This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

One-Bit Sigma-Delta MIMO Precoding

Mingjie Shao, Wing-Kin Ma, Qiang Li, and Lee Swindlehurst

***Abstract*— Coarsely quantized MIMO signalling methods have gained popularity in the recent developments of massive MIMO as they open up opportunities for massive MIMO implementation using cheap and power-efficient radio-frequency front-ends. This paper presents a new one-bit MIMO precoding approach using spatial Sigma-Delta ($\Sigma\Delta$) modulation. In previous one-bit MIMO precoding research, one mainly focuses on using optimization to tackle the difficult binary signal optimization problem that arises from the precoding design. Our approach attempts a different route. Assuming angular MIMO channels, we apply $\Sigma\Delta$ modulation—a classical concept in analog-to-digital conversion of temporal signals—in space. The resulting $\Sigma\Delta$ precoding approach has two main advantages: First, we no longer need to deal with binary optimization in $\Sigma\Delta$ precoding design. Particularly, the binary signal restriction is replaced by peak signal amplitude constraints. Second, the impact of the quantization error can be well controlled via modulator design and under appropriate operating conditions. Through symbol error probability analysis, we reveal that the very large number of antennas in massive MIMO provides favorable operating conditions for $\Sigma\Delta$ precoding. In addition, we develop a new $\Sigma\Delta$ modulation architecture that is capable of adapting the channel to achieve nearly zero quantization error for a targeted user. Furthermore, we consider multi-user $\Sigma\Delta$ precoding using the zero-forcing and symbol-level precoding schemes. These two $\Sigma\Delta$ precoding schemes perform considerably better than their direct one-bit quantized counterparts, as simulation results show.**

***Index Terms*— massive MIMO, one-bit MIMO, Sigma-Delta modulation, MIMO precoder design**

I. Introduction

Recently there has been growing interest in coarsely quantized multi-input multi-output (MIMO) transceiver implementations for massive MIMO communications systems that employ very large antenna arrays. These studies are strongly motivated by the need to reduce the hardware cost and power consumption of radio-frequency (RF) front-ends—which grow rapidly under massive MIMO—and the idea is to use low-resolution analog-to-digital converters (ADCs)/digital-to-analog converters (DACs) and energy-efficient low-dynamic-range power amplifiers. A number of researchers have investigated MIMO channel estimation and MIMO detection using one-bit or low-resolution ADCs [1]–[7], and it has been found that the very large number of antennas in massive MIMO indeed helps recover information lost due to the coarsely quantized signals.

MIMO precoding using one-bit DACs is another emerging topic in this area. A natural direction is to simply quantize the output of a conventional linear precoder, such as zero forcing (ZF), and the question is how the coarse quantization effects impact system performance [8]–[10] using, for example, the Bussgang decomposition as an analysis tool. More recently, there has been emphasis on directly designing a one-bit precoder, rather than following the aforementioned precode-then-quantize direction. The direct one-bit precoding designs use criteria such as minimum mean-square error and minimum symbol error probability [11]–[18], and numerically these designs were found to yield significantly improved performance. The challenge with direct one-bit precoding design is mainly centered on the optimization, which requires finding a good non-convex algorithm to handle a large-scale binary optimization problem. Promising numerical results have been reported with the direct one-bit precoding designs, but there is still much to be understood, *e.g.*, are the good numerical results an indication that most of the local minima have good quality, and if yes when can we guarantee this to happen? We refer the reader to [17], [18] for further descriptions of the various design approaches.

Since we have mentioned one-bit ADCs/DACs for MIMO, we should also mention the classical one-bit approach for analog-to-digital conversion—Sigma-Delta ($\Sigma\Delta$) modulation. The $\Sigma\Delta$ modulation approach exploits the use of oversampled, or low-frequency, signals in order to reduce the impact of the quantization noise. The $\Sigma\Delta$ principle is to employ a feedback loop to quantize the accumulated error between the input and the one-bit quantized output. The net effect is to shape the quantization noise to the high end of the frequency spectrum, where it can be separated from the signal of interest using a simple low-pass filter and decimator. For background on the $\Sigma\Delta$ approach and its various generalizations, the reader is referred to the tutorial article [19].

Alternatively, or in addition to quantization noise shaping in temporally oversampled systems, one can employ the $\Sigma\Delta$ effect using signals oversampled in space using an array of antennas. In such *spatial* $\Sigma\Delta$ architectures, the feedback signal is derived from the delayed and quantized outputs of adjacent antennas rather than or in addition to those of the given antenna. Oversampling in this context means that the elements of a uniform linear array would be spaced closer than one-half wavelength apart. As a result, the quantization error can be pushed to higher spatial frequencies, mitigating the distortion for signals of interest that might arrive from lower spatial frequencies, i.e., those near the broadside of the array. This idea has been exploited recently by a number of researchers [20]–[23]. Venkateswaran and van der Veen [24] use the concept in a different way, by beamforming the one-bit ADC outputs and using this as the feedback signal to each antenna, with the goal of removing interfering sources. The spatial $\Sigma\Delta$ approach should not be confused with the multi-antenna architecture of [25], in which each antenna output is modulated by a different Hadamard sequence prior to $\Sigma\Delta$ quantization in time. This is a variation of the approach originally proposed in [26], that uses a parallel bank of $\Sigma\Delta$ ADCs in order to obviate the need for temporal oversampling.

The $\Sigma\Delta$ idea has also been used for transmit signal pro-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

cessing. Scholnik *et al.* [27] use space-time $\Sigma\Delta$ DACs to generate one-bit outputs that directly drive each of the antennas, focusing the resulting quantization noise to directions and frequencies that do not impact the signal at the desired receiver. Krieger *et al.* [28] considered designs of analog beamforming weights for phased arrays when low-resolution phase shifters are employed. The goal there is to reduce the error that results from quantization of the weights of a transmit beamformer, and the weights are generated via $\Sigma\Delta$ quantization assuming a "dense" (oversampled) linear array.

Curiously, to the best of the authors' knowledge, the current developments of one-bit massive MIMO precoding do not seem to have touched upon the possibility of spatial $\Sigma\Delta$ modulation. It is therefore interesting to explore and understand what opportunities spatial $\Sigma\Delta$ modulation can bring to one-bit massive MIMO precoding—this is the main objective of this paper. We summarize our contributions, and compare them with existing literature, below.

1. Our study reveals that one-bit massive MIMO precoding using spatial $\Sigma\Delta$ modulation, or simply $\Sigma\Delta$ precoding for short, allows us to effectively mitigate the quantization noise effects. More precisely, we consider uniform linear arrays with user angles being within a certain "tolerable" range, say, $[-30^\circ, 30^\circ]$. We show that the quantization noise can be substantially suppressed when the number of antennas is large. This conclusion resembles that for analog beamforming by Krieger *et. al* [28], although the context of this work is completely different from that of [28].
2. We generalize the concept of spatial $\Sigma\Delta$ modulation. The spatial $\Sigma\Delta$ modulation concept used in the aforementioned literature usually considers direct application of the basic $\Sigma\Delta$ modulation for low-pass temporal signals. In this application, the best noise shaping result, in terms of nearly zero quantization noise, is possible only when the signal of interest comes from the broadside. We question whether the broadside angle can be altered. We develop a new $\Sigma\Delta$ modulation architecture whose angle for nearly zero quantization noise can be changed to any angle, and in the single-user case this new modulator allows us to adapt the user angle for achieving nearly zero quantization noise. Furthermore, we generalize this angle-steering concept to any type of channel, rather than just the angular channel.
3. The $\Sigma\Delta$ precoding approach allows us to revisit the easier precode-then-quantize approach, this time with much better controlled quantization noise. We show that the "precode" part of the precode-then-quantize operation is to design precoders under peak amplitude constraints. Leveraging this advantage, we develop multi-user $\Sigma\Delta$ precoding schemes using ZF and symbol-level precoding (SLP) for both the PSK and QAM cases. Efficient optimization algorithms for SLP, with the design emphasis of operating under the assumption of a large number of antennas, are also derived.

The organization of this paper is as follows. Section II describes the massive MIMO one-bit precoding problem. Section III reviews the basics of $\Sigma\Delta$ modulation. Sections IV and V describe our $\Sigma\Delta$ precoding developments for the single-user and multi-user cases, respectively. Section VI provides simulation results. Section VII concludes this work.

II. Problem Settings

The scenario we consider is the multiuser MISO downlink over a quasi-static frequent-flat channel and under one-bit transmitted signal constraints. The model is given by

$$y_{i,t} = \sqrt{\frac{P}{2N}} \boldsymbol{h}_i^T \boldsymbol{x}_t + v_{i,t}, \quad t = 1, \ldots, T, \tag{1}$$

and for $i = 1, \ldots, K$, where $y_{i,t} \in \mathbb{C}$ represents the complex baseband received signal of the ith user at symbol time t; K denotes the number of users; T is the transmission block length; P is the total transmission power; N is the number of antennas of the BS; $\boldsymbol{h}_i \in \mathbb{C}^N$ is the channel from the BS to the ith user; $\sqrt{P/(2N)}\boldsymbol{x}_t$, with $\boldsymbol{x}_t \in \{\pm 1 \pm \mathrm{j}\}^N$, represents the complex baseband one-bit transmitted signal; $v_{i,t}$ is noise and is assumed to be i.i.d. circular complex Gaussian with mean zero and variance σ_v^2.

The BS aims to blast parallel data symbols to the users. To put into context, let $s_{i,t} \in \mathcal{S}$ denote the symbol to be transmitted to the ith user at symbol time t, where $\mathcal{S}$ denotes the symbol constellation set. For convenience with our development later, we will assume that

$$\max_{s \in \mathcal{S}} |s| = 1;$$

or, the symbols are normalized such that the above equation holds. The challenge is to find $\boldsymbol{x}_t \in \{\pm 1 \pm \mathrm{j}\}^N$, for $t = 1, \ldots, T$, such that

$$\boldsymbol{h}_i^T \boldsymbol{x}_t \approx c_{i,t} s_{i,t}, \quad \text{for all } i, t, \tag{2}$$

where $c_{i,t} > 0$ denotes a scaling factor; or, in words, we aim to shape the symbols at the user side under the one-bit transmitted signal constraints. As a more technical note, we should mention that i) if the decision of the symbols at the user side depends on the signal amplitude, e.g., M-ary QAM, we should also make $c_{i,1} = \cdots = c_{i,T}$ for every i; see [17], [18], [29] (also [15] for a further discussion); and that ii) if the decision involves only signal phase, e.g., M-ary PSK, the $c_{i,t}$'s are allowed to be different. In the currently available literature, this one-bit precoding challenge is formulated as a binary optimization problem—which is hard to solve by nature. For details, read the recently growing body of literature [12], [13], [16]–[18], [30], [31].

We are interested in the single-path angular array channel. The settings that lead to such channels are that the antennas at the BS are arranged as a uniform linear array, and that there is only one propagation path from the BS to each user; the extension to other channels will be given later. For the single-path angular channel, each $\boldsymbol{h}_i$ is characterized as

$$\boldsymbol{h}_i = \alpha_i \boldsymbol{a}(\theta_i), \tag{3}$$

where $\alpha_i \in \mathbb{C}$ is the complex channel gain; $\theta_i \in [-\pi/2, \pi/2]$ denotes the angle of departure from the BS to the ith user;

$$\boldsymbol{a}(\theta) = [\ 1, e^{-\mathrm{j}\frac{2\pi d}{\lambda}\sin(\theta)}, \ldots, e^{-\mathrm{j}(N-1)\frac{2\pi d}{\lambda}\sin(\theta)}\ ]^T \tag{4}$$

denotes the array response vector at θ, in which λ is the carrier wavelength and $d \leq \lambda/2$ is the inter-antenna spacing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

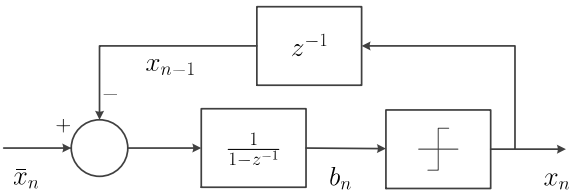


Fig. 1: The first-order $\Sigma\Delta$ modulator.

