IEEE Open Journal of Signal Processing

Received 19 October 2023; revised 5 February 2024; accepted 15 February 2024. Date of publication 11 March 2024; date of current version 23 April 2024. The review of this article was arranged by Associate Editor Christos Masouros.

Digital Object Identifier 10.1109/OJSP.2024.3375653

Spatial Sigma-Delta Modulation for Coarsely Quantized Massive MIMO Downlink: Flexible Designs by Convex Optimization

WAI-YIU KEUNG [1,2] AND WING-KIN MA [1] (Fellow, IEEE)

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong S.A.R., China
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong S.A.R., China

CORRESPONDING AUTHOR: WING-KIN MA (e-mail: wkma@ee.cuhk.edu.hk).

This work was supported by the General Research Fund (GRF) of Hong Kong Research Grant Council (RGC) under Project CUHK 14208819.

ABSTRACT This article considers the context of multiuser massive MIMO downlink precoding with low-resolution digital-to-analog converters (DACs) at the transmitter. This subject is motivated by the consideration that it is expensive to employ high-resolution DACs for practical massive MIMO implementations. The challenge with using low-resolution DACs is to overcome the detrimental quantization error effects. Recently, spatial Sigma-Delta ($\Sigma\Delta$) modulation has arisen as a viable way to put quantization errors under control. This approach takes insight from temporal $\Sigma\Delta$ modulation in classical DAC studies. Assuming a 1D uniform linear transmit antenna array, the principle is to shape the quantization errors in space such that the shaped quantization errors are pushed away from the user-serving angle sector. In the previous studies, spatial $\Sigma\Delta$ modulation was performed by direct application of the basic first- and second-order modulators from the $\Sigma\Delta$ literature. In this paper, we develop a general $\Sigma\Delta$ modulator design framework for any given order, for any given number of quantization levels, and for any given angle sector. We formulate our design as a problem of maximizing the signal-to-quantization-and-noise ratios (SQNRs) experienced by the users. The formulated problem is convex and can be efficiently solved by available solvers. Our proposed framework offers the alternative option of focused quantization error suppression in accordance with channel state information. Our framework can also be extended to 2D planar transmit antenna arrays. We perform numerical study under different operating conditions, and the numerical results suggest that, given a moderate number of quantization levels, say, 5 to 7 levels, our optimization-based $\Sigma\Delta$ modulation schemes can lead to bit error rate performance close to that of the unquantized counterpart.

INDEX TERMS Massive MIMO downlink, coarsely quantized MIMO, precoding, $\Sigma\Delta$ modulation, convex optimization.

I. INTRODUCTION

Physical-layer or signal-level transceiver techniques have been playing a key part in massive multi-input multi-output (MIMO) communications. They serve the crucial role of physically realizing the promise of massive MIMO, such as substantial gains in spectral efficiency and greatly improved spatial degrees of freedom for serving multiple users [1]. Recent research has focused on how MIMO transceiver techniques can allow us to better cope with practical limitations with the radio frequency (RF) front ends, specifically, issues with the energy efficiency and hardware cost of power amplifiers and analog-to-digital/digital-to-analog converters (ADCs/DACs). Let us narrow down our scope to the ADCs/DACs. We want fine signal resolution to support currently-used transceiver techniques. This calls for high resolution ADCs/DACs being employed at the receiver/transmitter, and a higher resolution means a higher hardware cost and energy consumption. Employing high-resolution ADCs/DACs would not be a serious issue if the MIMO scale (the number of antennas) is small. But, for massive MIMO, the total hardware cost and energy

consumption required by the high-resolution ADCs/DACs will be a burden. One solution is to replace the high-precision converters with lower precision ones [2], [3], [4], [5], [6], [7], [8].

The challenge with using low-resolution ADCs/DACs is that we need to deal with the undesirable error effects caused by coarse quantization. In this paper we are interested in the context of multiuser massive MIMO downlink with low-resolution DACs at the transmitter. It is important to mention that massive MIMO uplink with low-resolution ADCs at the receiver is another key topic; the reader is referred to the literature, such as [2], [3], [4], [5], [9], [10] and the references therein, for details. In coarsely quantized MIMO downlink precoding, the existing studies can be taxonomized into two types, namely, the precode-then-quantize type and the direct signal design type. The precode-then-quantize type takes a precoding scheme in the unquantized case, such as the popularly-used zero-forcing scheme, and then quantizes the precoded signals to produce the few-bit transmitted signals. This approach is straightforward, but the precoding schemes are not designed to resist the adverse effects of quantization errors. Performance analysis for the precode-then-quantize approach has been a subject of interest, helping us better understand the nature of coarsely quantized MIMO; see, e.g., [6], [11], [12]. The direct signal design type seeks to directly manipulate the few-bit signals by optimization, with the aim to optimize some symbol-level performance metric such as mean square error [8] and symbol error probability [13], [14]. Doing so requires us to handle a large-scale discrete optimization problem, which may not be easy. Also, this optimization-oriented approach is, by its nature, unable to leverage our community's rich understanding of MIMO precoding in the unquantized case. That being said, direct signal designs have been empirically found to provide significantly better performance than the precode-then-quantize methods [8], [13], [14], [15], [16], [17]. The advances of direct signal designs are mostly with the one-bit case and with the related context of constant envelope precoding [14], [17], [18], [19]. So far we have not seen direct signal designs for the general multi-bit case, due possibly to the difficulty of such optimization.

The traditional precode-then-quantize approach, which directly quantizes the precoded signals, has no control with the quantization noise. Lately, spatial Sigma-Delta ($\Sigma\Delta$) modulation has arisen as a new precode-then-quantize approach that features quantization noise control or containment [20], [21]. Spatial $\Sigma\Delta$ modulation draws inspiration from temporal $\Sigma\Delta$ modulation in the classical ADC/DAC literature [22]. The basic idea is to add an error feedback loop to the quantizer so that the quantization noise is shaped toward the high frequency band. Consequently, given a low-pass temporal signal, we can convert it to a few-bit signal whose frequency domain sees the signal and quantization noise well separated. In spatial $\Sigma\Delta$ modulation, we turn such noise shaping idea to space. To be specific, we consider a uniform linear transmit antenna array at the base station (BS). We pass the quantization noise of each antenna to the adjacent antenna, thereby forming a spatial $\Sigma\Delta$ feedback loop. This leads to the quantization noise being pushed toward high spatial frequencies, or angles. Consequently we can use a low angle sector to serve users, who will experience reduced quantization noise effects compared to the direct quantization case. While this means that we cannot use the high angle sectors, it is common in practice to consider an angle sector, rather than the full angle range, due to the directivity of antennas. As a precode-then-quantize approach, spatial $\Sigma\Delta$ modulation allows us to use precoding techniques established for the unquantized case—which is a merit. It is worth noting that, recently, spatial $\Sigma\Delta$ modulation has also been considered for MIMO uplink [23], [24], [25], [26], [27].

In the previous study of $\Sigma\Delta$ MIMO downlink [20], [21], the basic first- and second-order modulators from the temporal $\Sigma\Delta$ literature were directly applied to perform spatial $\Sigma\Delta$ modulation. An interesting question is whether we can build $\Sigma\Delta$ modulators that are general, flexible and specifically designed for the context of multiuser massive MIMO downlink precoding. In this paper, we develop a $\Sigma\Delta$ modulator design framework for such a purpose. Our framework considers a general $\Sigma\Delta$ error-feedback structure for any given modulator order and for any given number of quantization levels (or bits). We design $\Sigma\Delta$ modulators by optimization. By characterizing the signal-to-quantization-and-noise ratio (SQNR) experienced by the users, we formulate the $\Sigma\Delta$ modulator designs as some form of SQNR maximization problems. The formulated problems are convex and can be conveniently solved by calling available solvers. Our designs offer two options with quantization noise suppression, namely, (i) quantization noise suppression over a prescribed angle sector; and (ii) focused quantization noise suppression at the user angles, based on the instantaneous channel state information available at the BS. In particular, option (ii) is a new idea. Our framework can also be extended to the 2D uniform planar antenna array setting.

We should describe the relationship of this study to the prior studies in the temporal $\Sigma\Delta$ literature. We commonly see closed-form modulator designs in the temporal $\Sigma\Delta$ literature. While optimization-based modulator designs do not seem to be commonplace in the ADC/DAC literature, our background research found that, curiously, optimization-based modulator designs were considered in the signal processing literature; see [28] and the references therein. In particular, the work by Nagahara and Yamamoto [28] is worth noting, as it provides a convex optimization framework for Chebyshev-type filter designs for $\Sigma\Delta$ noise shaping. As we will elaborate upon in this paper, our spatial $\Sigma\Delta$ modulator designs happen to share some similarities with the temporal $\Sigma\Delta$ modulator designs by Nagahara and Yamamoto. We should however emphasize that, to the best of our knowledge, optimization-based designs have not been previously considered in spatial $\Sigma\Delta$ modulation for coarsely quantized MIMO precoding. Furthermore, our design philosophy differs in that we aim at maximization of the SQNRs experienced by the users, with the MIMO

application aspects taken into consideration, while Nagahara and Yamamoto consider noise shaping.

The organization of this paper is as follows. Section II reviews the background of spatial ΣΔ modulation for coarsely quantized massive MIMO precoding. Section III presents our ΣΔ modulator design framework. Section IV describes the extension of our framework to the 2D uniform planar array case. Section V provides numerical results to show how the ΣΔ modulators designed under our framework perform. Section VI concludes this work.

Our notations are as follows. The symbols $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{N}$ denote the sets of real numbers, complex numbers and non-negative integers, respectively. A scalar, a column vector and a matrix are represented by a lowercase normal letter, a lowercase boldfaced letter and a capital boldfaced letter, respectively; e.g., a, $\boldsymbol{a}$ and $\boldsymbol{A}$, respectively. The real and imaginary parts of a given vector $\boldsymbol{a}$ are denoted by $\Re(\boldsymbol{a})$ and $\Im(\boldsymbol{a})$, respectively. The transpose of a vector $\boldsymbol{a}$ is denoted by $\boldsymbol{a}^T$, and the same convention applies to matrices. The trace, inverse and pseudo-inverse of a matrix $\boldsymbol{A}$ are denoted by $\mathrm{tr}(\boldsymbol{A})$, $\boldsymbol{A}^{-1}$ and $\boldsymbol{A}^{\dagger}$, respectively. Given a vector $\boldsymbol{a}$, the notation $\mathrm{Diag}(\boldsymbol{a})$ denotes a diagonal matrix with the (i,i)th component given by the ith component of $\boldsymbol{a}$. Given a collection of scalars $a_1,\ldots,a_n$, the notation $(a_1,\ldots,a_n)$ denotes the concatenation of the a_i's as a vector, i.e., $(a_1,\ldots,a_n) = [\,a_1,\ldots,a_n\,]^T$. The same convention applies when the a_i's are vectors. We denote $\mathrm{j} = \sqrt{-1}$. Given a sequence $\{a_n\}_{n\in\mathcal{N}}$, where $\mathcal{N}$ equals either $\mathbb{N}$ or $\{0,1,\ldots,N-1\}$ for some positive integer N, the Fourier transform of $\{a_n\}_{n\in\mathcal{N}}$ is denoted by $A(\omega) = \sum_{n\in\mathcal{N}} a_n e^{-\mathrm{j}n\omega}$. Given a vector $\boldsymbol{a}$, the notations $\|\boldsymbol{a}\|_1$, $\|\boldsymbol{a}\|_2$ and $\|\boldsymbol{a}\|_\infty$ denote the 1-norm, Euclidean norm and ∞-norm of $\boldsymbol{a}$, respectively. Given a complex vector $\boldsymbol{a}$, the notations $\|\boldsymbol{a}\|_{\mathrm{IQ}-1}$ and $\|\boldsymbol{a}\|_{\mathrm{IQ}-\infty}$ denote the 1-norm and ∞-norm with respect to $(\Re(\boldsymbol{a}),\Im(\boldsymbol{a}))$, respectively; i.e., $\|\boldsymbol{a}\|_{\mathrm{IQ}-1} = \|(\Re(\boldsymbol{a}),\Im(\boldsymbol{a}))\|_1$ and $\|\boldsymbol{a}\|_{\mathrm{IQ}-\infty} = \|(\Re(\boldsymbol{a}),\Im(\boldsymbol{a}))\|_\infty$. The same conventions apply to matrices.

II. BACKGROUND

This section intends to provide the background of this study. We review the basics of ΣΔ modulation in the first subsection, give the problem statement of coarsely quantized MIMO precoding in the second subsection, and describe the spatial ΣΔ modulation approach for the precoding problem in the third subsection.

A. ΣΔ MODULATION

We introduce the basics of ΣΔ modulation by considering the one-bit first-order modulator, the most basic scheme in ΣΔ modulation. The system architecture of the modulator is depicted in Fig. 1. Let $\{\bar{x}_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ be a real-valued time sequence. Let $\mathrm{sgn}:\mathbb{R}\to\{\pm1\}$ be the signum function. The modulator takes $\{\bar{x}_n\}_{n\in\mathbb{N}}$ as the input and generates a binary output $\{x_n\}_{n\in\mathbb{N}} \subset \{\pm1\}$ by

$$x_n = \mathrm{sgn}(\bar{x}_n - q_{n-1}) = \bar{x}_n - q_{n-1} + q_n,\ n\in\mathbb{N}, \qquad (1)$$

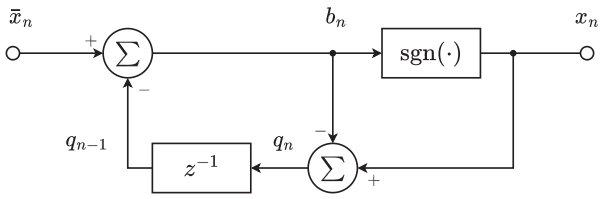


FIGURE 1. One-bit first-order ΣΔ modulator.

where q_n is the quantization error associated with $\bar{x}_n - q_{n-1}$, for $n\in\mathbb{N}$; and we have $q_{-1} = 0$. The rationale of this process should be described. The input $\{\bar{x}_n\}_{n\in\mathbb{N}}$ is a lowpass temporal signal. We want to coarsely quantize the input in such a way that the error signal at the output is weak in the low frequency band. We make the following assumption which is used in nearly every ΣΔ literature.

Assumption 1: *Consider the modulator in Fig. 1 or the system in (1). Each quantization error q_n is $[-1,1]$-supported, uniformly distributed on its support, and independent of any other random variables.*

Let $v_n = q_n - q_{n-1}$ be the error at the output x_n. The magnitude spectrum of $\{v_n\}_{n\in\mathbb{N}}$ equals

$$|V(\omega)|^2 = |Q(\omega) - e^{-\mathrm{j}\omega}Q(\omega)|^2 = |1-e^{-\mathrm{j}\omega}|^2|Q(\omega)|^2,$$

where $|1-e^{-\mathrm{j}\omega}|^2 = 4|\sin(\omega/2)|^2$ is a highpass response. Also, under Assumption 1 we can see $|Q(\omega)|^2$ as a flat spectrum; more precisely, the power spectral density of $\{q_n\}_{n\in\mathbb{N}}$ is flat. Hence, the modulator can be viewed as a quantizer that has the ability to shape the quantization error signal as highpass noise, and by doing so we reduce the undesirable interference effects of the quantization errors on the lowpass input signal over the low frequency band.

