selected Areas in Communications. 2016;**34**(4):998-1009. DOI: 10.1109/JSAC.2016.2549418

[30] Liang L, Xu W, Dong X. Low-complexity hybrid precoding in massive multiuser MIMO systems. IEEE Wireless Communications Letters. 2014;**3**(6): 653-656. DOI: 10.1109/LWC.2014.2363831

[31] Li A, Masouros C. Hybrid Analog-digital Millimeter-wave MU-MIMO transmission with virtual path selection. IEEE Communications Letters. 2016; **21**(2):438-441. DOI: 10.1109/LCOMM.2016.2621741

[32] Li A, Masouros C. Hybrid precoding and combining Design for Millimeter-Wave Multi-User MIMO based on SVD. In: 2017 IEEE International Conference on Communications (ICC'17), 21-25 May 2017; Paris. New York: IEEE; 2017. pp. 1-6. DOI: 10.1109/ICC.2017.7996970

[33] Allen PE, Dobkin R, Holberg DR. CMOS Analog Circuit Design. Amsterdam: Elsevier; 2011

[34] Chen JC. Alternating minimization algorithms for one-bit precoding in massive multiuser MIMO systems. IEEE Transactions on Vehicular Technology. 2018;**67**(8):7394-7406. DOI: 10.1109/TVT.2018.2836335

[35] Castaneda O, Goldstein T, Studer C. POKEMON: A non-linear beamforming algorithm for 1-bit massive MIMO. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), 5-9 March, 2017; New Orleans, LA, USA. New York: IEEE; 2017. pp. 3464-3468. DOI: 10.1109/ICASSP.2017.7952800

[36] Castañeda O, Jacobsson S, Durisi G, et al. 1-bit massive MU-MIMO precoding in VLSI. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 2017;**7**(4):508-522. DOI: 10.1109/JETCAS.2017.2772191

[37] Jacobsson S, Durisi G, Coldrey M, et al. Quantized precoding for massive MU-MIMO. IEEE Transactions on Communications. 2017;**65**(11): 4670-4684. DOI: 10.1109/TCOMM.2017.2723000

[38] Landau LTN, de Lamare RC. Branch-and-bound precoding for multiuser MIMO systems with 1-bit quantization. IEEE Wireless Communications Letters. 2017;**6**(6):770-773. DOI: 10.1109/LWC.2017.2740386

[39] Swindlehurst A, Saxena A, Mezghani A, et al. Minimum probability-of-error perturbation precoding for the one-bit massive MIMO downlink. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), 5-9 March, 2017; New Orleans, LA, USA. New York: IEEE; 2017. pp. 6483-6487. DOI: 10.1109/ICASSP.2017.7953405

[40] Shao M, Li Q, Ma WK. One-bit massive MIMO precoding via a minimum symbol-error probability design. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18), 15-20 April, 2018; Calgary, AB, Canada. New York: IEEE; 2018. pp. 3579-3583. DOI: 10.1109/ICASSP.2018.8461980

[41] Jedda H, Mezghani A, Nossek JA, et al. Massive MIMO downlink 1-bit precoding with linear programming for PSK Signaling. In: 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'17), 3-6 July, 2017; Sapporo, Japan. New York: IEEE; 2017. pp. 1-5. DOI: 10.1109/SPAWC.2017.8227757

[42] Li A, Masouros C, Swindlehurst AL. 1-bit massive MIMO downlink based on constructive interference. In: 2018 26th European Signal Processing Conference (EUSIPCO'18), 3-7 September, 2018; Rome, Italy. New York: IEEE; 2018. pp. 927-931. DOI: 10.23919/EUSIPCO.2018.8553556

[43] Li A, Liu F, Masouros C, et al. Interference exploitation 1-bit massive MIMO precoding: A partial branch-and-bound solution with near-optimal performance. IEEE Transactions on Wireless Communications. 2020;**19**(5): 3474-3489. DOI: 10.1109/TWC.2020.2973987

[44] Tsinos CG, Kalantari A, Chatzinotas S, et al. Symbol-level precoding with low resolution DACs for large-scale Array MU-MIMO systems. In: 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'18), 25-28 June, 2018; Kalamata, Greece. New York: IEEE; 2018. pp. 1-5. DOI: 10.1109/SPAWC.2018.8445995

[45] Bertsekas DP. Nonlinear programming. Journal of the Operational Research Society. 1997;**48**(3):334-334. DOI: 10.1057/palgrave.jors.2600425

[46] Sohrabi F, Liu YF, Yu W. One-bit precoding and constellation range design for massive MIMO with QAM signaling. IEEE Journal of Selected Topics in Signal Processing. 2018;**12**(3):557-570. DOI: 10.1109/JSTSP.2018.2823267

[47] Jedda H, Mezghani A, Swindlehurst AL, et al. Precoding under instantaneous per-antenna peak power constraint. In: 2017 25th European Signal Processing Conference (EUSIPCO'17), 28 August-2 September, 2017; Kos, Greece. New York: IEEE; 2017. pp. 863-867. DOI: 10.23919/EUSIPCO.2017.8081330

[48] Ding L, Zhou GT. Effects of even-order nonlinear terms on Predistortion linearization. In: Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop, 16 October 2002; Pine Mountain, GA, USA. New York: IEEE; 2002. pp. 1-6. DOI: 10.1109/DSPWS.2002.1231064

[49] Aghdam SR, Jacobsson S, Eriksson T. Distortion-aware linear precoding for Millimeter-wave multiuser MISO downlink. In: 2019 IEEE International Conference on Communications Workshops (ICC Workshops'19), 20-24 May 2019; Shanghai. New York: IEEE. pp. 1-6. DOI: 10.1109/ICCW.2019.8757031

[50] Zayani R, Shaïek H, Roviras D. Efficient precoding for massive MIMO downlink under PA nonlinearities. IEEE Communications Letters. 2019;**23**(9): 1611-1615. DOI: 10.1109/LCOMM.2019.2924001

[51] Jee J, Kwon G, Park H. Precoding design and power control for SINR maximization of MISO system with nonlinear power amplifiers. IEEE Transactions on Vehicular Technology. 2020;**69**(11):14019-14024. DOI: 10.1109/TVT.2020.3026752