III. Basics of $\Sigma\Delta$ Modulation

In this section we review the basic concepts of $\Sigma\Delta$ modulation [19]. We will focus on the notion of noise shaping, and will pay less attention to aspects that have little relevance to the one-bit precoding context. Consider the system in Fig. 1, which is called the first-order $\Sigma\Delta$ modulator. We have a discrete-time real-valued signal sequence $\{\bar{x}_n\}_{n\in\mathbb{Z}_+}$ as the modulator input. In the application of temporal DACs, $\bar{x}_n$ is a significantly oversampled version of some signal. Here, it is sufficient to know that $\bar{x}_n$ is a low-pass signal. The problem is to one-bit quantize $\{\bar{x}_n\}_n$ in a way that the resulting quantization noise is high-pass. Doing so satisfactorily will result in negligible quantization noise effects on the low-pass frequency region of the desired signal $\bar{x}_n$. The $\Sigma\Delta$ modulator output sequence, denoted by $\{x_n\}_{n\in\mathbb{Z}_+}$, is generated as

$$x_n = \mathrm{sgn}(b_n), \tag{5a}$$

$$b_n = b_{n-1} + (\bar{x}_n - x_{n-1}), \tag{5b}$$

for $n = 0, 1, \ldots$, and with $b_{-1} = x_{-1} = 0$. Let $q_n = x_n - b_n$, $n \in \mathbb{Z}_+$, denote the quantization noise, and let $q_{-1} = 0$ for convenience. From (5) one can show that

$$x_n = \bar{x}_n + q_n - q_{n-1}, \quad n \in \mathbb{Z}_+,$$

and subsequently

$$X(z) = \bar{X}(z) + (1 - z^{-1})Q(z),$$

where $X(z) = \sum_{n=0}^{\infty} x_n z^{-n}$ denotes the z-transform. Since $1 - z^{-1}$ is a high-pass response, the quantization noise is suppressed at low frequency.

A key issue in $\Sigma\Delta$ modulation is the effect of overloading. Overloading refers to the situation when the quantizer input b_n has amplitude greater than 2. The consequence is that the corresponding quantization noise q_n goes beyond the range $[-1, 1]$. As an example of showing what problem overloading can bring, consider

$$\bar{x}_n = 1 + \epsilon, \quad \text{for all } n \in \mathbb{Z}_+,$$

where $\epsilon > 0$. This is an instance in which the signal amplitude is greater than one. One can verify from (5) that $b_n = 1 + (n+1)\epsilon$ and $q_n = -(n+1)\epsilon$. We see that the quantization noise is unbounded as $n \to \infty$. A sufficient condition under which overloading can be safely avoided is to limit the input signal range as

$$-1 \le \bar{x}_n \le 1, \quad \text{for all } n \in \mathbb{Z}_+. \tag{6}$$

Under the above condition it is guaranteed that $|b_n| \le 2$ for all $n \in \mathbb{Z}_+$, and consequently,

$$-1 \le q_n \le 1, \quad \text{for all } n \in \mathbb{Z}_+.$$

To see this, suppose $|b_{n-1}| \le 2$. Then, we see from (5b) that

$$|b_n| \le |\bar{x}_n| + |b_{n-1} - x_{n-1}| \le 2,$$

where we have used $|b_{n-1} - x_{n-1}| \le 1$, implied by (5a).

Under the no-overload condition (6), it is very common to assume that the quantization noise q_n is i.i.d., uniformly distributed on $[-1, 1]$, and independent of $\{\bar{x}_n\}$. This assumption is widely adopted for signal-to-quantization-noise ratio (SQNR) prediction in the $\Sigma\Delta$-DAC/ADC literature. We should, however, emphasize that the uniform i.i.d. assumption is only a convenient approximation for the sake of tractable analysis. Quantization noise analysis in $\Sigma\Delta$ modulation is a complicated topic, and we refer the reader to the Introduction of [32] which provides an excellent discussion. Simply speaking, from a theoretical viewpoint, $\Sigma\Delta$ quantization noise analysis is very difficult owing to the feedback and coarse quantization nature of the $\Sigma\Delta$ modulator. Some analysis results are available, e.g., in [32] and the references therein, but they are very complicated for practical use. From a practical viewpoint, it has been found by experiments and simulations that the uniform i.i.d. assumption yields reasonable approximations in many applications, but it can also be a poor approximation for some specific signals. For the latter the remedial solution is to apply dithering, which will be discussed later. In this paper we will apply the uniform i.i.d. assumption, and the reader should bear in mind that the uniform i.i.d. assumption can fail sometimes.

There are three further aspects we would like to discuss. First, while the no-overload condition (6) is widely adopted for ensuring bounded quantization noise, overloading does not necessarily imply unbounded quantization noise. An example is $\bar{x}_n = (-1)^n(1+\epsilon)$ for some $0 < \epsilon < 1$. It can be verified that $q_n = -\epsilon$ for even n, and $q_n = 0$ for odd n. In fact, one can argue that a moderate amount of overloading could be acceptable in practice, since not all kinds of overloaded input signals trigger the occurrence of unbounded quantization noise. For example, the second-order $\Sigma\Delta$ modulator [19] cannot avoid overloading for any input signal range (unless $\bar{x}_n = 0$ for all n) [32], and yet it is still used in practice. That being said, there seems to be little theoretical work on understanding the quantization noise bound under overloading.

Second, we previously mentioned that the uniform i.i.d. assumption is far from true for some specific signals. Among them, DC and pure sinusoidal signals are most well-known [32], [33]. A popular way to handle the non-i.i.d. issue is to apply dithering. For example, as described in [33], consider modifying (5a) as

$$x_n = \mathrm{sgn}(b_n + u_n), \tag{7}$$

where u_n, called a dither signal, is uniform i.i.d. generated on $[-\delta, \delta]$ for some constant $\delta > 0$. Intuitively, the idea is to use artificial noise to make the overall quantization noise $q_n = x_n - b_n$ more random, thereby attempting to destroy correlated patterns that q_n may exhibit in the no-dithering case. Empirically, it has been found that dithering works to a certain extent [33]. However, dithering also increases the quantization noise level. It can be verified from (5b), (6) and (7) that $-1 - \delta \le q_n \le 1 + \delta$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

Third, better noise-shaping, in terms of further suppressing the low-pass region of the quantization noise, can be achieved by employing more advanced $\Sigma\Delta$ modulators, e.g., the higher-order and multi-stage versions of the first-order $\Sigma\Delta$ modulator. The issue arising would be with overloading, however, and in some cases multi-bit quantization is used to avoid overloading.

Readers are referred to the literature [19], [32] for further details of the above three aspects. To keep our forthcoming development simple, we will consider only the first-order $\Sigma\Delta$ modulator without overloading and without dithering, unless otherwise specified.

IV. $\Sigma\Delta$ Precoding: Single-User Case

This section and the subsequent sections describe how we apply $\Sigma\Delta$ modulation to perform one-bit precoding. In this section we consider the single-user case.

A. Spatial $\Sigma\Delta$ Modulation

Consider the basic model (1) for the single-user case. For the sake of notational simplicity, we remove the time index t and user index i from (1) and write

$$y = \sqrt{\frac{P}{2N}}\boldsymbol{h}^T\boldsymbol{x} + v, \tag{8}$$

with $\boldsymbol{h} = \alpha\boldsymbol{a}(\theta)$; θ is the user's angle. Let $\bar{\boldsymbol{x}} = [\ \bar{x}_1, \ldots, \bar{x}_N\ ]^T$, with $-1 \le \Re(\bar{x}_n) \le 1$ and $-1 \le \Im(\bar{x}_n) \le 1$ for all n, be the signal we wish to $\Sigma\Delta$-modulate. We apply first-order $\Sigma\Delta$ modulation (as described in the preceding section) to $\{\bar{x}_n\}_{n=1}^N$ to obtain $\{x_n\}_{n=1}^N$. The resulting $\boldsymbol{x} = [\ x_1, \ldots, x_N\ ]^T$ then serves as the one-bit transmitted signal. More precisely, we use two first-order $\Sigma\Delta$ modulators, one for the real part and another for the imaginary part, to get $\boldsymbol{x}$. By doing so, we perform $\Sigma\Delta$ modulation in space. The advantages of doing so will become clear as we analyze the subsequent quantization noise effects below.

Following the preceding section, we can write

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{q} - \boldsymbol{q}^- \tag{9}$$

where $\boldsymbol{q} = [\ q_1, q_2 \ldots, q_N\ ]^T$; $\boldsymbol{q}^- = [\ 0, q_1, \ldots, q_{N-1}\ ]^T$; each q_i is complex quantization noise with $-1 \le \Re(q_n) \le 1$ and $-1 \le \Im(q_n) \le 1$ (the aforesaid noise range is guaranteed when $-1 \le \Re(\bar{x}_n) \le 1$ and $-1 \le \Im(\bar{x}_n) \le 1$). For the sake of analysis, we model the q_n's as i.i.d. uniform noise on the unit box interval $\{q = a + \mathrm{j}b \mid a, b \in [-1, 1]\}$. Putting (9) into (8) gives

$$y = \sqrt{\frac{P}{2N}}\boldsymbol{h}^T\bar{\boldsymbol{x}} + w, \tag{10a}$$

$$w = \sqrt{\frac{P}{2N}}\boldsymbol{h}^T(\boldsymbol{q} - \boldsymbol{q}^-) + v, \tag{10b}$$

where w denotes a noise term that combines quantization noise and background noise. We are interested in knowing how the noise power scales with the system parameters. Let $z = e^{\mathrm{j}\frac{2\pi d}{\lambda}\sin(\theta)}$ for convenience. We see that

$$\boldsymbol{a}^T(\boldsymbol{q} - \boldsymbol{q}^-) = (1 - z^{-1})\sum_{n=0}^{N-2} z^{-n}q_{n+1} + z^{-(N-1)}q_N,$$

and consequently, $\mathbb{E}[\boldsymbol{a}^T(\boldsymbol{q} - \boldsymbol{q}^-)] = 0$ and

$$\mathbb{E}[|\boldsymbol{a}^T(\boldsymbol{q} - \boldsymbol{q}^-)|^2] = |1 - z^{-1}|^2(N-1)\sigma_q^2 + \sigma_q^2,$$

where $\sigma_q^2 = \mathbb{E}[|q_n|^2] = 2/3$ due to the assumption of uniform i.i.d. quantization noise. It follows that $\mathbb{E}[w] = 0$ and

$$\sigma_w^2 = \mathbb{E}[|w|^2] = \frac{|\alpha|^2 P}{3N}(|1 - z^{-1}|^2(N-1) + 1) + \sigma_v^2$$

By assuming large N, the above quantization noise variance formula can be simplified to

$$\sigma_w^2 \approx \frac{|\alpha|^2 P}{3}|1 - z^{-1}|^2 + \sigma_v^2 \tag{11a}$$

$$= \frac{4|\alpha|^2 P}{3}\left|\sin\left(\frac{\pi d}{\lambda}\sin(\theta)\right)\right|^2 + \sigma_v^2. \tag{11b}$$

Eq. (11b) reveals interesting behaviors with the quantization noise effects at the user side.

1. First, the quantization noise power at the user side is independent of the number of antennas N. This will give us substantial advantages in using massive MIMO to suppress the quantization noise, as we will further show in the next subsection.
2. Second, the quantization noise power increases as the absolute value of the angle $|\theta|$ increases; broadside ($\theta = 0$) is the best, while endfire ($\theta = \pi/2$ or $\theta = -\pi/2$) is the worst. This suggests that spatial $\Sigma\Delta$ modulation serves users with smaller $|\theta|$ better. This also suggests that if we work on sectored antenna arrays, where we only need to deal with a restricted angular range, say, from -30° to 30°, spatial $\Sigma\Delta$ modulation has an advantage.
3. Third, the quantization noise power decreases as we decrease the inter-antenna spacing d. This means that we may want to employ more densely spaced antennas. In practice, however, it is infeasible to have very small inter-antenna spacing as that will introduce mutual coupling effects. Also, the physical dimensions of the antennas prevent small spacing. We will have to rely on large N and smaller operating angular ranges to reduce the quantization noise.

A further comment is as follows.

Remark 1 We should also draw connections between conventional $\Sigma\Delta$ modulation for discrete-time signals and the spatial $\Sigma\Delta$ modulation proposed above. Simply speaking, frequency in the temporal case becomes angle in the spatial case. $\Sigma\Delta$ modulation in time and space serve low frequency and low angle signals better, respectively. Also, applying small d in the spatial case is essentially the same as oversampling in the temporal case. In fact, the latter typically considers a very large oversampling factor, such as 128, such that quantization noise becomes almost negligible [19]. Such extreme oversampling is however inapplicable to the spatial case; as mentioned above, mutual coupling and the physical dimension constraint prevent us from doing so.