In Assumption 1 we assume that every quantization error q_n is bounded, lying in $[-1,1]$. We want to discuss how this can be guaranteed. It can be easily shown that if the pre-quantized signal $b_n = \bar{x}_n - q_{n-1}$ has amplitude greater than 2, then the associated quantization error q_n will have $|q_n| > 1$—such phenomena are called overloading in the literature. Overloading can lead to large q_n in terms of the amplitude, and mathematically one can show that there exists an input $\{\bar{x}_n\}_{n\in\mathbb{N}}$ such that $|q_n|\to\infty$ as $n\to\infty$ [20]. Overloading can be prevented by restricting the input to be amplitude limited:

Fact 1: *Consider the modulator in Fig. 1 or the system in (1). Let $A>0$ be the maximum input amplitude, i.e., $|\bar{x}_n|\le A$ for all n. If $A\le 1$, then $|q_n|\le 1$ for all n.*

The proof of Fact 1 is simple: Suppose $|q_{n-1}|\le 1$. Then $|\bar{x}_n - q_{n-1}| \le A+1 \le 2$, and we have $|q_n|\le 1$. The proof is complete.

The one-bit first-order ΣΔ modulation scheme introduced above is basic. There are many other ΣΔ modulation schemes, as well as a variety of aspects related to the modulator designs. We refer the reader to the literature (e.g., [22], [29]) for details such as the multi-level and higher-order generalizations of the above ΣΔ modulator; the various ΣΔ

modulator architectures; reasonability of the independent and identical distributed (i.i.d.) assumption in Assumption 1 in practice, and the practical trick of dithering to try to make the quantization error more i.i.d.; and the impact of overloading in practice. We also refer the reader to the mathematical studies in [30], [31], which analyze the reconstruction accuracy of temporal $\Sigma\Delta$ modulation schemes without Assumption 1.

Before we finish our review, we should note a basic implementation aspect. For DACs with $\Sigma\Delta$ modulation (the case of interest here), the modulators appear in digital domain and can be flexibly implemented by digital signal processing. For ADCs (outside the scope of this study), $\Sigma\Delta$ modulators are implemented in analog domain and require dedicated analog and digital hardware to build.

B. COARSELY QUANTIZED MIMO PRECODING

Consider the following multiuser MIMO downlink communication problem. The base station (BS) serves a number of K users and has N transmit antennas. The users have a single antenna. Assuming frequency-flat time-invariant channels over a finite time frame of transmission, the transmit-receive relation from the BS to the users is modeled as

$$y_{i,t} = \sqrt{\rho}\boldsymbol{h}_i^T \boldsymbol{x}_t + \eta_{i,t},\ t = 1, \ldots, T, \tag{2}$$

where $y_{i,t}$ is the received signal of user i at symbol time t; $\sqrt{\rho}\boldsymbol{x}_t \in \mathbb{C}^N$ is the transmitted signal at symbol time t, with its nth component $\sqrt{\rho}x_{n,t}$ being the transmitted signal at the nth antenna; $\boldsymbol{x}_t$ is the transmitted signal before power amplification; $\rho > 0$ is a power scaling factor; $\eta_{i,t}$ is i.i.d. circular complex Gaussian noise with mean zero and variance σ_η^2; T is the transmission block length. Assume that the BS is informed of $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K$. The problem, called precoding, is to design the transmitted signals $\{\boldsymbol{x}_t\}_{t=1}^T$ such that each user will receive its own data symbol stream with minimal distortions. Specifically, we want the noise-free part of $y_{i,t}$ to take the form

$$\boldsymbol{h}_i^T \boldsymbol{x}_t \approx c_i s_{i,t}, \tag{3}$$

where $\{s_{i,t}\}_{t=1}^T$ is the symbol stream for user i; c_i represents the signal gain. For example, the zero-forcing (ZF) scheme performs precoding by

$$\boldsymbol{x}_t = \boldsymbol{H}^\dagger \boldsymbol{s}_t,\ t = 1, \ldots, T,$$

where $\boldsymbol{H} = [\,\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K\,]^T$; $\boldsymbol{s}_t = (s_{1,t}, \ldots, s_{K,t})$. It is easy to see that the ZF scheme leads to $y_{i,t} = \sqrt{\rho}s_{i,t} + \eta_{i,t}$.

In precoding, it is common to assume that the transmitted signals $x_{n,t}$'s are continuous valued. The problem of interest in this paper is coarsely quantized precoding, wherein the $x_{n,t}$'s are discrete valued. For example, for the one-bit case, the real and imaginary components of every $x_{n,t}$ are binary. The motivation, as discussed in the Introduction, is to reduce massive MIMO hardware costs and power consumption by replacing high-resolution DACs with low-resolution ones. A straightforward solution to coarsely quantized precoding is to directly quantize the precoded signals. For example, we can directly quantize the ZF scheme by $\boldsymbol{x}_t = \mathcal{Q}_c(\boldsymbol{H}^\dagger \boldsymbol{s}_t)$, where $\mathcal{Q}_c$ denotes the quantization function associated with the low-resolution DACs. But such a precode-then-quantized scheme can significantly suffer from quantization error effects. Some recent studies seek a different approach, namely, by directly optimizing the discrete variables $x_{n,t}$'s to shape symbols at the user side (cf. (3)) [8], [13], [15], [16], [17]. This direct design approach was found to be able to provide promising performance by numerical experiments. It however requires us to solve a large-scale discrete optimization problem. Also, its optimization-oriented design principle largely disallows us from reusing precoding concepts in the unquantized case, such as the simple ZF scheme.

C. SPATIAL $\Sigma\Delta$ MODULATION

We recently proposed a spatial $\Sigma\Delta$ modulation approach for the above stated coarsely quantized precoding problem [20]. It falls into the precode-then-quantized scope, and the spirit is to use $\Sigma\Delta$ modulation to shape the noise spectrum—spatially—such that users are less affected by the quantization error effects. The spatial $\Sigma\Delta$ modulation approach is described as follows. We assume angular channels

$$\boldsymbol{h}_i = \alpha_i \boldsymbol{a}(\theta_i),$$

where $\alpha_i \in \mathbb{C}$ is a complex channel gain; $\theta_i \in (-\pi/2, \pi/2)$ is the user angle;

$$\boldsymbol{a}(\theta) = (1, e^{-\mathrm{j}\frac{2\pi d}{\lambda}\sin(\theta)}, \ldots, e^{-\mathrm{j}(N-1)\frac{2\pi d}{\lambda}\sin(\theta)})$$

is the angular response, with λ being the carrier wavelength, $d \le \lambda/2$ being the inter-antenna spacing, and $\theta \in (-\pi/2, \pi/2)$ being the angle. The angular channels are based on the operating assumptions that the transmit antennas are arranged as a uniform linear array, and we consider a single-path far-field channel from the BS to each user; the reader is referred to the literature (e.g., [32]) for details. Consider, at each symbol time t, that we apply the $\Sigma\Delta$ modulator in Section II-A to the transmitted signals. To be careful, let $\bar{\boldsymbol{x}}_t$ and $\boldsymbol{x}_t$ be the transmitted signals before and after $\Sigma\Delta$ modulation, respectively. Also, as a slight abuse of notations, let $\bar{x}_{n,t}$ and $x_{n,t}$ denote the $(n+1)$th elements of $\bar{\boldsymbol{x}}_t$ and $\boldsymbol{x}_t$, respectively. We apply the one-bit first-order $\Sigma\Delta$ modulator in Section II-A to $\{\Re(\bar{x}_{n,t})\}_{n=0}^{N-1}$ to obtain $\{\Re(x_{n,t})\}_{n=0}^{N-1}$, and we apply another one-bit first-order $\Sigma\Delta$ modulator to $\{\Im(\bar{x}_{n,t})\}_{n=0}^{N-1}$ to obtain $\{\Im(x_{n,t})\}_{n=0}^{N-1}$. The appealing result goes as follows: for any $\theta \in (-\pi/2, \pi/2)$,

$$\begin{aligned}\boldsymbol{a}(\theta)^T \boldsymbol{x}_t &= \boldsymbol{a}(\theta)^T \bar{\boldsymbol{x}}_t + \sum_{n=0}^{N-1}(q_{n,t} - q_{n-1,t})e^{-\mathrm{j}n\omega} \\ &\simeq \boldsymbol{a}(\theta)^T \bar{\boldsymbol{x}}_t + (1 - e^{\mathrm{j}\omega})Q_t(\omega),\end{aligned} \tag{4}$$

where $\omega = \frac{2\pi d}{\lambda}\sin(\theta)$; $\{q_{n,t}\}_{n=0}^{N-1} \subset \mathbb{C}$ is the quantization error sequence; $Q_t(\omega)$ is the Fourier transform of $\{q_{n,t}\}_{n=0}^{N-1}$. In the second equation in (4), we assume that N is large, and we will continue to assume this without explicit mentioning. We observe from (4) that the quantization error term is a

highpass response—its magnitude is expected to be smaller if the frequency ω, or its respective angle θ, is closer to 0.

The above observation suggests the following possibility: Consider a sectored antenna array setting wherein we serve users within a lowpass angle sector, say, $[-30^\circ, 30^\circ]$. Then, the spatial $\Sigma\Delta$ modulation introduced above can lead to reduced quantization error effects on the users. Specifically, by plugging (4) into the signal model (2), we see that the received signals can be written as

$$y_{i,t} = \sqrt{\rho}\boldsymbol{h}_i^T \bar{\boldsymbol{x}}_t + \sqrt{\rho}\alpha_i \underbrace{(1 - e^{\mathrm{j}\omega_i})Q_t(\omega_i)}_{:=v_{i,t}} + \eta_{i,t}, \qquad (5)$$

where $\omega_i = \frac{2\pi d}{\lambda}\sin(\theta_i)$. By applying Assumption 1 to the real and imaginary components of $q_{n,t}$, it can be shown that the power of the quantization noise term $v_{i,t}$ is

$$\mathbb{E}[|v_{i,t}|^2] = |1 - e^{\mathrm{j}\omega_i}|^2 \frac{2N}{3} = 4\left|\sin\left(\frac{\pi d}{\lambda}\sin(\theta_i)\right)\right|^2 \frac{2N}{3},$$

which reduces with $|\theta_i|$. Note that we can also reduce the quantization noise power by reducing the inter-antenna spacing d, but in practice we cannot make d too small due to mutual coupling effects; the reader is referred to our previous work [20] for further discussion.

It is also necessary to describe the precoding part of the spatial $\Sigma\Delta$ modulation approach. The idea is nothing more than treating the second and third term on the right-hand side of the received signal model (5) as a single noise term, and then designing $\{\bar{\boldsymbol{x}}_t\}_{t=1}^T$ by an existing unquantized precoding scheme. But there is a new constraint unique to spatial $\Sigma\Delta$ modulation. To guarantee no overloading with the $\Sigma\Delta$ modulator, it is suggested by Fact 1 that we should limit the amplitude of the real and imaginary components of $\bar{\boldsymbol{x}}_t$, specifically,

$$\|\bar{\boldsymbol{x}}_t\|_{\mathrm{IQ}-\infty} := \max\{\|\Re(\bar{\boldsymbol{x}}_t)\|_\infty, \|\Im(\bar{\boldsymbol{x}}_t)\|_\infty\} \le 1, \qquad (6)$$

for all t. Hence, the precoding problem in spatial $\Sigma\Delta$ modulation is an amplitude-limited unquantized precoding problem, which is still not exactly the same as the popular unquantized precoding problem which typically considers average power constraints. But some precoding schemes can be easily modified to fit into the amplitude-limited case. For example, for the ZF scheme, we can do normalization

$$\bar{\boldsymbol{x}}_t = \frac{\boldsymbol{H}^\dagger \boldsymbol{s}_t}{C},\ t = 1, \ldots, T, \qquad (7)$$

where $C = \max_{t=1,\ldots,T} \|\boldsymbol{H}^\dagger \boldsymbol{s}_t\|_{\mathrm{IQ}-\infty}$, such that the amplitude constraints $\|\bar{\boldsymbol{x}}_t\|_{\mathrm{IQ}-\infty} \le 1$ are satisfied [20]. The reader is referred to our previous work [20] for more amplitude-limited precoding designs.

III. GENERAL AND FLEXIBLE DESIGNS FOR SPATIAL $\Sigma\Delta$ MODULATION

In our previous study with the spatial $\Sigma\Delta$ modulation approach, we mainly applied an existing $\Sigma\Delta$ modulator; we used the one-bit first-order modulator in [20], and later we adopted the two-bit second-order modulator in the $\Sigma\Delta$ literature [21]. From this section we set our sight on designing our own $\Sigma\Delta$ modulator. The study to be described revolve around the following questions.

1) Can we have a general and flexible design for $\Sigma\Delta$ modulation of any quantization level number and of any order?
2) Can we make the designs a better fit to coarsely quantized MIMO precoding, specifically, by explicitly working on the signal-to-quantization-and-noise ratios (SQNRs)?
3) Given an angle sector $[\theta_l, \theta_u] \subset (-\pi/2, \pi/2)$, a modulator order L, and a quantization level number M, can we design a $\Sigma\Delta$ modulator that works better than the standard $\Sigma\Delta$ modulators in the $\Sigma\Delta$ literature?
4) Can we lift the angle sector restriction and allow users to freely lie in any angles?

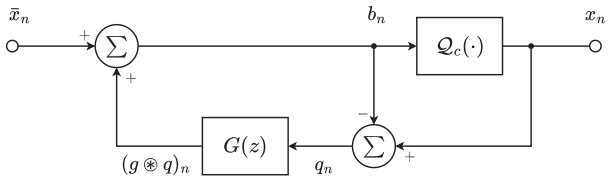


FIGURE 2. General $\Sigma\Delta$ modulator.

A. A GENERAL $\Sigma\Delta$ MODULATOR STRUCTURE

We consider a multi-level, higher-order and complex-valued generalization of the one-bit first-order $\Sigma\Delta$ modulator in Section II-A. The system architecture is depicted in Fig. 2. It should be noted that this generalized structure was mentioned or considered in the literature [22], [28], often for the real-valued case. The rationale of this modulator is identical to that of its predecessor in Section II-A, and we shall be concise with our description. The input $\{\bar{x}_n\}_{n\in\mathbb{N}}$ is a complex-valued sequence. The function $\mathcal{Q}_c$ applies M-level quantization to the real and imaginary components. To be specific, let

$$\mathcal{X} = \begin{cases} \{\pm 1, \pm 3, \ldots, \pm(M-1)\}, & M \text{ is even} \\ \{0, \pm 2, \ldots, \pm(M-1)\}, & M \text{ is odd} \end{cases} \qquad (8)$$

be the multi-level signal set, and let $\mathcal{Q} : \mathbb{R} \to \mathcal{X}$ be the quantizer associated with $\mathcal{X}$. The quantizer $\mathcal{Q}_c$ is given by $\mathcal{Q}_c(x) = \mathcal{Q}(\Re(x)) + \mathrm{j}\,\mathcal{Q}(\Im(x))$. The error feedback is given by

$$(g \circledast q)_n = \sum_{l=1}^{L} g_l q_{n-l},$$

which is the convolution of the quantization error sequence $\{q_n\}_{n\in\mathbb{N}}$ and an impulse response $\{g_l\}_{l=1}^L$ of a filter. The filter coefficients $g_1, \ldots, g_L$ are complex-valued and are to be designed. The input-output relation of the modulator is

$$x_n = \mathcal{Q}_c(\bar{x}_n + (g \circledast q)_n) = \bar{x}_n + (g \circledast q)_n + q_n,\ n \in \mathbb{N}. \qquad (9)$$

We assume that

Assumption 2: *Consider the modulator in Fig. 2 or the system in (9). Each quantization error component $\Re(q_n)$ or $\Im(q_n)$ is*

$[-1, 1]$-*supported, uniformly distributed on its support, and independent of any other random variables.*

Let $v_n = (g \circledast q)_n + q_n$ be the quantization noise term at the output. Its magnitude spectrum is

$$|V(\omega)|^2 = |1 + G(\omega)|^2 |Q(\omega)|^2,$$

where the response $1 + G(\omega)$ plays the key role of shaping the noise magnitude spectrum.

Furthermore, we are concerned with overloading. We will adopt the following no-overload condition.

Fact 2: *Consider the modulator in Fig. 2 or the system in (9). Let $A > 0$ be the maximum input amplitude, specifically, $|\bar{x}_n|_{lQ-\infty} := \max\{|\Re(\bar{x}_n)|, |\Im(\bar{x}_n)|\} \le A$ for all n. Let $\boldsymbol{g} = (g_1, \ldots, g_L)$, and $\|\boldsymbol{g}\|_{lQ-1} = \sum_{i=1}^{L} |\Re(g_i)| + |\Im(g_i)|$. If*

$$A + \|\boldsymbol{g}\|_{lQ-1} \le M,$$

then $|q_n|_{lQ-\infty} \le 1$ for all n.

Fact 2 is the multi-level higher-order complex-valued counterpart of Fact 1. The result is considered known in the literature; see e.g., [33], [22, Section 4.2.2], [28], [31] for the real-valued case. We provide the proof of Fact 2 in Appendix A for the reader's reference.

To give the reader some insight, we show some examples covered by the general $\Sigma\Delta$ modulator.

Example 1 (first-order modulator): *Consider $L = 1$, $g_1 = -1$. This is the previously studied first-order modulator, which has a noise shaping response $1 + G(\omega) = 1 - e^{-\mathrm{j}\omega}$. According to Fact 2, the modulator is guaranteed to have no overload if $A \le M - 1$.*

Example 2 (second-order modulator): *Consider $L = 2$, $g_1 = -2$, $g_2 = 1$. This modulator is called the second-order modulator in the $\Sigma\Delta$ literature. It has a shaping response $1 + G(\omega) = (1 - e^{-\mathrm{j}\omega})^2$, which is a stronger highpass response than the first-order. By Fact 2, the modulator has no overload if $A \le M - 3$. This further implies that the second-order modulator requires at least $M = 4$, or two bits, to achieve the no-overload condition.*

Example 3 (frequency-shifted modulator): *Consider $L = 1$, $g_1 = -e^{\mathrm{j}\omega_c}$ for some given frequency ω_c. The shaping response is $1 + G(\omega) = 1 - e^{-\mathrm{j}(\omega - \omega_c)}$, which is a band-stop response centered at ω_c [20]. By Fact 2, the modulator has no overload if $A \le M - |\sin(\omega_c)| - |\cos(\omega_c)|$, or, more conservatively, if $A \le M - \sqrt{2}$.*

B. SPATIAL $\Sigma\Delta$ MODULATION BY THE GENERAL STRUCTURE

Consider the spatial $\Sigma\Delta$ modulation for coarsely quantized MIMO precoding in Sections II-B and II-C. We want to replace the previous one-bit first-order $\Sigma\Delta$ modulator by the general $\Sigma\Delta$ modulator in the last subsection, specifically, by applying the general $\Sigma\Delta$ modulator to $\{\bar{x}_{n,t}\}_{n=0}^{N-1}$ to yield $\{x_{n,t}\}_{n=0}^{N-1}$. Following the same derivations in Sections II-B and II-C, we can show that the received signals can be modeled as

$$y_{i,t} = \sqrt{\rho}\boldsymbol{h}_i^T \bar{\boldsymbol{x}}_t + \sqrt{\rho}\alpha_i \underbrace{(1 + G(\omega_i))Q_t(\omega_i)}_{:=v_{i,t}} + \eta_{i,t}, \qquad (10)$$

where we recall $\omega_i = \frac{2\pi d}{\lambda}\sin(\theta_i)$. Also, by applying Assumption 2 to $\{q_{n,t}\}_{n=0}^{N-1}$, the quantization noise power is

$$\mathbb{E}[|v_{i,t}|^2] = |1 + G(\omega_i)|^2 \frac{2N}{3}. \qquad (11)$$

Our problems, to be studied in the subsequent subsections, is to design the filter coefficients $g_1, \ldots, g_L$ such that the quantization noise powers of the users are mitigated, while, at the same time, the no-overload condition in Fact 2 is satisfied.

C. ZERO QUANTIZATION NOISE?

It is natural to question this: Can we have zero quantization noise for all the users? In raising this question, we allow the angle θ_i of each user to lie freely in the admissible angle region $(-\pi/2, \pi/2)$. Achieving zero quantization noise means that $1 + G(\omega_i) = 0$ for all i, and this can be made possible by setting the shaping response as

$$1 + G(\omega) = \prod_{k=1}^{K} (1 - e^{-\mathrm{j}(\omega - \omega_k)}). \qquad (12)$$

It should be noted that $1 + G(\omega)$ produces a multiple notch filter response, with nulls placed at the ω_k's. The above shaping response corresponds to a K-th order $\Sigma\Delta$ modulator with coefficients

$$g_k = \sum_{1 \le i_1 < \cdots < i_k \le K} (-e^{\mathrm{j}\omega_{i_1}}) \cdots (-e^{\mathrm{j}\omega_{i_k}}), \qquad (13)$$

for $k = 1, \ldots, K$. This zero quantization noise design, however, has a serious limitation.

Proposition 1: *Consider the $\Sigma\Delta$ modulator with coefficients given by (13), which achieves zero quantization noise with the users. We have the following results.*

1) *It holds that*

$$\|\boldsymbol{g}\|_{lQ-1} \le \sqrt{2}(2^K - 1),$$

and equality is attained when $\omega_1 = \cdots = \omega_K \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$.

2) *As a simplifying assumption, assume each ω_i to be i.i.d. uniformly distributed on $(-\pi, \pi)$. Then,*

$$2^{(K-1)/2} \le \sqrt{\mathbb{E}[\|\boldsymbol{g}\|_{lQ-1}^2]} \le 2^K.$$

We show the proof of Proposition 1 in Appendix B. Proposition 1 suggests that $\|\boldsymbol{g}\|_{lQ-1}$ may increase exponentially with K. To give some idea, Table 1 shows some empirical evaluation results for $\|\boldsymbol{g}\|_{lQ-1}$.

We observe that the empirical results are in agreement with our theoretical prediction. By also considering the no-overload requirement in Fact 2, we see the following implication: the quantization level number M may need to increase

TABLE 1. Minimum, Mean, Root Mean Square (RMS), and Maximum Values of $\|g\|_{\mathrm{IQ}-1}$ for (13) and for a Number of Randomly Generated ω_i's. The ω_i's are I.i.d. $(-\pi, \pi)$-Uniform Distributed

$\|g\|_{\mathrm{IQ}-1}$ \ K	2	3	4	5	6	7	8
min.	1.00006	1.078	1.23	1.26	1.86	2.53	3.09
mean	2.89	5.28	8.37	12.85	18.83	27.17	38.06
RMS	3.01	5.60	9.27	14.72	22.53	33.78	50.37
max.	4.09	9.46	20.23	41.33	81.96	160.34	315.88

exponentially with the number of users K to achieve zero quantization noise at the user side.

D. SQNR MAXIMIZATION IN A USER TARGETED FASHION

Since zero quantization noise is practically infeasible even for a moderate number of users, we turn to the alternative of maximizing the SQNRs experienced by the users. Note that, as in the previous problem, we allow the user angles θ_i's to freely lie in $(-\pi/2, \pi/2)$. Our tasks are divided into three parts: define a suitable SQNR for the problem at hand, properly formulate the $\Sigma\Delta$ modulator design as an optimization problem, and develop a solution.

We start with the SQNR. From the received signal model (10), we see that the signal part $\sqrt{\rho}\boldsymbol{h}_i^T\bar{\boldsymbol{x}}_t$ scales with $\sqrt{\rho}|\alpha_i|A$. Here, it is important to note that A describes the maximum input signal amplitude, i.e., $\|\bar{\boldsymbol{x}}_t\|_{\mathrm{IQ}-\infty} \le A$ for all t. We define the SQNR of user i as the ratio of the square of the received signal scale factor $\sqrt{\rho}|\alpha_i|A$ to the quantization and noise power, which can be shown to be

$$\mathrm{SQNR}_i = \frac{\rho|\alpha_i|^2 A^2}{\frac{2N\rho|\alpha_i|^2}{3}|1+G(\omega_i)|^2 + \sigma_\eta^2}. \tag{14}$$

Next, we formulate the $\Sigma\Delta$ modulator design. Our underlying assumption is that the BS is informed of the channels $\boldsymbol{h}_i$'s, or, the complex gains α_i's and angles θ_i's of all the users. The BS is assumed to know the background noise power σ_η^2, too. Also, the modulator order L and the quantization level number M of the $\Sigma\Delta$ modulator are prespecified. We design the $\Sigma\Delta$ filter coefficients by the max-min-fair criterion, subject to the no-overload condition in Fact 2:

$$\max_{\boldsymbol{g}\in\mathbb{C}^L, A\in\mathbb{R}} \min_{i=1,\ldots,K} \mathrm{SQNR}_i$$
$$\text{s.t.}\ A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M,\ A > 0. \tag{15}$$

Here, fairness is achieved by maximizing the weakest user's SQNR, thereby sacrificing no one in the interest of others. It is worth noting that we also optimize the maximum input signal amplitude A, rather than prefixing it, to give the design more degrees of freedom.

The max-min-fair design (15) can be converted to a convex problem and can be efficiently solved. To see how this is done, we substitute (14) into problem (15) and rewrite the problem as

$$\min_{\boldsymbol{g}\in\mathbb{C}^L, A\in\mathbb{R}} \max_{i=1,\ldots,K} \frac{\sqrt{|1+G(\omega_i)|^2+\gamma_i}}{A}$$
$$\text{s.t.}\ A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M,\ A > 0, \tag{16}$$

where $\gamma_i = 3\sigma_\eta^2/(2N\rho|\alpha_i|^2)$. Problem (16) is quasi-convex, but not convex. Consider the following transformation

$$\boldsymbol{v} = \boldsymbol{g}/A,\ \xi = 1/A, \tag{17}$$

which is known as the Charnes-Cooper transformation in optimization [34]. The transformation (17) is one-to-one if A and ξ are positive. Using (17), problem (16) can be transformed as

$$\min_{\boldsymbol{v}\in\mathbb{C}^L, \xi\in\mathbb{R}} \max_{i=1,\ldots,K} \sqrt{|\xi + \boldsymbol{a}(\omega_i)^T\boldsymbol{v}|^2 + \gamma_i\xi^2}$$
$$\text{s.t.}\ 1 + \|\boldsymbol{v}\|_{\mathrm{IQ}-1} \le M\xi,\ \xi > 0, \tag{18}$$

where we redefine $\boldsymbol{a}(\omega) = (1, e^{-\mathrm{j}\omega}, \ldots, e^{-\mathrm{j}(N-1)\omega})$. Moreover, problem (18) is equivalent to

$$\min_{\boldsymbol{v}\in\mathbb{C}^L, \xi\in\mathbb{R}} \max_{i=1,\ldots,K} \sqrt{|\xi + \boldsymbol{a}(\omega_i)^T\boldsymbol{v}|^2 + \gamma_i\xi^2}$$
$$\text{s.t.}\ 1 + \|\boldsymbol{v}\|_{\mathrm{IQ}-1} \le M\xi,\ \xi \ge 0, \tag{19}$$

where we replace $\xi > 0$ with $\xi \ge 0$. This is without loss, because the first constraint in (19) implies $1 \le M\xi$, and with the second constraint $\xi \ge 0$ we further get $\xi \ge 1/M > 0$. Problem (19) is convex, and its solution can be conveniently and efficiently obtained by using a convex optimization software, such as the widely-used CVX [35].

Let us provide some numerical illustration. Fig. 3 shows the relative noise shaping responses, defined as

$$\mathrm{RNSR}(\theta) = \frac{|1+G(\frac{2\pi d}{\lambda}\sin(\theta))|^2}{A^2}, \tag{20}$$

of the user-targeted design (15). The red vertical lines in the figure indicate the user angles, and the system settings are $N = 1024$, $K = 6$, $L = 16$, $d = \lambda/4$, $|\alpha_i| = 1$ for all i, and $\sigma_\eta^2 = 0$. We see that, as the quantization level number M increases, the user-targeted design provides sharper notches, and therefore better quantization noise suppression, at the user angles.

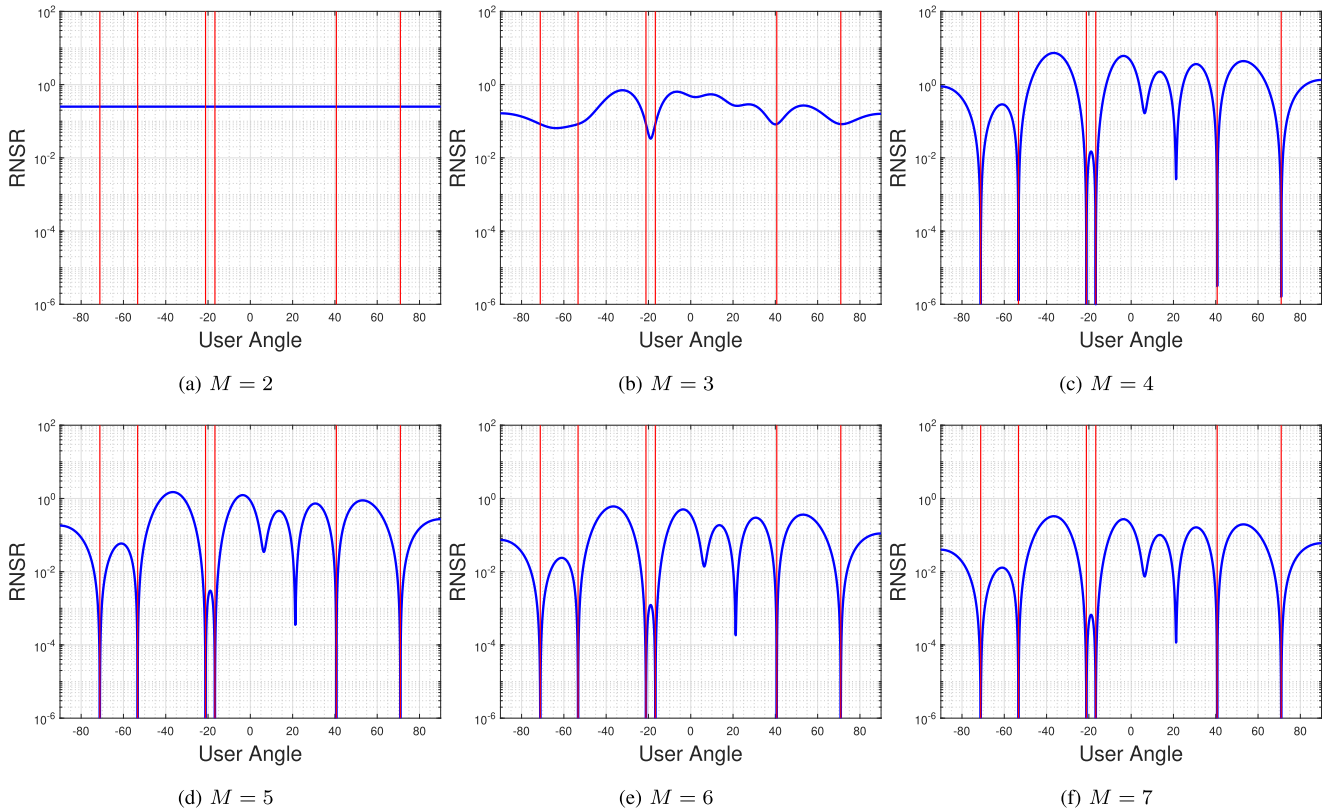


FIGURE 3. Illustration of the relative noise-shaping responses of the ΣΔ modulators designed by the user-targeted formulation (15). Red line: user angles.