B. $\Sigma\Delta$ Maximum Ratio Transmission

In the preceding subsection we have presented a different paradigm to deal with one-bit precoding: Using spatial $\Sigma\Delta$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

modulation, we can convert the one-bit precoding problem to a precoding problem for an amplitude-limited signal $\bar{\boldsymbol{x}}$, specifically, $-\mathbf{1} \leq \Re(\bar{\boldsymbol{x}}) \leq \mathbf{1}$ and $-\mathbf{1} \leq \Im(\bar{\boldsymbol{x}}) \leq \mathbf{1}$. Let us consider a simple precoding scheme, namely, the maximum ratio transmission (MRT) approach

$$\bar{\boldsymbol{x}} = \frac{\alpha^* s}{|\alpha|} \boldsymbol{a}^*(\theta), \tag{12}$$

where $s \in \mathcal{S}$ is a symbol. Note that $\bar{\boldsymbol{x}}$ satisfies the aforementioned amplitude-limit constraints since $|[\boldsymbol{a}(\theta)]_n| = 1$ for all n and $|s| \leq 1$. We are interested in performing a symbol-error probability (SEP) analysis of this $\Sigma\Delta$ MRT scheme. Plugging (12) into the model (10), we get

$$y = c \cdot s + w, \quad c = |\alpha| \sqrt{\frac{PN}{2}}.$$

Let us make an approximation, namely, that w is circular Gaussian distributed with mean zero and variance given by (11b). Let $\hat{s} = \mathrm{dec}(y)$ be the decision of s, where dec denotes the decision function associated with $\mathcal{S}$. The SEP can be characterized as

$$\mathrm{P}(\hat{s} \neq s) \leq \beta Q\left(\chi_{_M} \sqrt{\mathsf{SNR}_{\mathsf{eff}}}\right), \tag{13}$$

where $(\beta, \chi_{_M}) = (2, \sqrt{2}\sin(\pi/M))$ if $\mathcal{S}$ is the M-ary PSK constellation set, and $(\beta, \chi_{_M}) = (4, 1/(\sqrt{M}-1))$ if $\mathcal{S}$ is the M-ary QAM constellation set and M is a power of 4; $Q(t) = \int_t^\infty (e^{-z^2/2}/\sqrt{2\pi}) dz$;

$$\mathsf{SNR}_{\mathsf{eff}} = \frac{c^2}{\sigma_w^2}$$

denotes the effective SNR [34]. The effective SNR plays the main role in determining the SEP performance. From the above derivations, we see that

$$\mathsf{SNR}_{\mathsf{eff}} = \frac{|\alpha|^2 PN}{\frac{8|\alpha|^2 P}{3} \left|\sin\left(\frac{\pi d}{\lambda}\sin(\theta)\right)\right|^2 + 2\sigma_v^2}. \tag{14}$$

Let us extract some insights from the effective SNR derivation (14).

1. First, increasing the power P is not helpful in reducing quantization noise power. In fact, we have $\lim_{P\to\infty} \mathsf{SNR}_{\mathsf{eff}} = 3N/(8\left|\sin\left(\frac{\pi d}{\lambda}\sin(\theta)\right)\right|^2))$.
2. Second, the effective SNR increases linearly with the number of antennas N. In particular we observe that under a fixed power P, increasing N—which also means less power per antenna—is effective in improving the effective SNR. This suggests that $\Sigma\Delta$ precoding is particularly suitable for massive MIMO.

In the supplementary material of this paper, we provide additional numerical results to give readers some intuitive feeling on the noise shaping performance of $\Sigma\Delta$ MRT. One will see that, in general, the symbol shaping error of $\Sigma\Delta$ MRT reduces with N and increases with $|\theta|$.

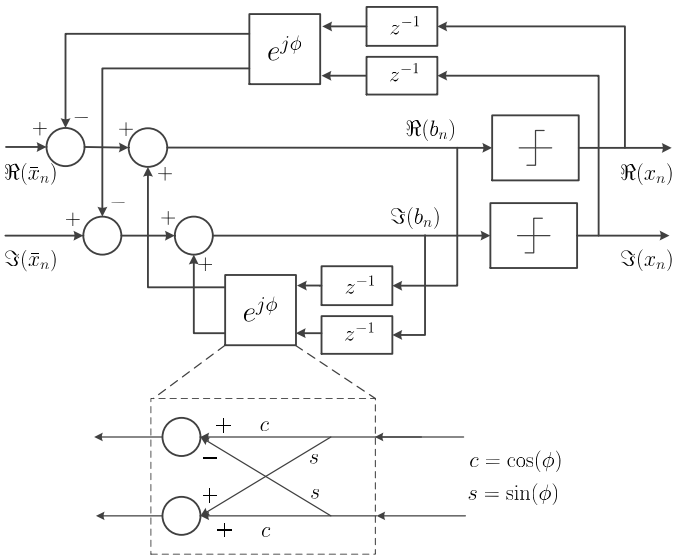


Fig. 2: The angle-steered first-order $\Sigma\Delta$ modulator.

C. Quantization Noise Zeroing by $\Sigma\Delta$ Angle Steering

We have seen that the quantization noise tends to increase as the angle θ is further away from 0. It is natural to question whether we can reduce the quantization noise by re-designing the $\Sigma\Delta$ modulator. The answer turns out to be yes.

Our idea borrows insight from bandpass $\Sigma\Delta$ modulation [19], although our task is still different from that of the latter. Consider the modified first-order $\Sigma\Delta$ modulator in Fig. 2, which we refer to as an *angle-steered* $\Sigma\Delta$ modulator. In this system, $\bar{x}_n$, b_n and x_n are all complex-valued, and $\phi \in [-\pi, \pi]$ is given. The modulation process is described by

$$x_n = \mathrm{sgn}(\Re(b_n)) + \mathrm{j} \cdot \mathrm{sgn}(\Im(b_n)), \tag{15a}$$
$$b_n = e^{\mathrm{j}\phi} b_{n-1} + (\bar{x}_n - e^{\mathrm{j}\phi} x_{n-1}), \tag{15b}$$

Let $q_0 = 0$, and let $q_n = x_n - b_n$ be the quantization noise. From (15) one can show that

$$x_n = \bar{x}_n + q_n - e^{\mathrm{j}\phi} q_{n-1}, \tag{16}$$

where the difference compared with the previous first-order $\Sigma\Delta$ modulator is the inclusion of the phase shift term $e^{\mathrm{j}\phi}$. We are concerned with the range of $\bar{x}_n$ under which no overloading will occur.

Fact 1 *Consider the angle-steered $\Sigma\Delta$ modulator in Fig. 2 or in* (15)*. Let*

$$A = 2 - |\cos(\phi)| - |\sin(\phi)|. \tag{17}$$

If $|\Re(\bar{x}_n)| \leq A$ and $|\Im(\bar{x}_n)| \leq A$ for all n, then b_n is not overloaded, and the quantization noise q_n is bounded with $|\Re(q_n)| \leq 1$ and $|\Im(q_n)| \leq 1$.

Proof: We prove Fact 1 by induction. Assume $|\Re(\bar{x}_n)| \leq A$ and $|\Im(\bar{x}_n)| \leq A$ for all n. It is easy to see that $|\Re(q_1)| \leq 1$ and $|\Im(q_1)| \leq 1$. Now, suppose that $|\Re(q_{n-1})| \leq 1$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

$|\Im(q_{n-1})| \le 1$ are true. Using $b_n = \bar{x}_n - e^{\mathrm{j}\phi} q_{n-1}$, which can be shown from (15), we have

$$\begin{aligned}|\Re(b_n)| &\le |\Re(\bar{x}_n)| + |\cos(\phi)\Re(q_{n-1})| + |\sin(\phi)\Im(q_{n-1})| \\ &\le A + |\cos(\phi)| + |\sin(\phi)| = 2,\end{aligned}$$

and similarly, $|\Im(b_n)| \le 2$. Consequently, we must have $|\Re(q_n)| \le 1$ and $|\Im(q_n)| \le 1$. The proof is complete. ■

We should mention that the largest value of A is $A = 1$, which happens when $\phi \in \{0, \pm\pi/2, \pm\pi\}$. The smallest value of A is $A = 0.59$, which happens when $\phi \in \{\pm\pi/4, \pm 3\pi/4\}$. This means that there is a mild compromise with the signal range if no overloading is desired.

However, the aforementioned compromise brings a significant advantage, namely, quantization noise zeroing. Following the same noise analysis in Section IV-A, we can show that

$$\begin{aligned}\sigma_w^2 &\approx \frac{|\alpha|^2 P}{3}|1 - e^{\mathrm{j}\phi} z^{-1}|^2 + \sigma_v^2 \\ &= \frac{4|\alpha|^2 P}{3}\left|\sin\left(\frac{\phi - \frac{2\pi d}{\lambda}\sin(\theta)}{2}\right)\right|^2 + \sigma_v^2.\end{aligned}$$

Hence, by selecting $\phi = 2\pi d \sin(\theta)/\lambda$, we can eliminate the quantization noise effects. To get more insight, let us consider MRT under such angle-steered $\Sigma\Delta$ modulation. The corresponding MRT scheme is $\bar{\boldsymbol{x}} = \frac{A\alpha^* s}{|\alpha|}\boldsymbol{a}(\theta)$. The effective SNR under angle steering is

$$\mathsf{SNR}_{\mathsf{eff}} = \frac{A^2|\alpha|^2 PN}{2\sigma_v^2}, \tag{18}$$

with $A = 2 - |\cos(2\pi d\sin(\theta)/\lambda)| - |\sin(2\pi d\sin(\theta)/\lambda)|$. We see that the sole factor of performance reduction is A, which is reduced to 0.59 (equivalently, -4.64dB SNR loss relative to $A = 1$) in the worst case. Thus, we see that the angle corresponding to the minimum quantization noise in the previous $\Sigma\Delta$ modulator, that is, the broadside angle $\theta = 0$, can be steered to any desired angle using the angle-steered $\Sigma\Delta$ modulation approach.

Again, to give readers some intuition, the supplementary material of this paper provides an additional numerical result that shows that the angle-steered $\Sigma\Delta$ modulation approach leads to almost zero symbol shaping error.

Remark 2 It is worthwhile to note that the angle-steered $\Sigma\Delta$ MRT scheme described above does not require the uniform i.i.d. assumption with the quantization noise. From (10), (16), and with $\phi = 2\pi d\sin(\theta)/\lambda$, one can show that the overall noise term w is actually given by

$$w = \sqrt{\frac{P}{2N}}\alpha z^{N-1} q_N + v;$$

we will show the details and insight of the above expression under a more general setting in the subsequent subsection. Note that the same phenomenon also happens with the basic $\Sigma\Delta$ MRT scheme when the user angle is $\theta = 0$. As such, there is no need to assume that the q_n's are i.i.d., and the remaining factor lies only in the surviving quantization noise term q_N in the above equation. That surviving term is small compared with the signal term for large N, and thus may be ignored.

Remark 3 The angle-steered $\Sigma\Delta$ modulation architecture can be used to change the angular range the system serves. Previously, we mentioned that the basic spatial $\Sigma\Delta$ modulation is more appropriate for serving users under a smaller angular range, say, from -30° to 30°. Now, with angle steering, we can easily alter the center of the angular range, say, to 60°, thereby serving users from 30° to 90°.

D. Angle-Steered $\Sigma\Delta$ Modulation for Any Channels

It is intriguing to further question whether the angle steering idea in the last subsection can be generalized to any arbitrary channel $\boldsymbol{h}$, rather than just the one-path angular channel under uniform linear arrays. The answer turns to be also yes.