E. SQNR MAXIMIZATION FOR A FIXED ANGLE SECTOR

The user-targeted SQNR maximization design in the last subsection assumes that we can change the ΣΔ modulator whenever the user angles θ_i's and channel gains $|\alpha_i|^2$ change. Suppose that we are prohibited to do so due to implementation reasons, and we can only re-design the ΣΔ modulator once in a while. We hence return to the angle sector setting wherein we serve users in a prespecified angle sector $[\theta_l, \theta_u] \subset (-\pi/2, \pi/2)$ (which can be lowpass or bandpass). Our problem is to adapt the preceding ΣΔ modulator design to this fixed sector setting.

We start with off-the-shelf designs from the ΣΔ literature. Consider the following shaping response

$$1 + G(\omega) = (1 - e^{-\mathrm{j}(\omega-\omega_c)})^L \tag{21}$$

for a given positive integer L, where $\omega_c = (\omega_l + \omega_u)/2$, $\omega_l = \frac{2\pi d}{\lambda}\sin(\theta_l)$, $\omega_u = \frac{2\pi d}{\lambda}\sin(\theta_u)$. This is a band-stop response with center frequency ω_c. The corresponding coefficients are

$$g_l = \binom{L}{l}(-e^{\mathrm{j}\omega_c})^l,\ l = 1, \ldots, L. \tag{22}$$

This modulator is essentially the combination of the standard L-th order modulator and the frequency-shifted modulator; see Examples 2–3. Increasing the order L makes the band-stop response sharper, but this comes with a limitation.

Proposition 2: *Consider the ΣΔ modulator with coefficients given by (22), and with the shaping response given by (21). We have*

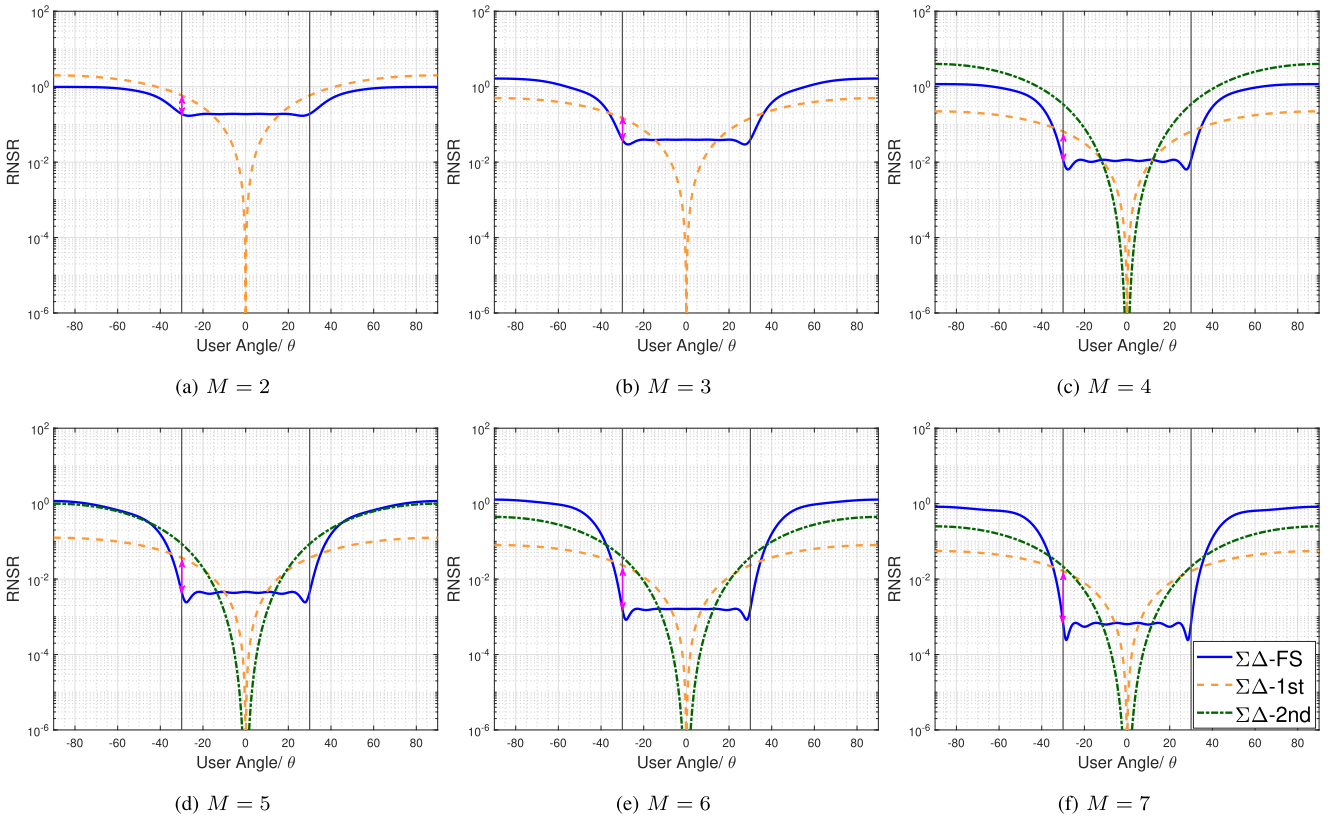
$$2^L - 1 \le \|\boldsymbol{g}\|_{lQ-1} \le \sqrt{2}(2^L - 1).$$

The proof of Proposition 2 is shown in Appendix B. Proposition 2, together with Fact 2, indicate that the quantization level number M needs to increase exponentially with L to achieve the no-overload condition.

Alternatively, we can repurpose the SQNR-based design in Section III-D. Suppose that the channel gains $|\alpha_i|$'s are known to lie in a range $[r_{\min}, r_{\max}]$. Our goal is to design the ΣΔ modulator in accordance with the prespecified angle sector $[\theta_l, \theta_u]$ and the channel gain range $[r_{\min}, r_{\max}]$. Following the SQNR definition in (14), a user with angle $\theta \in [\theta_l, \theta_u]$ and channel gain $|\alpha| \in [r_{\min}, r_{\max}]$ will experience an SQNR

$$\begin{aligned}\mathrm{SQNR} &= \frac{\rho|\alpha|^2 A^2}{\frac{2N\rho|\alpha|^2}{3}|1+G(\omega)|^2 + \sigma_\eta^2} \\ &\ge \frac{\rho r_{\min}^2 A^2}{\frac{2N\rho r_{\max}^2}{3}|1+G(\omega)|^2 + \sigma_\eta^2} \\ &:= \widetilde{\mathrm{SQNR}}(\omega),\end{aligned}$$

where $\omega = \frac{2\pi d}{\lambda}\sin(\theta)$. With the above expression, we consider the following adaptation of the max-min-fair design in



(a) $M = 2$ (b) $M = 3$ (c) $M = 4$

(d) $M = 5$ (e) $M = 6$ (f) $M = 7$

FIGURE 4. Illustration of the relative noise-shaping responses of the ΣΔ modulators designed by the fixed sector formulation (24). Black line: the boundary of the angle sector $[\theta_l, \theta_u]$. ΣΔ-FS: the fixed-sector design (24). ΣΔ-1st: the standard first-order ΣΔ modulator. ΣΔ-2nd: the standard second-order ΣΔ modulator.

(15) to the angle sector setting:

$$\max_{\boldsymbol{g}\in\mathbb{C}^L, A\in\mathbb{R}} \min_{\omega\in[\omega_l,\omega_u]} \widetilde{\mathrm{SQNR}}(\omega)$$
$$\text{s.t.}\, A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M,\ A > 0, \tag{23}$$

where we maximize the worst SQNR lower bound over the angle sector; recall that $\omega_l = \frac{2\pi d}{\lambda}\sin(\theta_l)$, $\omega_u = \frac{2\pi d}{\lambda}\sin(\theta_u)$. We deal with problem (23) by discretization:

$$\max_{\boldsymbol{g}\in\mathbb{C}^L, A\in\mathbb{R}} \min_{i=1,\ldots,I} \widetilde{\mathrm{SQNR}}(\omega_i)$$
$$\text{s.t.}\, A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M,\ A > 0, \tag{24}$$

where, with an abuse of notations, we redefine $\omega_l \le \omega_1 < \omega_2 < \cdots < \omega_I \le \omega_u$ as sample points of $[\omega_l, \omega_u]$ (e.g., by uniform sampling). Problem (24) takes the same form as problem (15), and the same method in Section III-D can be used to solve problem (24). We shall not repeat the details.

We give a numerical illustration by plotting the relative noise shaping responses of the fixed-sector design (24) in Fig. 4. To benchmark, we also plot the relative noise shaping responses of the first-order and second-order ΣΔ modulators in Examples 1 and 2. The settings are $N = 1024$, $L = 16$, $d = \lambda/4$, $\sigma_\eta^2 = 0$, $r_{\min} = r_{\max} = 1$, and $[\theta_l, \theta_u] = [-30^\circ, 30^\circ]$. The first- and second-order ΣΔ modulators have the maximum input amplitude A set to be the largest under the no-overload condition, i.e. $A = M - 1$ and $A = M - 3$, respectively (see Examples 1 and 2). In the plots in Fig. 4, the vertical black lines indicate the angle sector. The magenta double-headed arrows indicate the gap between the worst-case relative noise-shaping response, $\max_{\theta\in[\theta_l,\theta_u]} \mathrm{RNSR}(\theta)$, of the fixed-sector design and the worst-case relative noise-shaping response of the first-order and second-order ΣΔ modulators. We see that the fixed-sector design provides a uniform quantization noise suppression over the angle sector of interest. We also see that, for larger quantization level numbers M's, the fixed-sector design provides considerably improved quantization noise suppression in an angle-sector uniform sense.

F. COMPARISON WITH EXISTING TEMPORAL ΣΔ MODULATOR DESIGNS

It is interesting to compare our optimization-based spatial ΣΔ modulator designs with relevant designs in the temporal ΣΔ literature. To put this into perspective, let us write down the user-targeted and fixed-sector designs, shown in (15) and (23), respectively, as a single formulation:

$$\min_{\boldsymbol{g}\in\mathbb{C}^L, A\in\mathbb{R}} \max_{\omega\in\Omega} \frac{\sqrt{|1+G(\omega)|^2+\gamma_i}}{A}$$
$$\text{s.t.}\, A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M,\ A > 0, \tag{25}$$

where $\Omega = \{\omega_1, \ldots, \omega_K\}$ for the user-targeted case and $\Omega = [\omega_l, \omega_u]$ for the fixed-sector case; note that the constants γ_i's scale with the background noise power σ_η^2. Suppose we prefix the maximum input signal amplitude A and set $\gamma_i = 0$ for all i. The above problem then reduces to

$$\min_{\boldsymbol{g}\in\mathbb{C}^L} \max_{\omega\in\Omega} |1 + G(\omega)|$$
$$\text{s.t. } \|\boldsymbol{g}\|_{\text{IQ}-1} \le M - A \tag{26}$$

which is a multiple notch filter design for the user-targeted case, and a band-stop filter design for the fixed-sector case. In fact, we have seen that in the illustrations in Figs. 3 and 4. In this connection, a formulation similar to (26) was considered by Nagahara and Yamaoto [28] to design temporal $\Sigma\Delta$ modulators for lowpass or bandpass signals. There are subtle differences; e.g., Nagahara and Yamaoto do not use the no-overload constraint in problem (26), and they replace it with a sufficient condition in the form of a linear matrix inequality. The distinctive difference with our designs, apart from being for a different application, is that we also optimize the input amplitude A to maximize the users' SQNRs.

IV. TWO-DIMENSIONAL SPATIAL $\Sigma\Delta$ MODULATION

The spatial $\Sigma\Delta$ modulator designs developed in the preceding sections can be extended to the case of two-dimensional (2D) uniform planar arrays. It should be noted that, to the best of our knowledge, spatial $\Sigma\Delta$ modulation for coarsely quantized MIMO precoding with 2D uniform planar arrays has not been considered before. In the following subsections we will concisely describe how this is done.