Without loss of generality, assume $h_n \neq 0$ for all n. Also, assume the elements of the antenna array to be indexed such that $0 < |h_1| \le |h_2| \le \cdots \le |h_N|$. Consider modifying the angle-steered $\Sigma\Delta$ modulator (15) as follows:

$$x_n = \mathrm{sgn}(\Re(b_n)) + \mathrm{j}\cdot\mathrm{sgn}(\Im(b_n)), \tag{19a}$$

$$b_n = \tfrac{h_{n-1}}{h_n} b_{n-1} + \left(\bar{x}_n - \tfrac{h_{n-1}}{h_n} x_{n-1}\right), \tag{19b}$$

for $n = 1, \ldots, N$ and with $h_0 = 0$. From the above equations, one can readily show that

$$x_n = \bar{x}_n + q_n - \tfrac{h_{n-1}}{h_n} q_{n-1}, \tag{20}$$

where $q_0 = 0$; $q_n = x_n - b_n$ for $n = 1, \ldots, N$. By observing

$$\begin{aligned}\boldsymbol{h}^T\boldsymbol{x} &= \sum_{n=1}^{N} h_n \bar{x}_n + \sum_{n=1}^{N} h_n\left(q_n - \tfrac{h_{n-1}}{h_n} q_{n-1}\right) \\ &= \boldsymbol{h}^T\bar{\boldsymbol{x}} + h_N q_N,\end{aligned}$$

where the quantization noise terms $q_1, \ldots, q_{N-1}$ are successively canceled, the signal model reduces to

$$y = \sqrt{\frac{P}{2N}}\boldsymbol{h}^T\bar{\boldsymbol{x}} + w, \tag{21a}$$

$$w = \sqrt{\frac{P}{2N}} h_N q_N + v. \tag{21b}$$

Suppose that the $\Sigma\Delta$ modulator is not overloaded such that $|q_N| \le 1$. Then, for most massive MIMO cases of interest in which $|h_N| \ll \sum_{n=1}^{N-1}|h_n|$, the quantization noise term in w can be neglected. We call this modulator a *generalized angle-steered* $\Sigma\Delta$ modulator. The sufficient condition for no overloading is as follows.

Fact 2 *Consider the generalized angle-steered $\Sigma\Delta$ modulator in* (15a) *and* (19). *Let, for $n = 1, \ldots, N$,*

$$A_n = 2 - \tfrac{|h_{n-1}|}{|h_n|}(|\cos(\phi_n)| + |\sin(\phi_n)|), \tag{22}$$

where ϕ_n denotes the phase of h_{n-1}/h_n. If $|\Re(\bar{x}_n)| \le A_n$ and $|\Im(\bar{x}_n)| \le A_n$ for all n, then b_n is not overloaded, and the quantization noise q_n is bounded with $|\Re(q_n)| \le 1$ and $|\Im(q_n)| \le 1$.

The proof of Fact 2 is essentially the same as that of Fact 1, and we shall thus omit it. Note that $0.59 \le A_n < 2$. Also,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

since the signal range (22) varies with n, it makes sense to modify the MRT scheme accordingly:

$$\bar{\boldsymbol{x}}_n = \boldsymbol{r}s, \tag{23}$$

where $r_n = A_n h_n^* / \max\{|\Re(h_n)|, |\Im(h_n)|\}$ for all n.

V. ΣΔ Precoding: Multi-User Case

The study in the preceding section provides us with vital insights into how the performance of ΣΔ precoding scales with the system parameters, assuming a single user. Now we turn to the multi-user case.

The development follows exactly the same spirit as the preceding section. We simplify the notation of the basic signal model (1) by removing the index t, i.e.,

$$y_i = \sqrt{\frac{P}{2N}} \boldsymbol{h}_i^T \boldsymbol{x} + v_i, \quad i = 1, \ldots, K.$$

For simplicity, we apply ΣΔ modulation without angle steering. Adaptation to the angle-steered case is straightforward. The corresponding model is

$$y_i = \sqrt{\frac{P}{2N}} \boldsymbol{h}_i^T \bar{\boldsymbol{x}} + w_i, \quad i = 1, \ldots, K, \tag{24}$$

where $\bar{\boldsymbol{x}} \in \mathbb{C}^N$ is an amplitude-limited desired signal, with $-\mathbf{1} \le \Re(\bar{\boldsymbol{x}}) \le \mathbf{1}$ and $-\mathbf{1} \le \Im(\bar{\boldsymbol{x}}) \le \mathbf{1}$; w_i is a term combining quantization noise and background noise. The noise term w_i is modeled as mean-zero circular complex Gaussian. The variance of w_i, denoted by $\sigma_{w,i}^2$, is evaluated as

$$\sigma_{w,i}^2 = \frac{4|\alpha_i|^2 P}{3} \left| \sin\left(\frac{\pi d}{\lambda} \sin(\theta_i) \right) \right|^2 + \sigma_v^2, \tag{25}$$

where large N has again been assumed; note that (25) directly follows from the noise variance formula (11).

In the first two subsections below, we will describe two design schemes for $\bar{\boldsymbol{x}}$ under the assumption of M-ary PSK constellations. Then, the third subsection will consider the adaptation of the two schemes to the M-ary QAM constellation case. The final subsection will discuss the extension to the multi-path angular channel case.

A. ΣΔ Zero-Forcing

The first scheme we consider is ZF. For notational convenience, define

$$\|\boldsymbol{x}\|_{IQ-\infty} = \max\{|\Re(x_1)|, |\Im(x_1)|, \ldots, |\Re(x_N)|, |\Im(x_N)|\};$$

that is, the infinity norm applied on the in-phase and quadrature-phase components of a vector. Also, assume M-ary PSK constellations. The ZF precoding scheme implements

$$\bar{\boldsymbol{x}} = \gamma \boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}, \tag{26}$$

where $\boldsymbol{s} \in \mathcal{S}^K$ is the symbol vector, with s_i representing the symbol for the ith user;

$$\boldsymbol{D} = \mathrm{Diag}(\sigma_{w,1}\alpha_1^*/|\alpha_1|^2, \ldots, \sigma_{w,K}\alpha_K^*/|\alpha_K|^2),$$
$$\boldsymbol{A} = [\ \boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\ ]^T, \quad \boldsymbol{a}_i = \boldsymbol{a}(\theta_i);$$

and γ is a normalization constant such that $\|\bar{\boldsymbol{x}}\|_{IQ-\infty} = 1$. It is easy to see that

$$\gamma = \frac{1}{\|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}\|_{IQ-\infty}}. \tag{27}$$

This ZF precoding scheme is designed such that every user has the same effective SNR, and consequently, uniform SEP performance. To see this, consider putting (26) into (24). It can be shown that

$$y_i = c_i \cdot s_i + w_i, \quad c_i = \sqrt{\frac{P}{2N}} \gamma \sigma_{w,i}.$$

Following the effective SNR concept used in the preceding section, the effective SNR of the ith user is

$$\mathsf{SNR}_{\mathsf{eff},i} = \frac{c_i^2}{\sigma_{w,i}^2} = \frac{P}{2N}\gamma^2. \tag{28}$$

Clearly, the effective SNRs of all the users are identical.

In the simulation results section we will show the performance of this ΣΔ ZF precoding scheme. Here, we are interested in analyzing how the effective SNRs scale with the system parameters. The result is as follows.

Proposition 1 *Let $k = \arg\max_{i=1,\ldots,K} \sigma_{w,i}/|\alpha_i|$. The users' effective SNRs are bounded by*

$$\begin{aligned} \mathsf{SNR}_{\mathsf{eff},i} &\ge \frac{PN|\alpha_k|^2 \lambda_{\min}^2(\boldsymbol{R})}{2K^3 \sigma_{w,k}^2} \\ &= \frac{PN|\alpha_k|^2 \lambda_{\min}^2(\boldsymbol{R})}{2K^3 \left(\frac{4|\alpha_k|^2 P}{3} \left| \sin\left(\frac{\pi d}{\lambda} \sin(\theta_k) \right) \right|^2 + \sigma_v^2 \right)}, \end{aligned} \tag{29}$$

for all i, where $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}^H/N$; $\lambda_{\min}(\boldsymbol{R})$ denotes the smallest eigenvalue of $\boldsymbol{R}$. Also, it holds that

$$1 \ge \lambda_{\min}(\boldsymbol{R}) \ge 1 - (K-1)\rho, \tag{30}$$

where

$$\rho = \max_{i \ne j} \left| D_N\left(\frac{\pi d}{\lambda} (\sin(\theta_i) - \sin(\theta_j)) \right) \right|,$$

and $D_N(\phi) = \sin(N\phi)/(N\sin(\phi))$ is the digital sinc function.

Proof: From (27)–(28), we see that the problem is to analyze $\|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}\|_{IQ-\infty}$. Let $\|\cdot\|_p$ denote either the p-norm for vectors or the induced p-norm for matrices. We have

$$\begin{aligned} \|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}\|_{IQ-\infty} &\le \|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}\|_\infty \le \|\boldsymbol{A}^\dagger\|_\infty \|\boldsymbol{D}\boldsymbol{s}\|_\infty \\ &= \|\boldsymbol{A}^\dagger\|_\infty \max_i \sigma_{w,i}/|\alpha_i|, \end{aligned}$$

where we have used $\|\boldsymbol{x}\|_{IQ-\infty} \le \|\boldsymbol{x}\|_\infty$, $\|\boldsymbol{A}x\|_\infty \le \|\boldsymbol{A}\|_\infty \|\boldsymbol{x}\|_\infty$, and $|s_i| \le 1$ for all i. By using

$$\boldsymbol{A}^\dagger = \boldsymbol{A}^H (\boldsymbol{A}\boldsymbol{A}^H)^{-1} = \frac{1}{N} \boldsymbol{A}^H \boldsymbol{R}^{-1},$$

we further get

$$\begin{aligned} \|\boldsymbol{A}^\dagger\|_\infty &\le \frac{1}{N} \|\boldsymbol{A}^H\|_\infty \|\boldsymbol{R}^{-1}\|_\infty \\ &\le \frac{1}{N} K(\sqrt{K} \|\boldsymbol{R}^{-1}\|_2) \\ &= \frac{K^{3/2}}{N} \lambda_{\min}^{-1}(\boldsymbol{R}), \end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

where the first inequality is due to $\|\boldsymbol{XY}\|_\infty \leq \|\boldsymbol{X}\|_\infty \|\boldsymbol{Y}\|_\infty$; the second inequality is due to $\|\boldsymbol{A}^H\|_\infty = \max_i \sum_{j=1}^K |a_{ji}| = K$ and due to $\|\boldsymbol{X}\|_\infty \leq \sqrt{n}\|\boldsymbol{X}\|_2$ for $\boldsymbol{X} \in \mathbb{C}^{m\times n}$; the third inequality is due to the fact that for a positive definite $\boldsymbol{X}$, it is true that $\|\boldsymbol{X}\|_2 = \lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\max}(\boldsymbol{X}^{-1}) = 1/\lambda_{\min}(\boldsymbol{X})$. By putting the above inequalities into (27) and (28), the desired inequality (29) is obtained. The proof of (30) is relegated to Appendix A. ■

Let us discuss the implications of the theoretical result in (29)–(30). First, the quantization noise effects are the same as what we see in the single-user case; a larger absolute value of the angle means a larger quantization noise power. Second, the lower bound of the effective SNRs increases linearly with the number of antennas N. Again, this suggests that $\Sigma\Delta$ precoding is favorable for massive MIMO. Third, $\lambda_{\min}(\boldsymbol{R})$, which appears in the signal power part of the effective SNR, is large if the user angles are well separated, but small if some of the angles are close. This factor is relative to N. Fixing the angles, larger N brings $\lambda_{\min}(\boldsymbol{R})$ closer to its largest value, 1. Third, we are interested in how N should scale with the number of users K. Very intuitively, by reading (29), there is an indication that N should increase cubically with K; doing so keeps N/K^3 constant in the effective SNR bound. However, note that this is a prediction from a performance bound that is safe, but also pessimistic, by its nature. For instances where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K$ are orthogonal—which one can expect it to be approximately true when N is very large, one can redo the proof of Proposition 1 to obtain a better bound

$$\mathsf{SNR}_{\mathsf{eff},i} \geq \frac{PN|\alpha_k|^2}{2K\left(\frac{4|\alpha_k|^2 P}{3}\left|\sin\left(\frac{\pi d}{\lambda}\sin(\theta_k)\right)\right|^2 + \sigma_v^2\right)},$$

which is merely the single-user effective SNR (14) downscaled by K. In such instances it suffices to scale N linearly with K.