A. A 2D $\Sigma\Delta$ MODULATOR

Before we proceed, we should mention that 2D $\Sigma\Delta$ modulation was considered in, and finds important applications to, image half-toning [36]. Here, we first consider the 2D extension of the general $\Sigma\Delta$ modulator in Section III-A. The input-output relation of the 2D modulator is

$$\begin{aligned} x_{n_1,n_2} &= \mathcal{Q}_c(\bar{x}_{n_1,n_2} + (g \circledast q)_{n_1,n_2}) \\ &= \bar{x}_{n_1,n_2} + (g \circledast q)_{n_1,n_2} + q_{n_1,n_2}, \\ (g \circledast q)_{n_1,n_2} &= \sum_{l_1=0}^{L_1}\sum_{l_2=0}^{L_2} g_{l_1,l_2} g_{n_1-l_1,n_2-l_2}, \end{aligned}$$

where $\{\bar{x}_{n_1,n_2}\}_{n_1,n_2\in\mathbb{N}} \subset \mathbb{C}$ is the input; $\{x_{n_1,n_2}\}_{n_1,n_2\in\mathbb{N}} \subset \mathcal{X} + \mathrm{j}\mathcal{X}$ is the output; each $q_{n_1,n_2} \in \mathbb{C}$ is a quantization error and is assumed to be follow the i.i.d. assumption in Assumption 2; the g_{l_1,l_2}'s, $l_1 = 0, \ldots, L_1$, $l_2 = 0, \ldots, L_2$, with $g_{0,0} = 0$, are the filter coefficients. The filter plays the role of shaping the noise magnitude spectrum according to $|1 + G(\omega_1, \omega_2)|^2$, where

$$G(\omega_1, \omega_2) = \sum_{n_1=0}^{L_1}\sum_{n_2=0}^{L_2} g_{n_1,n_2} e^{-\mathrm{j}(n_1\omega_1 + n_2\omega_2)}$$

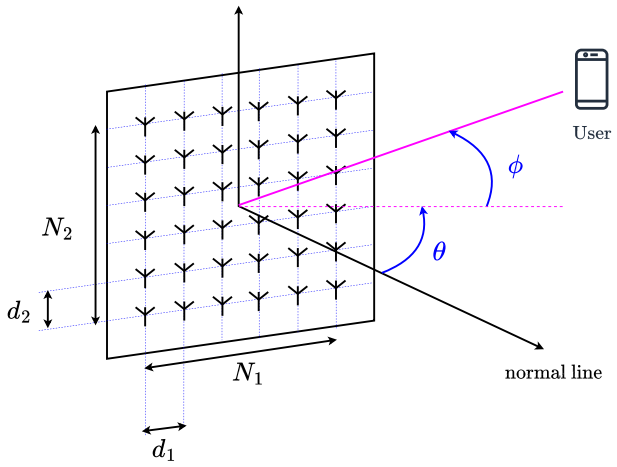


FIGURE 5. Uniform planar array.

is the 2D Fourier transform of $\{g_{l_1,l_2}\}$. Let $\boldsymbol{G} \in \mathbb{C}^{(L_1+1)\times(L_2+1)}$ be a matrix with its (i, j)th element given by $g_{i-1,j-1}$. As the 2D extension of the no-overload condition in Fact 2, the modulator has no overload if

$$A + \|\boldsymbol{G}\|_{\text{IQ}-1} \le M,$$

where $\|\boldsymbol{G}\|_{\text{IQ}-1} = \sum_{l_1=0}^{L_1}\sum_{l_2=0}^{L_2} |\Re(g_{l_1,l_2})| + |\Im(g_{l_1,l_2})|$; $A > 0$ is the maximum input amplitude, i.e., $|x_{n_1,n_2}|_{\text{IQ}-\infty} \le 1$ for all n_1, n_2.

B. UNIFORM PLANAR ARRAY

Second, we review some concepts with the uniform planar array. As illustrated in Fig. 5, a uniform planar array has the antennas arranged in a equi-spaced rectangular fashion [37]. It has N_1 and N_2 antennas in the horizontal and vertical directions, respectively. Under the same set of operating assumptions as uniform linear arrays, the uniform planar array has an array response

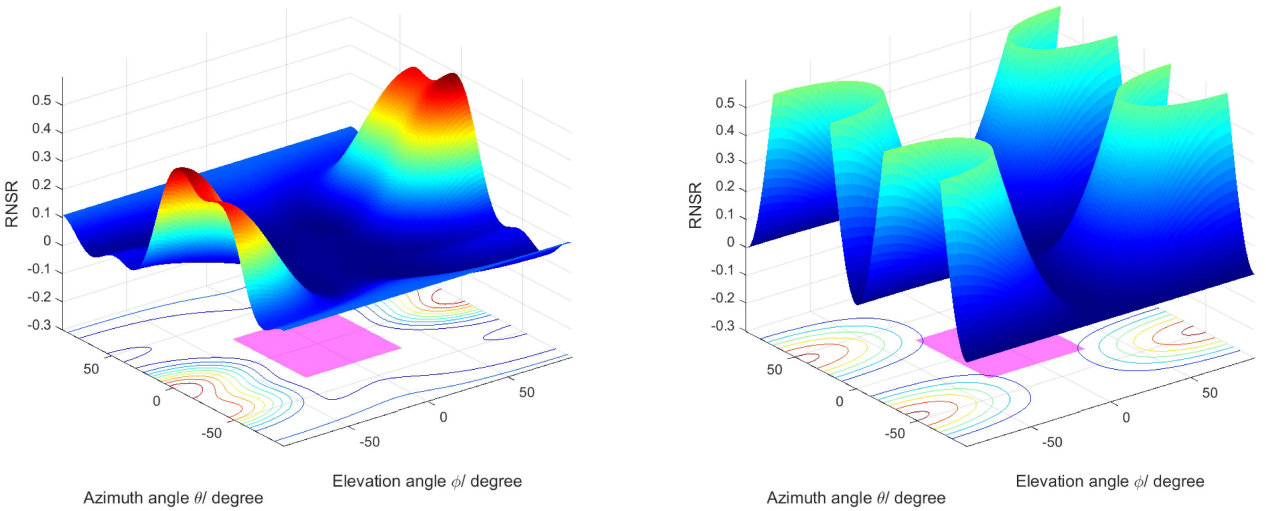
$$\boldsymbol{A}(\theta, \phi) = \boldsymbol{a}_1(\theta, \phi)\boldsymbol{a}_2^T(\phi) \in \mathbb{C}^{N_1\times N_2},$$

where $\theta \in (-\pi/2, \pi/2)$ and $\phi \in (-\pi/2, \pi/2)$ are the azimuth and elevation angles, respectively; we have

$$\begin{aligned} \boldsymbol{a}_1(\theta, \phi) &= (1, e^{-\mathrm{j}\omega_1}, \ldots, e^{-\mathrm{j}(N_1-1)\omega_1}), \\ \boldsymbol{a}_2(\phi) &= (1, e^{-\mathrm{j}\omega_2}, \ldots, e^{-\mathrm{j}(N_2-1)\omega_2}), \\ \omega_1 &= \frac{2\pi d_1}{\lambda}\cos(\phi)\sin(\theta),\ \omega_2 = \frac{2\pi d_2}{\lambda}\sin(\phi); \end{aligned}$$

$d_1 \le \lambda/2$ and $d_2 \le \lambda/2$ are horizontal and vertical inter-antenna spacings, respectively; λ is the carrier wavelength. Let x_{n_1,n_2} be the transmitted signal from the $(n_1 + 1, n_2 + 1)$th antenna of the array, and let $\boldsymbol{X} \in \mathbb{C}^{N_1\times N_2}$ be a matrix with its (i, j)th element given by $x_{i-1,j-1}$. The array exhibits a transmit directional pattern

$$\begin{aligned} \mathrm{tr}(\boldsymbol{A}^T(\theta, \phi)\boldsymbol{X}) &= \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} x_{n_1,n_2} e^{-\mathrm{j}(n_1\omega_1 + n_2\omega_2)} \\ &= X(\omega_1, \omega_2). \end{aligned}$$



(a) The fixed-sector design (27).

(b) The first-order ΣΔ modulator.

FIGURE 6. Illustration of the relative noise-shaping response in the 2D case.

C. ΣΔ MIMO PRECODING FOR UNIFORM PLANAR ARRAYS

Third, we consider spatial ΣΔ modulation for coarsely quantized MIMO precoding in Sections II-C and III when the 1D uniform linear array is replaced by the 2D uniform planar array. Under the 2D uniform planar array setting, the basic signal model (2) is modified as

$$y_{i,t} = \sqrt{\rho}\,\mathrm{tr}(\boldsymbol{H}_i^T \boldsymbol{X}_t) + \eta_{i,t},$$

where $\boldsymbol{X}_t \in \mathbb{C}^{N_1 \times N_2}$ is the transmitted signal; $\boldsymbol{H}_i \in \mathbb{C}^{N_1 \times N_2}$ is the channel of user i and is modeled as

$$\boldsymbol{H}_i = \alpha_i \boldsymbol{A}(\theta_i, \phi_i),$$

in which $\alpha_i, \theta_i, \phi_i$ are the complex channel gain, azimuth angle and elevation angle of user i, respectively. Also, the ΣΔ modulator is replaced by the 2D modulator in Section IV-A. Let $\bar{\boldsymbol{X}}_t \in \mathbb{C}^{N_1 \times N_2}$ be the 2D transmitted signal before ΣΔ modulation. It can be shown that

$$y_{i,t} \simeq \sqrt{\rho}\mathrm{tr}(\boldsymbol{H}_i^T \bar{\boldsymbol{X}}_t) + \sqrt{\rho}\alpha_i(1 + G(\omega_1, \omega_2))Q_t(\omega_1, \omega_2) + \eta_{i,t},$$

which, as before, is the sum of signal, quantization noise, and background noise components; here, $Q_t(\omega_1, \omega_2)$ is the 2D Fourier transform of the quantization error $\{q_{n_1,n_2,t}\}_{n_1,n_2}$. Subsequently, we can further show that the SQNR, under the definition in (14), is

$$\mathrm{SQNR}_i = \frac{\rho|\alpha_i|^2 A^2}{\frac{2N_1N_2\rho|\alpha_i|^2}{3}|1 + G(\omega_{1,i}, \omega_{2,i})|^2 + \sigma_\eta^2}.$$

where $\omega_{1,i} = \frac{2\pi d}{\lambda}\cos(\phi_i)\sin(\theta_i)$, $\omega_{2,i} = \frac{2\pi d}{\lambda}\sin(\phi_i)$.

Let us describe the modulator designs. We can follow the user-targeted ΣΔ modulator design in problem (15) in Section III-D, which maximizes the users' SQNRs in the max-min-fair fashion and in a user targeted fashion. The 2D extension of the design is

$$\max_{\boldsymbol{G}, A \in \mathbb{R}} \min_{i=1,\ldots,K} \mathrm{SQNR}_i$$
$$\text{s.t.}\, A + \|\boldsymbol{G}\|_{\mathrm{IQ}-1} \le M,\ A > 0,$$

where the domain of $\boldsymbol{G}$ is $\mathbb{C}^{(L_1+1)\times(L_2+1)}$, $g_{0,0} = 0$. The above problem takes the same form as its predecessor, problem (15), and it can be solved by the exactly same way as in Section II-I-D. We can also adopt the fixed-sector design in problem (23) in Section III-E, which designs a fixed modulator for an angle sector by maximizing the worst SQNR lower bound over that sector. Let $[\theta_l, \theta_u] \times [\phi_l, \phi_u]$ be the angle sector of interest. The 2D extension of the design is

$$\max_{\boldsymbol{G}, A \in \mathbb{R}} \min_{(\omega_1,\omega_2)\in\Omega} \widetilde{\mathrm{SQNR}}(\omega_1, \omega_2)$$
$$\text{s.t.}\, A + \|\boldsymbol{G}\|_{\mathrm{IQ}-1} \le M,\ A > 0, \tag{27}$$

where

$$\Omega = \left\{ \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} = \begin{bmatrix} \frac{2\pi d}{\lambda}\cos(\phi)\sin(\theta) \\ \frac{2\pi d}{\lambda}\sin(\phi) \end{bmatrix} \middle| \begin{matrix} \theta \in [\theta_l, \theta_u] \\ \phi \in [\phi_l, \phi_u]' \end{matrix} \right\};$$

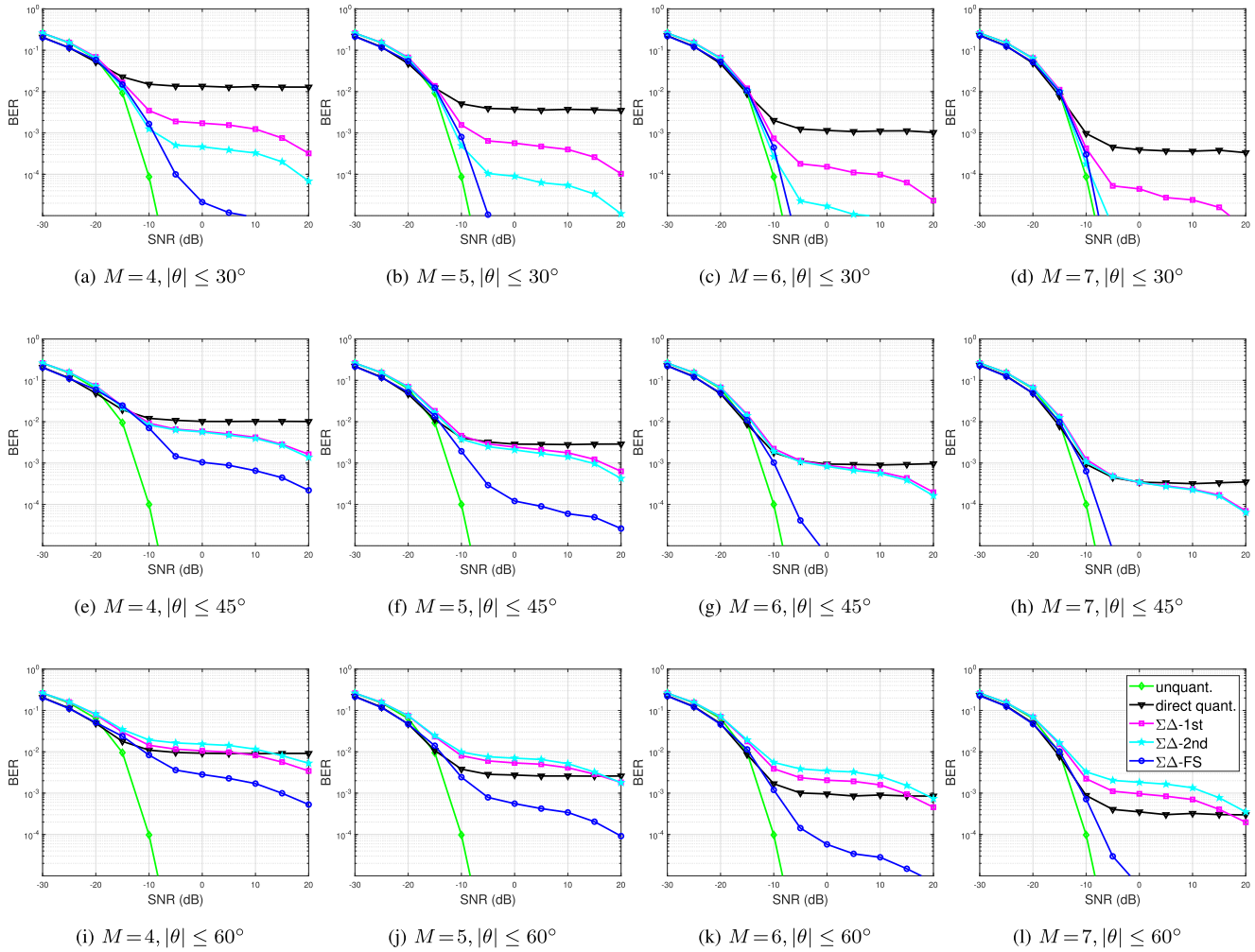
$$\widetilde{\mathrm{SQNR}}(\omega_1, \omega_2) = \frac{\rho r_{\min}^2 A^2}{\frac{2N_1N_2\rho r_{\max}^2}{3}|1 + G(\omega_1, \omega_2)|^2 + \sigma_\eta^2}.$$

The above problem can be handled by the same way as in Section III-E.