B. $\Sigma\Delta$ Symbol-Level Precoding

The second scheme we consider is SLP. The idea is to design, on a per-symbol-time basis, an amplitude-limited $\bar{\boldsymbol{x}}$ such that the SEP performance of the users is improved. It is interesting to first draw a connection between SLP and ZF. As shown in [35], any $\bar{\boldsymbol{x}} \in \mathbb{C}^N$ can be expressed as

$$\bar{\boldsymbol{x}} = \boldsymbol{A}^\dagger(\boldsymbol{Ds} + \boldsymbol{u}) + \boldsymbol{\eta}, \tag{31}$$

where $\boldsymbol{\eta}$ lies in the nullspace of $\boldsymbol{A}$, $\boldsymbol{D} = \mathrm{Diag}(\beta_1, \ldots, \beta_K)$, with $\beta_i > 0$ for all i, and $\boldsymbol{u} \in \mathbb{C}^K$. Putting (31) into the model (24) gives

$$y_i = \sqrt{\frac{P}{2N}}\alpha_i(\beta_i s_i + u_i) + w_i, \quad i = 1, \ldots, K, \tag{32}$$

where the nullspace term $\boldsymbol{\eta}$ has no impact on the received signals, the u_i's appear as symbol perturbation terms, and the β_i's appear as symbol gains. There are two main ideas. First, conditioned on s_i, we can use u_i to purposely push the shaped symbol away from the decision boundaries. SEP performance can thereby be improved. Second, while the nullspace term $\boldsymbol{\eta}$ seems useless at first glance, it plays a hidden role in improving energy efficiency. Intuitively, from (31), we may hope that some particular $\boldsymbol{\eta}$ can cancel some of the signal components of $\boldsymbol{A}^\dagger(\boldsymbol{Ds} + \boldsymbol{u})$, possibly reducing the subsequent IQ amplitude limit $\|\bar{\boldsymbol{x}}\|_{IQ-\infty}$. In the related context of per-antenna power constrained linear precoding, it has been alluded to that using the nullspace term can be beneficial [36].

Having shed light on the intuition, we turn to the design. We formulate the design as a minimax SEP problem. Assume M-ary PSK constellations. Let $\mathsf{SEP}_i = \mathrm{P}(\hat{s}_i \neq s_i)$, with $\hat{s}_i = \mathrm{dec}(y_i)$, be the SEP of the ith user. The problem is

$$\min_{\|\bar{\boldsymbol{x}}\|_{IQ-\infty}\leq 1} \max_{i=1,\ldots,K} \mathsf{SEP}_i. \tag{33}$$

Our first challenge is to find a tractable characterization of SEP_i. Consider the following result.

Lemma 1 ([37]) [1] *Let $\mathcal{S}$ be the M-ary PSK constellation set. Let $y = z + w$, where w is circular complex Gaussian with mean zero and variance σ_w^2, and $z \in \mathbb{C}$ is arbitrary. Let $s \in \mathcal{S}$, and let $\hat{s} = \mathrm{dec}(y)$. It holds that*

$$\mathrm{P}(\hat{s} \neq s) \leq 2Q\left(\chi_{_M}\frac{\psi}{\sigma_w}\right),$$

where $\chi_{_M} = \sqrt{2}\sin(\pi/M)$, and

$$\psi = \Re(zs^*) - |\Im(zs^*)|\cot(\pi/M).$$

Applying Lemma 1 to the signal model (24), we characterize the users' SEPs as

$$\mathsf{SEP}_i \leq 2Q(\chi_{_M}\sqrt{\mathsf{SNR}_{\mathsf{eff},i}}),$$

where

$$\mathsf{SNR}_{\mathsf{eff},i} = \frac{P(\Re(\boldsymbol{h}_i^T\bar{\boldsymbol{x}}s_i^*) - |\Im(\boldsymbol{h}_i^T\bar{\boldsymbol{x}}s_i^*)|\cot(\pi/M))}{2N\sigma_{w,i}^2}.$$

Since the above bound on SEP_i decreases as $\mathsf{SNR}_{\mathsf{eff},i}$ increases, and since this relationship is monotone, it makes sense to consider

$$\max_{\|\bar{\boldsymbol{x}}\|_{IQ-\infty}\leq 1} \min_{i=1,\ldots,K} \mathsf{SNR}_{\mathsf{eff},i} \tag{34}$$

as a convenient and reasonable approximation of the minimax SEP problem (33). As a slight abuse of notation, redefine the variable $\boldsymbol{x}$ as

$$\boldsymbol{x} = [\ \Re(\bar{\boldsymbol{x}})^T, \Im(\bar{\boldsymbol{x}})^T\ ]^T.$$

Through some efforts, problem (34) can be rewritten as

$$\min_{\boldsymbol{x}\in[-1,1]^{2N}} f(\boldsymbol{x}) \triangleq \max\{\boldsymbol{c}_1^T\boldsymbol{x}, \cdots, \boldsymbol{c}_{2K}^T\boldsymbol{x}\}, \tag{35}$$

where

$$\boldsymbol{c}_i = \begin{cases} -\boldsymbol{b}_i + \boldsymbol{r}_i & i = 1, \ldots, K \\ -\boldsymbol{b}_i - \boldsymbol{r}_i & i = K+1, \ldots, 2K \end{cases}$$
$$\boldsymbol{b}_i = \sigma_{w,i}^{-1}[\ \Re(s_i^*\boldsymbol{h}_i^T), -\Im(s_i^*\boldsymbol{h}_i^T)\ ]^T,$$
$$\boldsymbol{r}_i = \sigma_{w,i}^{-1}\cot(\pi/M)[\ \Im(s_i^*\boldsymbol{h}_i^T), \Re(s_i^*\boldsymbol{h}_i^T)\ ]^T.$$

It is worthwhile to note that problem (35) is convex.

[1] As a technically subtle note, the SEP result for the case of $z = c\cdot s$, where $c > 0$, is available in the classical communications literature. However, the same result for arbitrary z does not seem to be as readily available.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

Our second challenge is to find a suitable algorithm for computing the optimal solution to problem (35); note that we consider large N. Since problem (35) can be formulated as a linear program, one could use general-purpose conic optimization software to complete the task. However, we argue that this is not preferred for large N. Here we give two solutions; both exploit the problem structure. One is to apply the smoothed accelerated projected gradient (APG) method, previously developed for the non-convex one-bit precoding problem [18]. Concisely, the method works as follows. We first approximate the non-differentiable f by

$$\hat{f}(\boldsymbol{x}) = \mu \log \left(\sum_{i=1}^{2K} e^{\boldsymbol{c}_i^T \boldsymbol{x}/\mu} \right),$$

where $\mu > 0$; note that $\hat{f}$ is smooth and it is a tight approximation of f when $\mu \to 0$. Then, we apply the APG method [38]–[40] on the smoothed problem. This gives rise to the following algorithm

$$\boldsymbol{x}^{k+1} = \left[\boldsymbol{x}_{\mathsf{ex}}^k - \beta^k \nabla \hat{f}(\boldsymbol{x}_{\mathsf{ex}}^k) \right]_{-\mathbf{1}}^{\mathbf{1}}, \quad k = 0, 1, 2, \ldots \tag{36}$$

where $\beta^k > 0$ is the step size; $\nabla \hat{f}(\boldsymbol{x})$ is the gradient of $\hat{f}$ at $\boldsymbol{x}$; $\boldsymbol{x}_{\mathsf{ex}}^k$, called an extrapolated point, is given by

$$\boldsymbol{x}_{\mathsf{ex}}^k = \boldsymbol{x}^k + \gamma_k (\boldsymbol{x}^k - \boldsymbol{x}^{k-1}),$$

with $\gamma_k = (\xi_{k-1} - 1)/\xi_k$, $\xi_i = (1 + \sqrt{1 + 4\xi_{k-1}^2})/2$ and $\xi_{-1} = 0$; the notation $[\cdot]_{-\mathbf{1}}^{\mathbf{1}}$ denotes the projection onto $[-1, 1]^n$. Note that $[\cdot]_{-\mathbf{1}}^{\mathbf{1}}$ is merely an element-wise clipping function; i.e., if $\boldsymbol{y} = [\boldsymbol{x}]_{-\mathbf{1}}^{\mathbf{1}}$ then $y_i = \max\{-1, \min\{x, 1\}\}$ for all i. We choose β^k as the reciprocal of the Lipschitz constant of $\nabla \hat{f}$, a choice that guarantees convergence to an optimal solution. The Lipschitz constant of $\nabla \hat{f}$ can be shown to be $\|\boldsymbol{C}\|_2^2/\mu$ [40], where $\boldsymbol{C} = [\ \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{2K}\ ]$, and $\|\cdot\|_2$ denotes the spectral matrix norm. We will call the algorithm in (36) the primal APG method.

The second solution considers a dual form of problem (35). The primal APG method has $2N$ decision variables, which is large, and the motivation of the dual method is to see if we can use a smaller number of variables to solve problem (35). More accurately, consider a regularized form of problem (35)

$$\min_{-\mathbf{1} \le \boldsymbol{x} \le \mathbf{1}} f(\boldsymbol{x}) + \frac{\tau}{2} \|\boldsymbol{x}\|_2^2 \tag{37}$$

for some small $\tau > 0$. As a key observation, we note that

$$f(\boldsymbol{x}) = \max_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\lambda}^T \mathbf{1} = 1} \boldsymbol{\lambda}^T \boldsymbol{C}^T \boldsymbol{x}.$$

The above alternative expression of f leads us to

$$(37) = \min_{-\mathbf{1} \le \boldsymbol{x} \le \mathbf{1}} \max_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\lambda}^T \mathbf{1} = 1} \boldsymbol{\lambda}^T \boldsymbol{C}^T \boldsymbol{x} + \frac{\tau}{2} \|\boldsymbol{x}\|_2^2 \tag{38a}$$

$$= \max_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\lambda}^T \mathbf{1} = 1} \min_{-\mathbf{1} \le \boldsymbol{x} \le \mathbf{1}} \boldsymbol{\lambda}^T \boldsymbol{C}^T \boldsymbol{x} + \frac{\tau}{2} \|\boldsymbol{x}\|_2^2 \tag{38b}$$

$$= \max_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\lambda}^T \mathbf{1} = 1} g(\boldsymbol{\lambda}) \triangleq \sum_{i=1}^{2N} -\varphi_\tau(\bar{\boldsymbol{c}}_i^T \boldsymbol{\lambda}), \tag{38c}$$

where $\bar{\boldsymbol{c}}_i$ denotes the ith row of $\boldsymbol{C}$; φ_τ is the Huber function and is given by

$$\varphi_\tau(y) = \begin{cases} y^2/(2\tau) & |y| \le \tau \\ |y| - \tau/2 & \text{otherwise} \end{cases}$$

Note that (38b) is due to Sion's minimax theorem [41], and (38c) is due to $\min_{-1 \le x \le 1}\ yx + \tau x^2/2 = -\varphi_\tau(y)$ which one can easily show. Consider the dual problem in (38c), which has $2K$ decision variables. In the same vein as the previously introduced APG method, we use APG to solve problem (38c)

$$\boldsymbol{\lambda}^{k+1} = \Pi_{\{\boldsymbol{\lambda} \ge \mathbf{0} | \boldsymbol{\lambda}^T \mathbf{1} = 1\}} \left(\boldsymbol{\lambda}_{\mathsf{ex}}^k + \beta^k \nabla g(\boldsymbol{\lambda}_{\mathsf{ex}}^k) \right), \tag{39}$$

for $k = 0, 1, 2, \ldots$, where $\Pi_{\{\boldsymbol{\lambda} \ge \mathbf{0} | \boldsymbol{\lambda}^T \mathbf{1} = 1\}}$ denotes the projection onto the unit simplex; $\boldsymbol{\lambda}_{\mathsf{ex}}^k$ is defined in the same way as $\boldsymbol{x}_{\mathsf{ex}}^k$; β^k is the step size. Note that there exist very efficient algorithms for computing the unit simplex projection [42]. Also, the Lipschitz constant of ∇g is shown to be $\|\boldsymbol{C}\|_2^2/\tau$.

Once we compute the optimal solution $\boldsymbol{\lambda}^\star$ to problem (38c), the question that remains is how we can use $\boldsymbol{\lambda}^\star$ to recover the optimal solution $\boldsymbol{x}^\star$ to problem (37). From the study of minimax theory [43], it is understood that $\boldsymbol{x}^\star$ must be an optimal solution to

$$\min_{-\mathbf{1} \le \boldsymbol{x} \le \mathbf{1}} (\boldsymbol{\lambda}^\star)^T \boldsymbol{C}^T \boldsymbol{x} + \frac{\tau}{2} \|\boldsymbol{x}\|_2^2. \tag{40}$$

Since problem (40) has one optimal solution only, owing to the strong convexity of its objective function, the optimal solution to problem (40) must be $\boldsymbol{x}^\star$ itself. The optimal solution to (40) is simply

$$\boldsymbol{x}^\star = \left[-\tfrac{1}{\tau} \boldsymbol{C} \boldsymbol{\lambda}^\star \right]_{-\mathbf{1}}^{\mathbf{1}}. \tag{41}$$

We will call the method in (39) and (41) the dual APG method.

From our numerical experience, the primal and dual APG methods are both competitive. This will be shown in the simulation results section.

C. The QAM Case

Having studied the $\Sigma\Delta$ ZF and SLP schemes for the M-ary PSK constellation case in the preceding subsections, we now consider the M-ary QAM constellation case. There is an aspect we need to discuss first. Previously we ignore the time index t in the basic signal model (1). This is without loss of generality since PSK symbols do not require signal amplitude information for detection, and it is unnecessary to coordinate the scalings of the received signals over time. But this is no longer true in the QAM case. More technically, in (2), we need to make the received signal scaling coefficients $c_{i,t}$'s to be consistent for every symbol, namely, by having $c_{i,1} = \cdots = c_{i,T} \triangleq c_i$ for every i [15], [17], [18], [29].

Let us consider the ZF scheme under the above design consideration. We modify the ZF scheme in Section V-A as

$$\bar{\boldsymbol{x}}_t = \frac{\boldsymbol{A}^\dagger \boldsymbol{D} \boldsymbol{s}_t}{\max_{t=1,\ldots,T} \|\boldsymbol{A}^\dagger \boldsymbol{D} \boldsymbol{s}_t\|_{IQ-\infty}}, \quad t = 1, \ldots, T. \tag{42}$$

Following the same development as before, one can show that the corresponding received signals are given by

$$y_{i,t} = c_i \cdot s_{i,t} + w_{i,t}, \quad c_i = \sqrt{\frac{P}{2N}} \gamma \sigma_{w,i}, \tag{43}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

where $\gamma = 1/\max_{t=1,\ldots,T} \|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}_t\|_{IQ-\infty}$. It can be shown that the same result in Proposition 1 applies here.

For the SLP scheme, essentially the same problem was considered in [18]. The latter considers the minimax SEP design in non-convex one-bit precoding, and does so by joint optimization of all $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_T$ and all scalings $c_1, \ldots, c_K$. This amounts to a large-scale problem, but enhanced SEP was also observed. The algorithm proposed there is similar to the primal APG method in Section V-B, with a non-convex penalty term for forcing a binary solution. By removing that penalty term, the algorithm will be applicable to our $\Sigma\Delta$ SLP design. We omit the details due to space limitation.

We propose one more scheme that strikes a balance between ZF and SLP. Consider the following nullspace-assisted ZF scheme

$$\bar{\boldsymbol{x}}_t = \frac{\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}_t + \boldsymbol{\eta}_t}{\max_{t=1,\ldots,T} \|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}_t + \boldsymbol{\eta}_t\|_{IQ-\infty}}, \ t = 1, \ldots, T, \quad (44)$$

where every $\boldsymbol{\eta}_t$ lies in the nullspace of $\boldsymbol{A}$. The scheme (44) is a more general version of the ZF scheme (42), taking advantage of the design simplicity of the latter. It is also a special case of the SLP scheme. From the alternative SLP interpretation (31)–(32), one can see that (44) is an SLP scheme that drops the symbol perturbation terms $\boldsymbol{u}_t$'s, adopts the simple way to decide the received signal gains β_i's in ZF, but keeps the nullspace term $\boldsymbol{\eta}_t$. The received signals of the scheme (44) are the same as (43), with γ replaced by $\gamma = 1/\max_{t=1,\ldots,T} \|\boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}_t + \boldsymbol{\eta}_t\|_{IQ-\infty}$. Now, the problem is find $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T$ such that γ is maximized. It is readily seen that we can achieve this by solving, in a time decoupled manner,

$$\min_{\boldsymbol{\xi}_t \in \mathbb{C}^{N-K}} \|\boldsymbol{r}_t + \boldsymbol{B}\boldsymbol{\xi}_t\|_{IQ-\infty}, \quad t = 1, \ldots, T, \quad (45)$$

where we apply change of variable $\boldsymbol{\eta}_t = \boldsymbol{B}\boldsymbol{\xi}_t$; $\boldsymbol{B} \in \mathbb{C}^{N\times(N-K)}$ is an orthogonal basis of the nullspace of $\boldsymbol{A}$; $\boldsymbol{r}_t = \boldsymbol{A}^\dagger \boldsymbol{D}\boldsymbol{s}_t$. We will show by simulation results that this nullspace-assisted ZF scheme provides order-of-magnitude SEP improvement over the ZF scheme.

We finish by mentioning how we solve the problems in (45). We first reformulate each problem in (45) in a form similar to problem (35), but without the constraint $\boldsymbol{x} \in [-1, 1]^{2N}$. Then we apply the smoothed APG method in (36) (without projection) to find the solution. We omit the details for the sake of brevity.

D. The Multi-Path Case

Our preceding developments can also be extended to the case of multi-path angular channels. Consider the multi-path channel model

$$\boldsymbol{h}_i = \sum_{l=1}^{L_i} \alpha_{il} \boldsymbol{a}(\theta_{il}), \quad (46)$$

where α_{il} and θ_{il} correspond to the complex channel gain and angle of the lth path to the ith user, respectively; L_i is the number of paths associated with the ith user. Following the same development as in the preceding sections, it can be shown that the basic signal model takes the same form as (24), i.e.,

$$y_i = \sqrt{\frac{P}{2N}} \boldsymbol{h}_i^T \bar{\boldsymbol{x}} + w_i, \quad i = 1, \ldots, K.$$

The difference is that the expression of the noise variance $\sigma_{w,i}^2$ is replaced by

$$\sigma_{w,i}^2 \approx \frac{P}{3N} \left(\sum_{n=0}^{N-1} \left| \sum_{l=1}^{L_i} \alpha_{il} (z_{il}^{-n} - z_{il}^{-n-1}) \right|^2 \right) + \sigma_v^2, \quad (47)$$

where $z_{il} = e^{\mathrm{j}\frac{2\pi d}{\lambda}\sin(\theta_{il})}$. As a result, the ZF and SLP schemes developed above can be applied to the multi-path case (with minor modifications). The detailed derivations of (47) are relegated to Appendix B.

We should mention that $\sigma_{w,i}^2$ does not increase with N. Using $|x_1 + \cdots + x_L|^2 \le L(|x_1|^2 + \cdots + |x_L|^2)$, we show from (47) that

$$\sigma_{w,i}^2 \le \frac{4PL_i}{3} \sum_{l=1}^{L_i} |\alpha_{il}|^2 \left| \sin\left(\frac{\pi d}{\lambda} \sin(\theta_{il}) \right) \right|^2 + \sigma_v^2.$$

As can be seen, the above bound does not depend on N.

VI. Simulation Results

This section shows our simulation results for $\Sigma\Delta$ precoding.

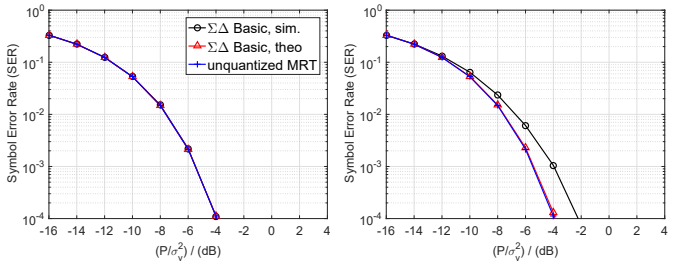
A. Single-User Case with Basic $\Sigma\Delta$ Modulation

We start with the single-user case, specifically, the basic $\Sigma\Delta$ MRT scheme in Section IV-B. The simulation settings are as follows. The number of antennas and the inter-antenna spacing are $N = 256$ and $d = \lambda/8$, respectively. The complex channel gain α has unit amplitude and phase uniformly drawn from $[-\pi, \pi]$ in each simulation trial. The symbol constellation is 8-ary PSK. For benchmarking we also evaluate the theoretical SEP bound of the basic $\Sigma\Delta$ MRT scheme, i.e., (13)–(14), and the simulated symbol error rate (SER) performance of the unquantized MRT scheme. Here, unquantized MRT (or any precoding) refers to the case where one applies MRT (or any precoding) without the one-bit signal restriction.

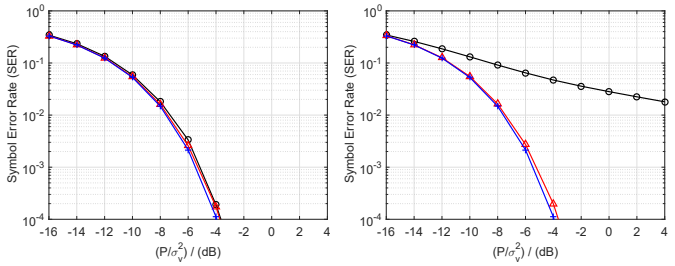
Fig. 3 shows the SER performance under several different values of the user angle θ. We see that, for the cases of $\theta = 0^\circ$ and $\theta = 60^\circ$, the simulated SER performance of the basic $\Sigma\Delta$ MRT scheme is almost the same as the theoretical. For the case of $\theta = 30^\circ$, we observe a small gap between the simulated and theoretical SER performance of the basic $\Sigma\Delta$ MRT scheme. The reason, as we found out, is that the quantization noise could have its behavior deviating from the i.i.d. assumption in some specific cases, and $\theta = 30^\circ$ happens to fall into one such case. We may mitigate the non-i.i.d. effect by dithering, although it may not be worthwhile to try dithering in this case since the performance gap is small and dithering increases the quantization noise level. Moreover, for the case of $\theta = 90^\circ$, we notice that the simulated and theoretical SER performance has a significant gap. Again, this is because the quantization noise is not i.i.d., and the non-i.i.d. effect is severe in this case.

Next, we show the radiation patterns of the basic $\Sigma\Delta$ MRT scheme. The simulation settings are the same as the previous.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

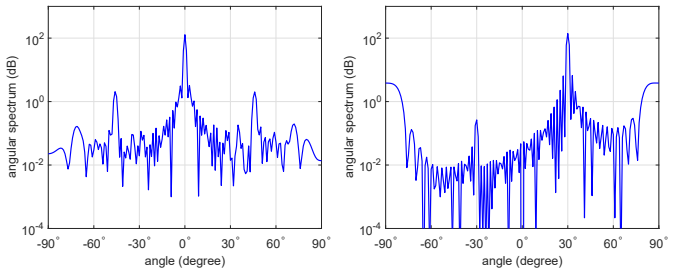


(a) $\theta = 0°$ (b) $\theta = 30°$

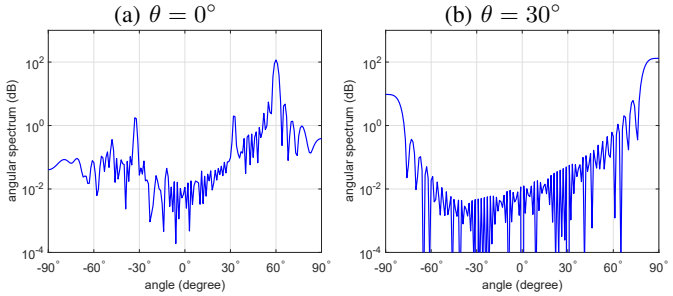


(c) $\theta = 60°$ (d) $\theta = 90°$

Fig. 3: SERs for different θ.



(a) $\theta = 0°$ (b) $\theta = 30°$



(c) $\theta = 60°$ (d) $\theta = 90°$

Fig. 4: The angular power spectrum for different θ.