As an illustration, Fig. 6(a) plots the relative noise shaping response of the fixed-sector design (27). The settings are $(N_1, N_2) = (60, 60)$, $d_1 = d_2 = \lambda/4$, $(L_1, L_2) = (4, 4)$, $r_{\max} = r_{\min} = 1$, $\sigma_\eta^2 = 0$, $[\theta_l, \theta_u] \times [\phi_l, \phi_u] = [-30^\circ, 30^\circ] \times [-30^\circ, 30^\circ]$, $M = 4$. To benchmark, we also consider a 2D first-order ΣΔ modulator whose shaping response is

$$1 + G(\omega_1, \omega_2) = (1 - e^{-\mathrm{j}\omega_1})(1 - e^{-\mathrm{j}\omega_2}), \tag{28}$$

and whose coefficients are $(g_{01}, g_{10}, g_{11}) = (-1, -1, 1)$; we set $A = M - 3$, the maximum under the no-overload condition. The relative noise shaping response of this first-order modulator is plotted in Fig. 6(b). Comparing Fig. 6(a) and (b),



(a) $M=4, |\theta| \le 30^\circ$ (b) $M=5, |\theta| \le 30^\circ$ (c) $M=6, |\theta| \le 30^\circ$ (d) $M=7, |\theta| \le 30^\circ$

(e) $M=4, |\theta| \le 45^\circ$ (f) $M=5, |\theta| \le 45^\circ$ (g) $M=6, |\theta| \le 45^\circ$ (h) $M=7, |\theta| \le 45^\circ$

(i) $M=4, |\theta| \le 60^\circ$ (j) $M=5, |\theta| \le 60^\circ$ (k) $M=6, |\theta| \le 60^\circ$ (l) $M=7, |\theta| \le 60^\circ$

FIGURE 7. BER performance of the fixed-sector optimized ΣΔ modulation scheme for various settings of the angle sector $[\theta_l, \theta_u]$. $N = 1024$, $d = \lambda/4$, $K = 8$, $L = 16$, $|\theta| \le \vartheta$ means that the angle sector is $[\theta_l, \theta_u] = [-\vartheta, \vartheta]$. ΣΔ-FS: the fixed-sector ΣΔ modulation scheme, ΣΔ-1st: the first-order ΣΔ modulation scheme, ΣΔ-2nd: the second-order ΣΔ modulation scheme, direct quant.: the direct quantization scheme, unquant.: the unquantized performance baseline.

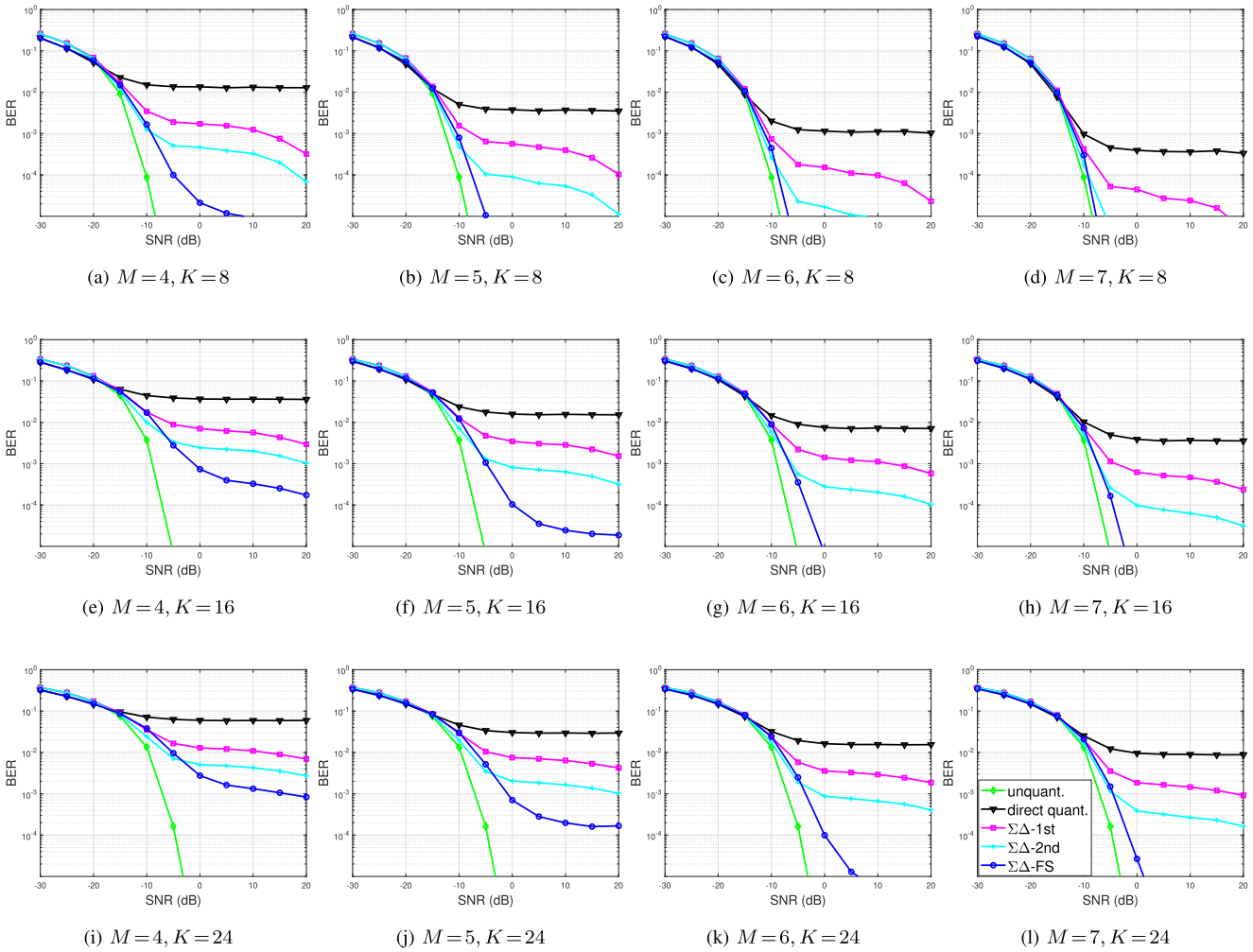
the fixed-sector design (27) appears to provide better quantization noise suppression over the given angle sector than the first-order modulator.

V. NUMERICAL RESULTS

In this section we provide numerical results. We simulate both spatial ΣΔ modulation and precoding at the signal level, and we evaluate users' bit error rates (BERs) as our way to assess the performance of our method. The symbol stream $\{s_{i,t}\}_{t=1}^{T}$ of each user is drawn from the 64-QAM constellation, with symbol stream length $T = 500$. The precoding scheme is the ZF scheme. To be specific, for a given spatial ΣΔ modulator, the ZF precoded signals are given by

$$\bar{\boldsymbol{x}}_t = \frac{A}{C}\boldsymbol{H}^{\dagger}\boldsymbol{D}\boldsymbol{s}_t,\ t = 1, \ldots, T, \tag{29}$$

where $C = \max_{t=1,\ldots,T} \|\boldsymbol{H}^{\dagger}\boldsymbol{D}\boldsymbol{s}_t\|_{\mathbb{Q}-\infty}$, $\boldsymbol{D} = \mathrm{Diag}(\sigma_{w,1}, \ldots, \sigma_{w,K})$, $\sigma_{w,i}^2 = \rho|\alpha_i|^2\mathbb{E}[|v_{i,t}|^2] + \sigma_\eta^2$, and $\mathbb{E}[|v_{i,t}|^2] = 2N|1 + G(\omega_i)|^2/3$; see [20]. Note that we scale the symbol streams such that the post-precoding SNRs of all the users are equal, and that the normalization with C is to enforce the peak signal amplitude constraint $\|\bar{\boldsymbol{x}}_t\|_{\mathbb{Q}-\infty} \le A$ for all t. For each symbol time t, the precoded signal $\bar{\boldsymbol{x}}_t$ is fed to the ΣΔ modulator to generate the transmitted signal $\boldsymbol{x}_t$. Our ΣΔ modulation scheme is designed either by the fixed-sector design in Section III-E or by the user-targeted design in Section III-D. As benchmarks, we also consider the first- and second-order ΣΔ modulation schemes in Examples 1 and 2, respectively, which are standard modulators in the ΣΔ literature. The maximum input signal amplitude A of the first- and second-order modulators is set to be the largest under the no-overload condition, which are $A = M - 1$ and $A = M - 3$ for the first- and second-order modulators, respectively. In addition we benchmark the direct quantization method. We employ the ZF precoding scheme, to be consistent with our benchmarking, and the transmitted signals for the directly quantized ZF



(a) $M=4, K=8$ (b) $M=5, K=8$ (c) $M=6, K=8$ (d) $M=7, K=8$

(e) $M=4, K=16$ (f) $M=5, K=16$ (g) $M=6, K=16$ (h) $M=7, K=16$

(i) $M=4, K=24$ (j) $M=5, K=24$ (k) $M=6, K=24$ (l) $M=7, K=24$

FIGURE 8. BER performance of the fixed-sector optimized ΣΔ modulation scheme for various values of the user number K. $N = 1024$, $d = \lambda/4$, $L = 16$, $[\theta_l, \theta_u] = [-30°, 30°]$. See the caption of Fig. 7 for a description of the legend labels.

scheme are given by

$$\boldsymbol{x}_t = \mathcal{Q}_c\left(M\frac{1}{C}\boldsymbol{H}^{\dagger}\boldsymbol{s}_t\right),\ t=1,\ldots,T,$$

where $C = \max_{t=1,\ldots,T} \|\boldsymbol{H}^{\dagger}\boldsymbol{D}\boldsymbol{s}_t\|_{\mathrm{IQ}-\infty}$. Furthermore we provide a performance baseline by evaluating the BER performance of the following unquantized ZF scheme:

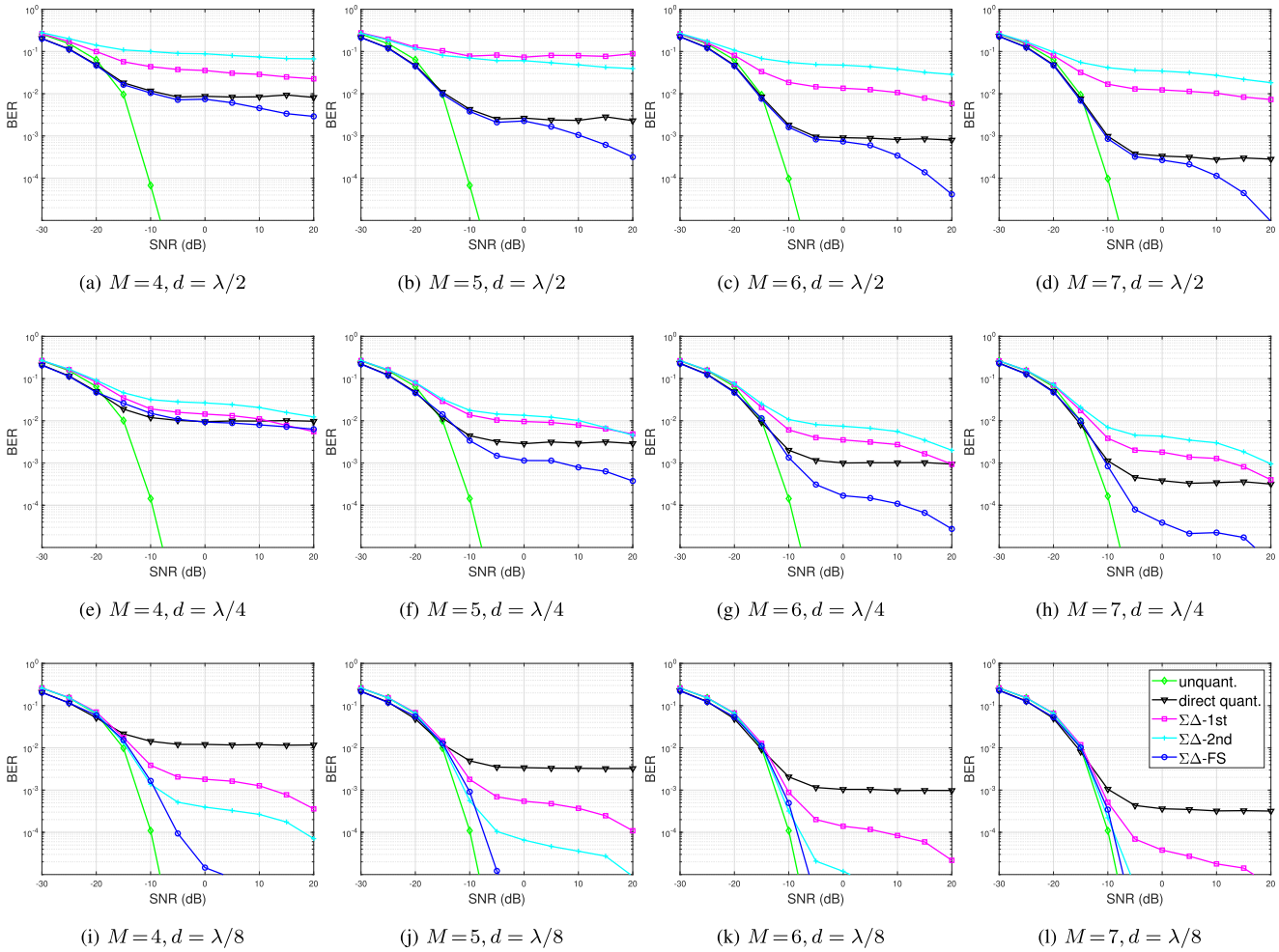
$$\boldsymbol{x}_t = \frac{M-1}{C}\boldsymbol{H}^{\dagger}\boldsymbol{s}_t,\ t=1,\ldots,T, \tag{30}$$

where $C = \max_{t=1,\ldots,T} \|\boldsymbol{H}^{\dagger}\boldsymbol{s}_t\|_{\mathrm{IQ}-\infty}$. Note that this unquantized scheme satisfies $\|\boldsymbol{x}_t\|_{\mathrm{IQ}-\infty} \leq M-1$ for all t, which complies with the peak signal amplitude constraint for the coarsely quantized case. We define the SNR as $\mathrm{SNR} = (M-1)^2\rho/\sigma_\eta^2$, which is the ratio of the per-antenna peak power to the background noise power.

The BER performance to be reported was obtained by Monte-Carlo simulations with 1,000 trials. At each trial, the user angles θ_i's and the complex channnel gains α_i's are generated by the following way. The user angles θ_i's are randomly drawn from a prespecified angle sector $[\theta_l, \theta_u]$, and they are separated by no less than $1°$. The phases of α_i's are uniformly drawn from $[-\pi, \pi]$. The amplitudes of α_i's are generated by $|\alpha_i| = r_0/r_1$, where $r_0 = 30$ and r_1 is randomly drawn from $[20, 100]$.