Fig. 4 plots the angular power spectrum $P(\vartheta) = \mathbb{E}[|\boldsymbol{a}(\vartheta)^T \bar{\boldsymbol{x}}|^2]$ for several values of the desired angle θ. We see that the actual angular power spectrum does not always look like what theory ideally suggests, i.e., superposition of the highpass quantization noise spectrum and the MRT signal spectrum, the latter of which appears as a spike at θ. We see a highpass response with the actual angular power spectrum for $\theta = 30°, 60°, 90°$, but this is not seen for $\theta = 0$. We expect a single peak at the desired angle θ, but we also see some smaller peaks at other angles for $\theta = 0°, 30°, 60°$. Also, we do not see a peak for $\theta = 90°$. The non-ideal phenomena we see are due to the non-i.i.d. effects, and they are identical to those in the temporal $\Sigma\Delta$ modulation of DC and sinusoidal input signals [32], [33]. Nonetheless we can also argue that the actual angular spectrum roughly follows the theoretical, say, for $\theta = 30°, 60°$. As an aside, while our interest is to apply precoding to a target user, the quantization noise of the $\Sigma\Delta$ scheme also causes interference to other angles—an issue one needs to be careful when operating under multi-cell interfering channel environments.

Finally, we examine the SER performance under different numbers of antennas N. The user angle is fixed at $\theta = 60°$. The result in Fig. 5 illustrates that, under the same SNR level P/σ_v^2, increasing N reduces the SERs substantially. This numerical observation is in agreement with the SEP analysis result in Section IV-B.

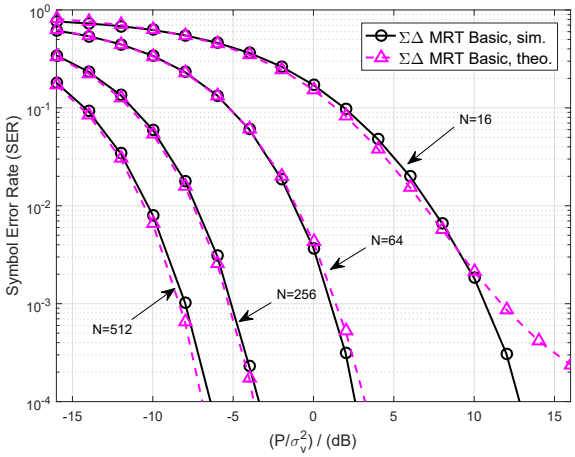


Fig. 5: SERs for different N.

B. Single-User Case with Angle-Steered $\Sigma\Delta$ Modulation

We turn our attention to the angle-steered $\Sigma\Delta$ MRT scheme in Section IV-C. The simulation settings are essentially identical to those in the preceding subsection, and the difference is that we reduce the number of antennas to $N = 128$, increase the inter-antenna spacing to $d = \lambda/2$, and try a large angle of $\theta = 90°$. The basic $\Sigma\Delta$ MRT scheme is expected to work poorly, as suggested by the analysis in Section IV-B. Also, as we have seen in the simulation results in the preceding subsection, the non-i.i.d. quantization noise effect may become significant. To mitigate the non-i.i.d. effect, we try the dithered $\Sigma\Delta$ MRT scheme, specifically, by applying the dithering procedure (7) to the basic spatial $\Sigma\Delta$ modulator. The dithering level δ in (7) is set to $\delta = 0.8$.

Fig. 6 shows the SER performance of the basic, dithered and angle-steered $\Sigma\Delta$ MRT schemes. As seen previously in Fig. 3(d), the basic $\Sigma\Delta$ MRT scheme suffers from the non-i.i.d. quantization noise effect when $\theta = 90°$. The situation now is even worse. We see from Fig. 6 that the basic $\Sigma\Delta$ MRT scheme completely fails, and does not perform as the theoretical SER performance says. The dithered $\Sigma\Delta$ MRT scheme yields significantly improved performance, and this indicates that dithering can reduce the non-i.i.d. effect. However, it is the angle-steered $\Sigma\Delta$ MRT scheme that gives the best performance. Also, the theoretical SER performance of the angle-steered $\Sigma\Delta$ MRT scheme accurately predicts the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

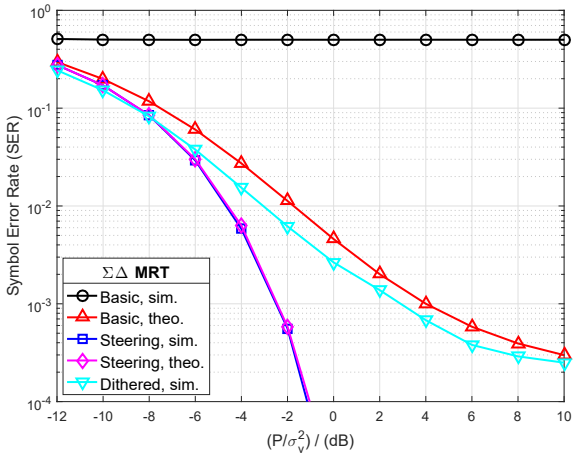


Fig. 6: SERs under angle-steering and dithering; $\theta = 90°$.

simulated SER performance.

Next, we consider the generalized angle-steered ΣΔ MRT scheme in Section IV-D under i.i.d. Gaussian channels. Specifically, in each simulation trial, the channel $\boldsymbol{h}$ is i.i.d. complex circular Gaussian generated with mean zero and unit variance. The number of antennas is $N = 256$, and the symbol constellation is 16-ary QAM. The benchmark scheme is the unquantized MRT. The unquantized MRT scheme we consider is the one under the peak IQ amplitude constraint $\|\boldsymbol{x}\|_{IQ-\infty} \leq 1$ and without one-bit quantization; precisely, it is implemented by (23) with $A_n = 1$ and with $\sigma_w^2 = \sigma_v^2$. Also, we try the direct one-bit quantization of the unquantized MRT scheme, which we call it the quantized MRT scheme; we will use the same convention to name other direct one-bit quantized algorithms in the sequel. In addition to the generalized angle-steered ΣΔ MRT scheme, we try a heuristic where we overload the generalized angle-steered ΣΔ MRT scheme by setting $A_n = 1$ for all n. Careful readers will see from Section IV-D that the issue will be that the surviving quantization noise term q_N in (21b) may become large. But if it does not in general, then the overloaded heuristic will have the advantage of enhanced SNR.

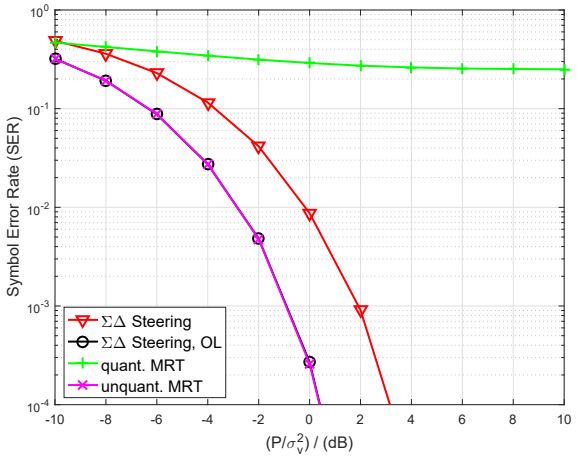


Fig. 7: SERs for the i.i.d. Gaussian channel.

Fig. 7 shows the results. We have the following observations. First, the quantized MRT scheme fails to work. Second, the generalized angle-steered ΣΔ MRT scheme ("ΣΔ Steering" in the figure) yields SER performance that is about 3dB away from that of the unquantized MRT scheme. This agrees with our analysis, which suggests 4.64dB as the worst case. Third, the overloaded generalized angle-steered ΣΔ MRT scheme ("ΣΔ Steering, OL") yields SER performance almost the same as that by the unquantized MRT. While our present work only considers the no-overload case, this simulation result suggests that overloading can be beneficial. We will leave overloading as a subject of future investigation.

C. The Multi-User Case

Now we consider the multi-user case. The simulation settings are as follows: The number of antennas is $N = 512$; the inter-antenna spacing is $d = \lambda/8$; the number of users is $K = 24$, and the users are within an angular range $[-30°, 30°]$; the angles θ_i's are randomly picked from $[-30°, 30°]$ with inter-angle difference no greater than $1°$; the complex channel gains α_i's have phases uniformly drawn from $[-\pi, \pi]$, and their amplitudes are generated as $|\alpha_i| = r_0/r_i$ where $r_0 = 30$ and r_i are uniformly drawn from $[20, 100]$ (this is a standard free-space path-loss model, with r_i being the distance from the BS to the ith user and r_0 being a reference value); the symbol constellation is 8-ary PSK.

The settings of the ΣΔ SLP scheme should also be mentioned. For the primal APG, the smoothing parameter $\mu = 0.05$, and the algorithm stops when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 \leq 10^{-5}$ or when a maximum iteration number of 2000 is reached. For the dual APG, the regularization parameter is $\tau = 0.005$, and the algorithm stops when $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|_2 \leq 10^{-7}$ or when a maximum iteration number of 3000 is reached.

Fig. 8 shows the results. In the legend, "unquant. ZF" is the unquantized ZF scheme under the average power constraint; "quant. ZF" is the direct one-bit quantization of the unquantized ZF scheme; "ΣΔ ZF" is the ΣΔ ZF scheme in Section V-A; "ΣΔ Primal APG" and "ΣΔ Dual APG" are the primal and dual ΣΔ SLP schemes in Section V-B; "unquant. SLP" is the unquantized version of the SLP scheme; "quant. SLP" is the direct one-bit quantization of the unquantized SLP scheme. We see that the proposed ΣΔ ZF and SLP schemes work well. The quantized ZF and SLP schemes do not, however.

Next, we perform benchmarking with some existing one-bit precoding designs. The simulation settings are the same as the previous, except that we reduce the number of antennas to $N = 256$ and the angular range to $[-22.5°, 22.5°]$. The compared algorithms are the SQUID algorithm [12] and the maximum safety margin (MSM) algorithm [30]. The results in Fig. 9 show that the ΣΔ SLP scheme outperforms SQUID and MSM, and the ΣΔ ZF scheme performs better than the latter when the SNR is greater than 25dB. In addition to SERs, we also compare the algorithm runtimes. Table I shows the runtime results; the results were obtained on MATLAB, and a desktop computer with Intel i7-4770 processor and 16GB memory was used to perform the runtime test. We can see that the proposed ΣΔ SLP designs yield competitive runtime performance compared to SQUID and MSM.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

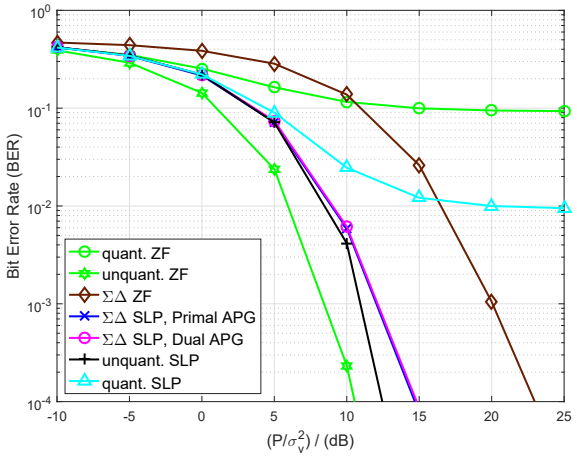


Fig. 8: Bit error rates (BERs) of the multi-user ΣΔ precoding schemes in the 8-ary PSK case.

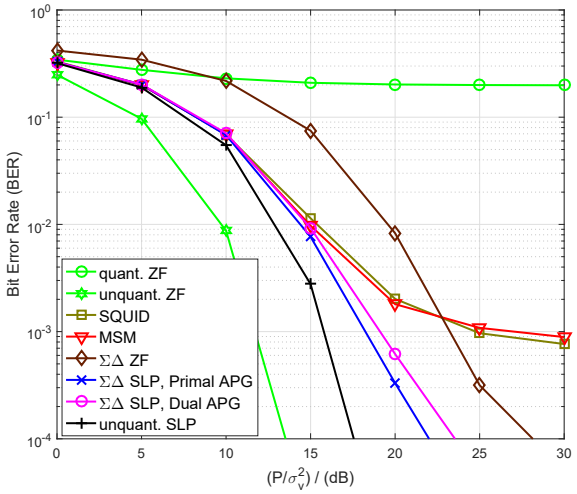


Fig. 9: BER comparison of the multi-user ΣΔ precoding schemes and some existing one-bit precoding schemes.

TABLE I: Average runtime (in Sec.) of different algorithms; $(N, K) = (256, 24)$, 8-ary PSK.