A. FIXED-SECTOR DESIGN

We consider the fixed-sector design in Section III-E. Here are the settings: The number of transmit antennas is $N = 1024$; the inter-antenna spacing is $d = \lambda/4$ (we remind the reader that a small d leads to a small quantization noise power, as described in Section II-C); the number of users is $K = 8$; the filter order of our fixed-sector optimized ΣΔ modulator is $L = 16$. Fig. 7 shows the BER-versus-SNR plots of our scheme and the benchmarked schemes for various settings of the quantization level number M and the angle sector $[\theta_l, \theta_u]$. In particular, Fig. 7(a)–(d) consider $[\theta_l, \theta_u] = [-30°, 30°]$, Fig. 7(e)–(h) consider $[\theta_l, \theta_u] = [-45°, 45°]$, and Fig. 7(i)–(l) consider $[\theta_l, \theta_u] = [-60°, 60°]$. We see that our fixed-sector optimized ΣΔ modulation scheme generally leads to better



(a) $M=4, d=\lambda/2$ (b) $M=5, d=\lambda/2$ (c) $M=6, d=\lambda/2$ (d) $M=7, d=\lambda/2$

(e) $M=4, d=\lambda/4$ (f) $M=5, d=\lambda/4$ (g) $M=6, d=\lambda/4$ (h) $M=7, d=\lambda/4$

(i) $M=4, d=\lambda/8$ (j) $M=5, d=\lambda/8$ (k) $M=6, d=\lambda/8$ (l) $M=7, d=\lambda/8$

FIGURE 9. BER performance of the fixed-sector optimized ΣΔ modulation scheme for various values of the inter-antenna spacing d. $N = 1024$, $K = 8$, $L = 16$, $[\theta_l, \theta_u] = [-75°, 75°]$. See the caption of Fig. 7 for a description of the legend labels.

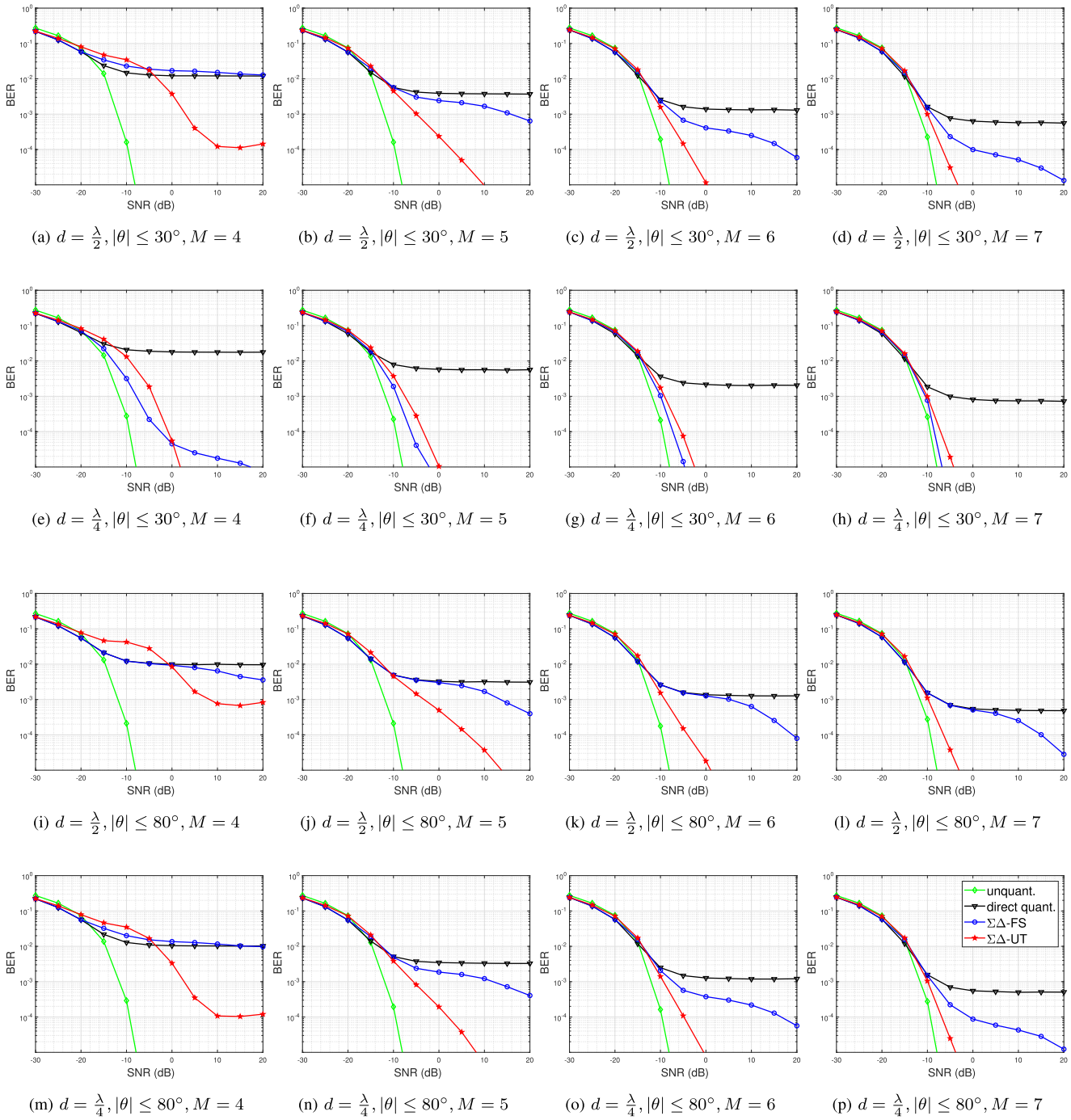
BER performance than the benchmarked schemes. We also see that, as the width of the angle sector increases, we need a larger quantization level number M to provide the same or similar BER performance level.

The next simulation result has the challenge raised by increasing the user number K. Fig. 8 displays a set of BER-versus-SNR plots for various values of K, wherein we fix the angle sector as $[\theta_l, \theta_u] = [-30°, 30°]$. Once again, the fixed-sector optimized ΣΔ modulation scheme is seen to lead to better BER performance than the benchmarked schemes. It is also noticed that, as the user number K increases, we need a larger quantization level number M to get close to the unquantized performance baseline.

We are also interested in how the performance changes with the inter-antenna spacing d. As discussed, the ΣΔ notion suggests that we want d to be as small as possible, but physical limitations disallow us from making d too small. Fig. 9 shows the results for various values of d, wherein we set the angle sector as $[\theta_l, \theta_u] = [-75°, 75°]$ which is relatively wide. We see that a smaller d leads to better performance for all the ΣΔ schemes, while a larger d requires us to use a larger quantization level number M to get reasonable performance. This simulation result, together with the previous results, indicate a tradeoff—if we want to have a wider angle sector and/or a larger inter-antenna spacing, the ΣΔ noise shaping problem becomes harder and we need a greater number of quantization levels to meet the challenge.

B. USER-TARGETED DESIGN

We turn our interest to the user-targeted design in Section II-I-D. The settings are identical to those in the last subsection, except that the filter order of our optimization-based ΣΔ modulator is $L = 24$. Fig. 10 displays a set of BER-versus-SNR plots when the user number is $K = 9$. We see that the user-targeted ΣΔ modulation scheme can improve upon the fixed-sector optimized ΣΔ modulation scheme, and the improvement is significant for larger values of the quantization level number M. For instance, for $d = \lambda/2$ and $[\theta_l, \theta_u] = [-80°, 80°]$, which is a challenging setting, Fig. 10(l) shows that the user-targeted ΣΔ modulation scheme can lead to BER



(a) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 4$ (b) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 5$ (c) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 6$ (d) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 7$

(e) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 4$ (f) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 5$ (g) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 6$ (h) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 7$

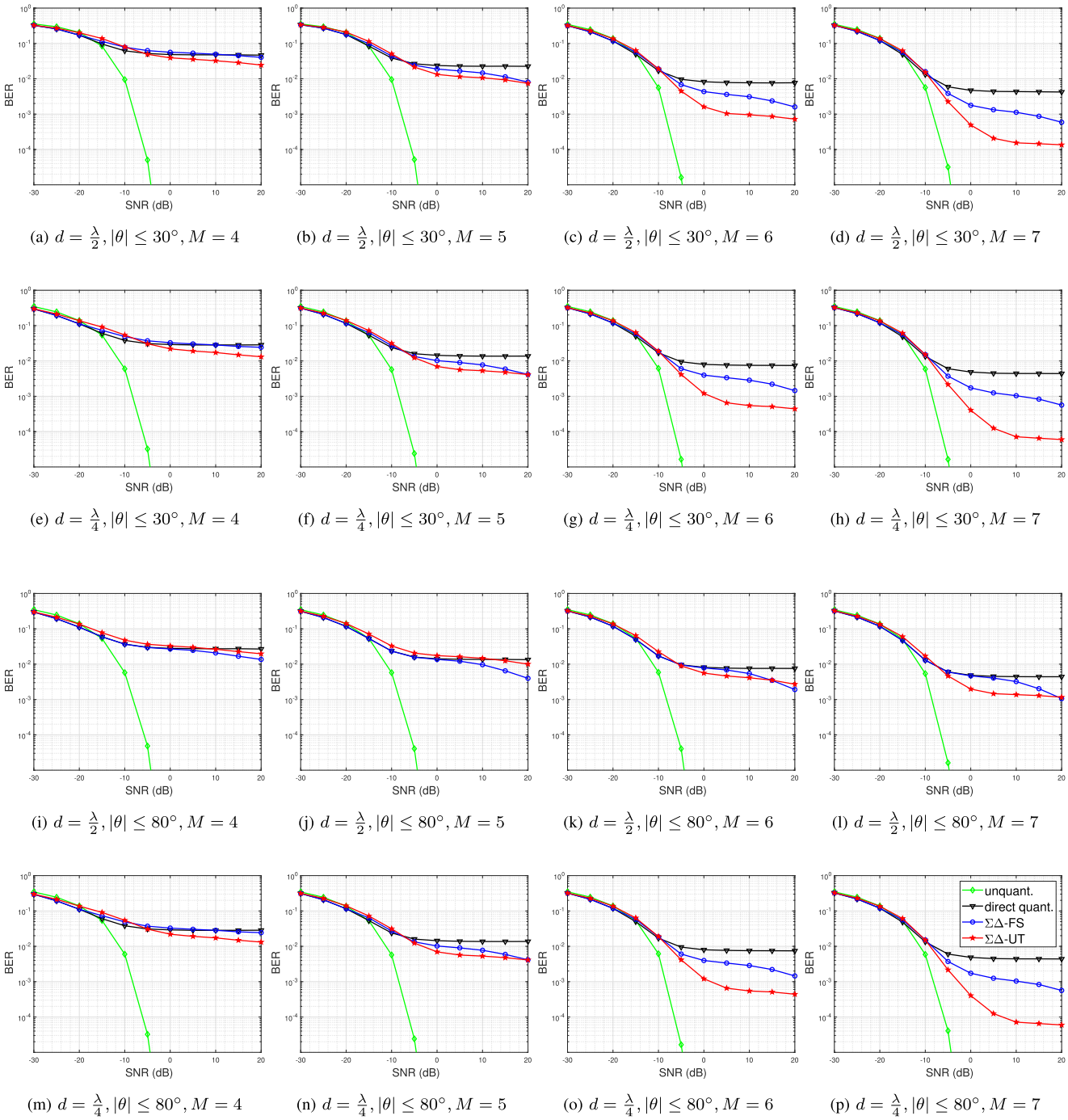
(i) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 4$ (j) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 5$ (k) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 6$ (l) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 7$

(m) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 4$ (n) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 5$ (o) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 6$ (p) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 7$

FIGURE 10. BER performance of the user-targeted ΣΔ modulation scheme for $K = 9$. $N = 1024$, $L = 24$, $|\theta| \leq \vartheta$ means that the angle sector is $[\theta_l, \theta_u] = [-\vartheta, \vartheta]$. ΣΔ-UT: the user-targeted ΣΔ modulation scheme. See the caption of Fig. 7 for a description of the other legend labels.

performance close to the unquantized performance baseline. Also we see that if d is larger and/or the angle sector width is larger, the user-targeted ΣΔ modulation scheme requires a larger M to provide good performance. This is in agreement with our observation with the fixed-sector ΣΔ modulation scheme in the last subsection.

Fig. 11 shows another set of plots wherein we increase the user number to $K = 18$. Comparing this result with the previous result ($K = 9$), we observe that (i) the performance behaviors of the current result appears to be consistent with those of the previous; (ii) the performance sees degradation as the user number increases. We argue that the second observation is an inevitable limitation, as alluded to by Proposition 1 which suggests that achieving zero quantization noise would require the quantization level number M to increase exponentially with the user number K.



(a) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 4$ (b) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 5$ (c) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 6$ (d) $d = \frac{\lambda}{2}, |\theta| \leq 30°, M = 7$

(e) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 4$ (f) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 5$ (g) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 6$ (h) $d = \frac{\lambda}{4}, |\theta| \leq 30°, M = 7$

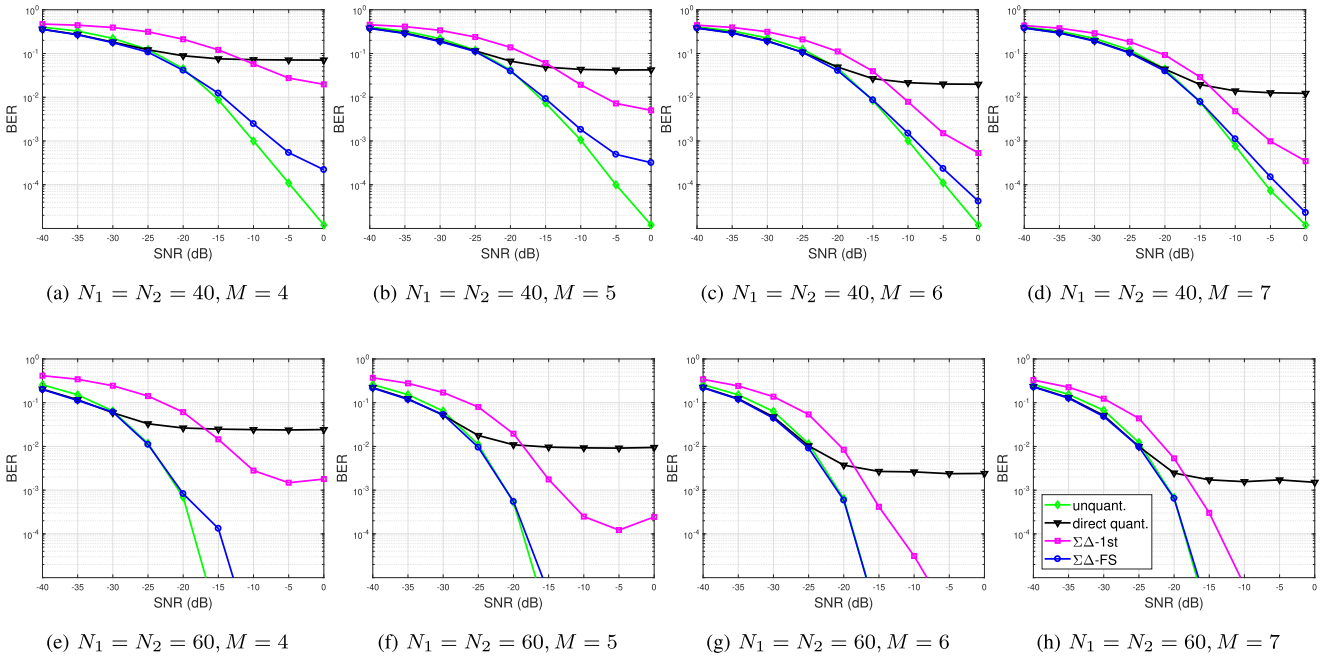
(i) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 4$ (j) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 5$ (k) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 6$ (l) $d = \frac{\lambda}{2}, |\theta| \leq 80°, M = 7$

(m) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 4$ (n) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 5$ (o) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 6$ (p) $d = \frac{\lambda}{4}, |\theta| \leq 80°, M = 7$

FIGURE 11. BER performance of the user-targeted ΣΔ modulation scheme for $K = 18$. The settings are identical to those in Fig. 10.

C. 2D SPATIAL ΣΔ MODULATION FOR UNIFORM PLANAR ARRAYS

We consider the 2D spatial ΣΔ modulation schemes for uniform planar arrays, described in Section IV. The simulation workflow is identical to the above. The simulation settings are as follows: the user number is $K = 8$; the inter-antenna spacings are $d_1 = d_2 = \lambda/4$; the angle sector is $[\theta_l, \theta_u] \times [\phi_l, \phi_u] = [-30°, 30°] \times [0°, 20°]$; the filter order of our optimized ΣΔ modulator is $(L_1, L_2) = (5, 5)$. We consider the fixed-sector design, and we use the 2D first-order ΣΔ modulator (cf. (28)) as our main benchmark. Fig. 12 shows the results for two different settings of the transmit antenna size (N_1, N_2). The results demonstrate that the 2D spatial ΣΔ modulation schemes are viable. We should remark that, to the best of our knowledge, 2D spatial ΣΔ modulation for coarsely quantized MIMO precoding with uniform planar arrays was not attempted before; even the 2D first-order ΣΔ modulation scheme is a new attempt.



(a) $N_1 = N_2 = 40, M = 4$ (b) $N_1 = N_2 = 40, M = 5$ (c) $N_1 = N_2 = 40, M = 6$ (d) $N_1 = N_2 = 40, M = 7$

(e) $N_1 = N_2 = 60, M = 4$ (f) $N_1 = N_2 = 60, M = 5$ (g) $N_1 = N_2 = 60, M = 6$ (h) $N_1 = N_2 = 60, M = 7$

FIGURE 12. BER performance of the 2D fixed-sector optimized ΣΔ modulation scheme. $d_1 = d_2 = \lambda/4$, $K = 8$, $L_1 = L_2 = 5$, $[\theta_l, \theta_u] \times [\phi_l, \phi_u] = [-30°, 30°] \times [0, 20°]$. See the caption of Fig. 7 for a description of the legend labels.

VI. CONCLUSION

To summarize, we developed a spatial ΣΔ modulator design framework for coarsely quantized massive MIMO downlink precoding. Our framework is flexible. It can handle any ΣΔ filter order and any number of quantization levels. It can deal with various SQNR requirements, such as max-min-fair SQNR enhancement over a prescribed angle sector, or SQNR enhancement in accordance with the user angles in an instantaneous fashion. It can also be extended to 2D uniform planar arrays. Our design framework is based on convex optimization. Numerical results showed that ΣΔ modulators designed under our framework outperform the existing ΣΔ modulators, and may lead to near-ideal (unquantized) performance under certain operating conditions.

APPENDIX A

A. PROOF OF FACT 2

Suppose $|q_{n-l}|_{\mathrm{IQ}-\infty} \le 1$ for all $l \ge 1$. For convenience, let $b_n = \bar{x}_n + (g \circledast q)_n$. We have

$$\begin{aligned}|\Re(b_n)| &\le |\Re(\bar{x}_n)| + \left|\Re\left(\sum_{l=1}^{L} g_l q_{n-l}\right)\right| \\ &\le |\Re(\bar{x}_n)| + \sum_{l=1}^{L}(|\Re(g)_l||\Re(q_{n-l})| + |\Im(g)_l||\Im(q_{n-l})|) \\ &\le A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1}.\end{aligned}$$

Recall that $\mathcal{Q}$ is the quantizer associated with $\mathcal{X}$ in (8), and note $\Re(q_n) = \mathcal{Q}(\Re(b_n)) - \Re(b_n)$. It can be verified that $|\mathcal{Q}(y) - y| \le 1$ if $|y| \le M$. Hence, if $A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M$ holds, we have $|\Re(q_n)| \le 1$. Similarly, one can show that if $A + \|\boldsymbol{g}\|_{\mathrm{IQ}-1} \le M$, then $|\Im(q_n)| \le 1$. The proof is done.

B. PROOF OF PROPOSITIONS 1 AND 2

First we show Proposition 1.(a) and Proposition 2. For convenience, rewrite the coefficients g_k's in (13) as

$$g_k = \sum_{1 \le i_1 < \cdots < i_k \le K} \beta_{i_1} \cdots \beta_{i_k},$$

where $\beta_i = -e^{\mathrm{j}\omega_i}$. Note that the coefficients g_k's in (22) is a special case of the above where $\beta_1 = \cdots = \beta_K = e^{\mathrm{j}\omega_c}$, $K = L$. It can be shown that, for $x \in \mathbb{C}$,

$$|x|_{\mathrm{IQ}-1} := |\Re(x)| + |\Im(x)| \ge |x|, \tag{31}$$

$$|x|_{\mathrm{IQ}-1} \le \sqrt{2}|x|, \tag{32}$$

where equality in (32) is attained if x takes the form $x = |x|e^{j\omega}$, $\omega \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$. This leads to

$$\begin{aligned}|g_k|_{\mathrm{IQ}-1} &\le \sqrt{2}\left|\sum_{1 \le i_1 < \cdots < i_k \le K} \beta_{i_1} \cdots \beta_{i_k}\right| \\ &\le \sqrt{2} \sum_{1 \le i_1 < \cdots < i_k \le K} |\beta_{i_1} \cdots \beta_{i_k}| \\ &= \sqrt{2}\binom{K}{k},\end{aligned}$$

where equality above is attained if $\beta_1 = \cdots = \beta_K = e^{\mathrm{j}\omega}$, $\omega \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$. Hence we have

$$\|\boldsymbol{g}\|_{\mathrm{IQ}-1} \leq \sqrt{2}\sum_{k=1}^{K}\binom{K}{k} = \sqrt{2}(2^K - 1),$$

which is the inequality in Proposition 1.(a) and the upper bound inequality in Proposition 2. Furthermore, for the case of $\beta_1 = \cdots = \beta_K := \beta$, we use (31) to obtain

$$|g_k|_{\mathrm{IQ}-1} \geq \left|\sum_{1\leq i_1<\cdots<i_k\leq K}\beta^k\right| = \binom{K}{k}.$$

Consequently we have $\|\boldsymbol{g}\|_{\mathrm{IQ}-1} \geq 2^K - 1$, the lower bound inequality in Proposition 2.

Second we show Proposition 1.(b). If $\omega_1, \ldots, \omega_K$ are i.i.d. and $(-\pi, \pi)$-uniform distributed, one can show the following: given $1 \leq i_1 < \cdots < i_k \leq K$, $1 \leq j_1 < \cdots < j_k \leq K$,

$$\mathbb{E}[\beta_{i_1}\cdots\beta_{i_k}\beta_{j_1}^*\cdots\beta_{j_k}^*] = \begin{cases}1, & i_l = j_l \text{ for all } l\\ 0, & \text{otherwise}\end{cases}$$

Subsequently we have

$$\begin{aligned}&\mathbb{E}[|g_k|^2] =\\ &\textstyle\sum_{1\leq i_1<\cdots<i_k\leq K}\sum_{1\leq j_1<\cdots<j_k\leq K}\mathbb{E}[\beta_{i_1}\cdots\beta_{i_k}\beta_{j_1}^*\cdots\beta_{j_k}^*]\\ &= \textstyle\sum_{1\leq i_1<\cdots<i_k\leq K} 1\\ &= \tbinom{K}{k}.\end{aligned}$$

Also, it can be shown that, for $\boldsymbol{x} \in \mathbb{C}^K$, $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_{\mathrm{IQ}-1} \leq \sqrt{2K}\|\boldsymbol{x}\|_2$. This leads to

$$2^K - 1 \leq \mathbb{E}[\|\boldsymbol{g}\|_{\mathrm{IQ}-1}^2] \leq 2K(2^K - 1). \tag{33}$$

Our final step is to polish the above bounds to a simpler form. Consider the inequalities below:

$$\begin{aligned}2K = 2^{\log(K)+\log(2)} &\leq 2^{K-1+\log(2)} \leq 2^K\\ 2^K - 1 &\leq 2^K\\ 2^K - 1 &\geq 2^K - 2^{K-1} = 2^{K-1}\end{aligned}$$

where, in the first equation, we have used $\log(x) \leq x - 1$ for $x > 0$. Applying the above inequalities to (33) gives the result in Proposition 1.(b).

REFERENCES

[1] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[2] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," 2014, *arXiv:1404.7736*.

[3] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005–2018, May 2016.

[4] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Jan. 2017.

[5] 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.

[6] 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.

[7] A. Swindlehurst, A. 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.*, 2017, pp. 6483–6487.

[8] 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.

[9] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[10] M. Shao, W.-K. Ma, J. Liu, and Z. Huang, "Accelerated and deep expectation maximization for one-bit MIMO-OFDM detection," *IEEE Trans. Signal Process.*, vol. 72, pp. 1094–1113, 2024.

[11] Y. Li, C. Tao, A. Lee Swindlehurst, A. Mezghani, and L. Liu, "Downlink achievable rate analysis in massive MIMO systems with one-bit DACs," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1669–1672, Jul. 2017.

[12] 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.*, 2009, pp. 683–686.

[13] F. Sohrabi, Y.-F. Liu, and W. Yu, "One-bit precoding and constellation range design for massive MIMO with QAM signaling," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 557–570, Jun. 2018.

[14] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive MISO channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, Oct. 2019.

[15] 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.

[16] A. Li, C. Masouros, F. Liu, and A. L. Swindlehurst, "Massive MIMO 1-bit DAC transmission: A low-complexity symbol scaling approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7559–7575, Nov. 2018.

[17] H. Jedda, A. Mezghani, A. L. Swindlehurst, and J. A. Nossek, "Quantized constant envelope precoding with PSK and QAM signaling," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8022–8034, Dec. 2018.

[18] M. Kazemi, H. Aghaeinia, and T. M. Duman, "Discrete-phase constant envelope precoding for massive MIMO systems," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2011–2021, May 2017.

[19] Z. Wu, Y.-F. Liu, B. Jiang, and Y.-H. Dai, "Efficient quantized constant envelope precoding for multiuser downlink massive MIMO systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[20] M. Shao, W.-K. Ma, Q. Li, and A. L. Swindlehurst, "One-bit sigma-delta MIMO precoding," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1046–1061, Sep. 2019.

[21] M. Shao, W.-K. Ma, and L. Swindlehurst, "Multiuser massive MIMO downlink precoding using second-order spatial sigma-delta modulation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 8966–8970.

[22] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Piscataway, NJ, USA: Wiley, 2005, vol. 74.

[23] D. Barac and E. Lindqvist, "Spatial sigma-delta modulation in a massive MIMO cellular system," Master's Thesis, Dept. Comput. Sci. Eng., Chalmers Univ. Technol., Gothenburg, Sweden, 2016.

[24] R. M. Corey and A. C. Singer, "Spatial sigma-delta signal acquisition for wideband beamforming arrays," in *Proc. Int. ITG Workshop Smart Antennas*, 2016, pp. 1–7.

[25] H. Pirzadeh, G. Seco-Granados, S. Rao, and A. L. Swindlehurst, "Spectral efficiency of one-bit sigma-delta massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2215–2226, Sep. 2020.

[26] S. Rao, G. Seco-Granados, H. Pirzadeh, J. A. Nossek, and A. L. Swindlehurst, "Massive MIMO channel estimation with low-resolution spatial sigma-delta ADCs," *IEEE Access*, vol. 9, pp. 109320–109334, 2021.

[27] R. P. Sankar and S. P. Chepuri, "Channel estimation in MIMO systems with one-bit spatial sigma-delta ADCs," *IEEE Trans. Signal Process.*, vol. 70, pp. 4681–4696, 2022.

[28] M. Nagahara and Y. Yamamoto, "Frequency domain min-max optimization of noise-shaping delta-sigma modulators," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2828–2839, Jun. 2012.
[29] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1220–1244, Nov. 1990.
[30] I. Daubechies and R. DeVore, "Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Ann. Math.*, vol. 158, no. 2, pp. 679–710, 2003.
[31] C. S. Güntürk, "One-bit sigma-delta quantization with exponential accuracy," *Commun. Pure Appl. Math.*, vol. 56, no. 11, pp. 1608–1630, 2003.
[32] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[33] R. Schreier and M. Snelgrove, "Stability in a general sigma delta modulator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1991, pp. 1769–1772.
[34] A. Charnes and W. W. Cooper, "Programming with linear fractional functionals," *Nav. Res. Logistics Quart.*, vol. 9, no. 3/4, pp. 181–186, 1962.
[35] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014, [Online]. Available: http://cvxr.com/cvx
[36] T. D. Kite, B. L. Evans, A. C. Bovik, and T. L. Sculley, "Digital halftoning as 2-D delta-sigma modulation," in *Proc. IEEE Int. Conf. Image Process.*, 1997, vol. 1, pp. 799–802.
[37] C. A. Balanis, *Antenna Theory: Analysis and Design*. Hoboken, NJ, USA: Wiley, 2015.

WAI-YIU KEUNG received the B. Eng. (Hons.) and the M. Phil. degrees in 2018 and 2020, respectively, from the Chinese University of Hong Kong (CUHK), Hong Kong, where he is currently working toward the Ph.D. degree with the Department of Electronic Engineering under the supervision of Professor Wing-Kin Ma. He is currently an Assistant Lecturer with the Department of Computer Science and Engineering, CUHK. His research interests include signal processing and optimization, with a focus on applications in hardware-constrained physical-layer transceiver designs for massive MIMO.

WING-KIN MA (Fellow, IEEE) is currently a Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. His research interests include signal processing, optimization and communications, with recent focus on 1) optimization and statistical aspects with structured matrix factorization, with application to remote sensing and data science; and 2) coarsely quantized MIMO transceiver designs. Dr. Ma has rich experience in editorial service, such as an Associate Editor, and then later, a Senior Area Editor, and then from 2021 to 2023, an Editor-in-Chief of IEEE TRANSACTIONS ON SIGNAL PROCESSING, and many others. He was a Tutorial Speaker in EUSIPCO 2011 and ICASSP 2014, and was an IEEE Signal Processing Society (SPS) Distinguished Lecturer during 2018–2019. He was the recipient of the Research Excellence Award 2013–2014 by CUHK, 2015 IEEE Signal Processing Magazine Best Paper Award, 2016 IEEE Signal Processing Letters Best Paper Award, and 2018 IEEE SPS Best Paper Award. He was a Member of the Signal Processing for Communications and Networking Technical Committee (SPCOM-TC) during 2015–2020, Member of Signal Processing Theory and Methods Technical Committee (SPTM-TC) during 2012–2017, SPS Regional Director-at-Large for Region 10 during 2020–2021, and Technical Program Co-Chair of ICASSP 2023. He co-founded and co-organized One World Signal Processing in 2020, a virtual seminar series for signal processing.