Algorithm	Σ-Δ ZF	Σ-Δ Primal APG	Σ-Δ Dual APG	SQUID	MSM
runtime	**0.0021**	**0.0574**	**0.0496**	1.0324	0.9197

Finally, we consider the QAM case. The simulation settings are: 16-ary QAM, $N = 256$, $K = 16$, and the transmission block length $T = 100$. Also, the angular range is $[-30°, 30°]$, and the α_i's are generated in the same way as before. Fig. 10 shows the results. In the plot, "ΣΔ ZF" is the ΣΔ ZF scheme in (42); "ΣΔ null. ZF" is the nullspace-assisted ΣΔ ZF scheme in (44)–(45), and "GEMM" is the direct one-bit precoding design in [18]. We see that the ΣΔ ZF schemes, with and without nullspace assistance, work. Also we should pay particular attention to the nullspace-assisted ΣΔ ZF scheme. It has a 5dB gain compared to the ΣΔ ZF scheme, and it is only 3dB away from GEMM. We should mention that GEMM handles a more complicated design problem than the nullspace-assisted ΣΔ ZF scheme.

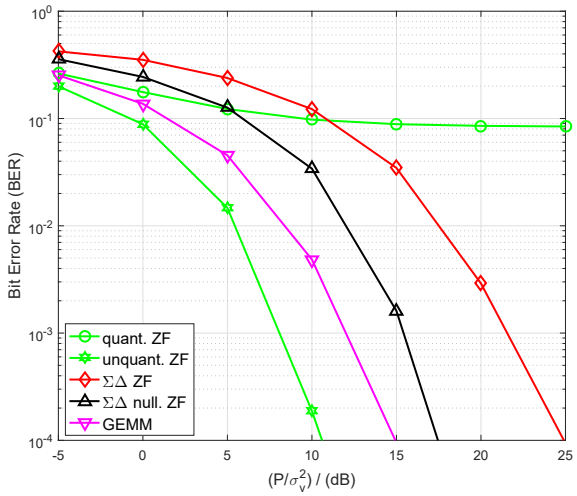


Fig. 10: BERs of the multi-user ΣΔ precoding schemes in the 16-ary QAM case.

VII. Conclusion

In this paper we studied the potential of spatial ΣΔ modulation for one-bit MIMO precoding. We showed that ΣΔ precoding is an excellent candidate when the system is equipped with a massive antenna array and when the users lie within a certain angular sector, which is a typical assumption in many cellular systems. The major advantage of ΣΔ precoding is that it can achieve good performance for relatively simple designs such as quantized linear precoding, whereas direct one-bit design requires complicated non-convex methods with binary signal constraints (or relaxed versions thereof) in order to obtain low error rates. While our initial ΣΔ precoder assumed a simple angular channel, we showed how to generalize the idea to any type of channel in the single-user case.

Appendix

A. *Proof of* (30)

The proof of (30) in Proposition 1 is as follows. Since $\boldsymbol{R}$ is Hermitian, we can write

$$\lambda_{\min}(\boldsymbol{R}) = \min_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^H \boldsymbol{R}\boldsymbol{x}.$$

Let $\phi_i = 2\pi d \sin(\theta_i)/\lambda$, and let $z_i = e^{\mathrm{j}\phi}$. Since $\boldsymbol{a}_i = [\,1, z_i^{-1}, \ldots, z_i^{-(N-1)}\,]^T$, we have

$$r_{ij} = \frac{1}{N}\boldsymbol{a}_i^T \boldsymbol{a}_j^* = \frac{1}{N}\sum_{n=0}^{N-1}(z_i^{-1}z_j)^n = \frac{1-(z_i^{-1}z_j)^N}{N(1-(z_i^{-1}z_j))}.$$

Thus, the elements of $\boldsymbol{R}$ satisfy $r_{ii} = 1$ and

$$|r_{ij}| = \left| D_N\left(\frac{\phi_1 - \phi_2}{2}\right)\right| \le \rho.$$

Since $r_{ii} = 1$, we have

$$\lambda_{\min}(\boldsymbol{R}) \le \boldsymbol{e}_1^H \boldsymbol{R}\boldsymbol{e}_1 = 1,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2938687, IEEE Journal of Selected Topics in Signal Processing

where $\boldsymbol{e}_1 = [\ 0, 1, \ldots, 1]^T$. Also, it holds that

$$\begin{aligned}\boldsymbol{x}^H \boldsymbol{R}\boldsymbol{x} &\geq \sum_{i=1}^{K} |x_i|^2 - \sum_{i\neq j} |r_{ij}||x_i||x_j| \\ &\geq \sum_{i=1}^{K} |x_i|^2 - \rho \sum_{i\neq j} |x_i||x_j| \\ &= |\boldsymbol{x}|^T((1+\rho)\boldsymbol{I} - \rho \mathbf{1}\mathbf{1}^T)|\boldsymbol{x}| \\ &\geq (1 + (1-K)\rho)\|\boldsymbol{x}\|_2^2,\end{aligned}$$

where we denote $|\boldsymbol{x}| = [\ |x_1|, \ldots, |x_n|\ ]^T$ in the third equation; the last equation is due to $\lambda_{\min}((1+\rho)\boldsymbol{I} - \rho\mathbf{1}\mathbf{1}^T) = 1 + \rho - \rho\|\mathbf{1}\|_2^2 = 1 + (1-K)\rho$. It therefore follows that $\lambda_{\min}(\boldsymbol{R}) \geq 1 + (1-K)\rho$. The proof of (30) is thus complete.

B. Derivation of (47)

Following the single-user development in Section IV-A, the noise term w_i in the model (24) of the multiuser case is given by

$$w_i = \sqrt{\frac{P}{2N}}\boldsymbol{h}^T(\boldsymbol{q} - \boldsymbol{q}^-) + v_i;$$

recall $\boldsymbol{q}^- = [\ 0, q_1, \ldots, q_{N-1}\ ]^T$. When the channel $\boldsymbol{h}_i$ takes the multipath form in (46), w_i can be expressed as

$$\begin{aligned}w_i = \sqrt{\frac{P}{2N}}\Bigg[&\sum_{n=0}^{N-2}\left(\sum_{l=1}^{L_i}\alpha_{il}(z_{il}^{-n} - z_{il}^{-n-1})\right) q_{n+1} \\ &+ \left(\sum_{l=1}^{L_i}\alpha_{il} z_{il}^{-N-1}\right) q_N\Bigg] + v_i,\end{aligned}$$

where $z_{il} = e^{\mathrm{j}\frac{2\pi d}{\lambda}\sin(\theta_{il})}$. Under the uniform i.i.d. assumption with $\boldsymbol{q}$, we have $\mathbb{E}[w_i] = 0$ and

$$\begin{aligned}\sigma_{w,i}^2 = \frac{P\sigma_q^2}{2N}\Bigg[&\sum_{n=0}^{N-2}\left|\sum_{l=1}^{L_i}\alpha_{il}(z_{il}^{-n} - z_{il}^{-n-1})\right|^2 \\ &+ \left|\sum_{l=1}^{L_i}\alpha_{il} z_{il}^{-N-1}\right|^2\Bigg] + \sigma_v^2,\end{aligned}$$

where $\sigma_q^2 = \mathbb{E}[|q_n|^2] = 2/3$. By approximating the above expression as

$$\sigma_{w,i}^2 \approx \frac{P\sigma_q^2}{2N}\left[\sum_{n=0}^{N-1}\left|\sum_{l=1}^{L_i}\alpha_{il}(z_{il}^{-n} - z_{il}^{-n-1})\right|^2\right] + \sigma_v^2,$$

which is reasonable for large N, we arrive at the noise variance formula in (47).

REFERENCES

[1] L. Fan, S. Jin, C. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Dec 2015.

[2] 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.

[3] 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.

[4] 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.

[5] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, June 2017.

[6] Z. Shao, R. C. de Lamare, and L. T. N. Landau, "Iterative detection and decoding for large-scale multiple-antenna systems with 1-bit ADCs," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 476–479, June 2018.

[7] Y. Jeon, N. Lee, S. Hong, and R. W. Heath, "One-bit sphere decoding for uplink massive MIMO systems with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4509–4521, July 2018.

[8] 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.*, Dec 2009, pp. 683–686.

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

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

[11] O. Castaneda, 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.

[12] 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.

[13] 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. Acous., Speech, Signal Process. (ICASSP)*, Mar. 2017, pp. 6483–6487.

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

[15] 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.

[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, 2018.

[17] 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, 2018.

[18] 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," *arXiv preprint arXiv:1810.03159*, 2018.

[19] P. M. Aziz, H. V. Sorensen, and J. Van Der Spiegel, "An overview of sigma-delta converters: How a 1-bit ADC achieves more than 16-bit resolution," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 61–84, 1996.

[20] R. M. Corey and A. C. Singer, "Spatial sigma-delta signal acquisition for wideband beamforming arrays," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, March 2016.

[21] D. Barac and E. Lindqvist, "Spatial sigma-delta modulation in a massive MIMO cellular system," Master's thesis, Department of Computer Science and Engineering, Chalmers University of Technology, 2016.

[22] A. Nikoofard, J. Liang, M. Twieg, S. Handagala, A. Madanayake, L. Belostotski, and S. Mandal, "Low-complexity N-port ADCs using 2-D sigma-delta noise-shaping for N-element array receivers," in *Proc. Int. Midwest Symposium Circuits Syst. (MWSCAS)*, 2017, pp. 301–304.

[23] A. Madanayake, N. Akram, S. Mandal, J. Liang, and L. Belostotski, "Improving ADC figure-of-merit in wideband antenna array receivers using multidimensional space-time delta-sigma multiport circuits," in *Proc. Int. Workshop Multidimensional (nD) Syst. (nDS)*, Sept 2017.

[24] V. Venkateswaran and A. van der Veen, "Multichannel $\Sigma\Delta$ ADCs with integrated feedback beamformers to cancel interfering communication signals," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2211–2222, May 2011.

[25] D. S. Palguna, D. J. Love, T. A. Thomas, and A. Ghosh, "Millimeter wave receiver design using low precision quantization and parallel $\Delta\Sigma$ architecture," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6556–6569, Oct 2016.

[26] I. Galton and H. T. Jensen, "Delta-sigma modulator based A/D conversion without oversampling," *IEEE Trans. Circuits Syst. II, Analog Digital Signal Process.*, vol. 42, no. 12, pp. 773–784, Dec 1995.

[27] D. P. Scholnik, J. O. Coleman, D. Bowling, and M. Neel, "Spatio-temporal delta-sigma modulation for shared wideband transmit arrays," in *Proc. IEEE Radar Conf.*, 2004, pp. 85–90.

[28] J. D. Krieger, C. P. Yeang, and G. W. Wornell, "Dense delta-sigma phased arrays," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1825–1837, April 2013.

[29] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *Proc. Asilomar Conf. Signals, Syst. Comp.*, Nov 2016, pp. 763–767.

[30] H. Jedda, A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, "Massive MIMO downlink 1-bit precoding with linear programming for PSK signaling," in *2017 IEEE 18th Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, July 2017, pp. 1–5.

[31] A. Swindlehurst, H. Jedda, and I. Fijalkow, "Reduced dimension minimum BER PSK precoding for constrained transmit signals in massive MIMO," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Process. (ICASSP)*, 2018, pp. 3584–3588.

[32] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1220–1244, 1990.

[33] S. R. Norsworthy, "Effective dithering of sigma-delta modulators," in *Proc. 1992 IEEE Int. Symposium Circuits Syst. (ISCAS)*, vol. 3, 1992, pp. 1304–1307.

[34] J. Proakis, *Digital Communications*, ser. Electrical engineering series. McGraw-Hill, 2001. [Online]. Available: https://books.google.com.hk/books?id=sbr8QwAACAAJ

[35] Y. Liu and W.-K. Ma, "Symbol-level precoding is symbol-perturbed ZF when energy efficiency is sought," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Process. (ICASSP)*, 2018, pp. 3869–3873.

[36] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, 2008.

[37] M. Shao, Q. Li, Y. Liu, and W.-K. Ma, "Multiuser one-bit massive MIMO precoding under MPSK signaling," in *Proc. IEEE Global Conf. Signal and Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 833–837.

[38] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$," in *Doklady AN USSR*, vol. 269, 1983, pp. 543–547.

[39] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[40] A. Beck, *First-Order Methods in Optimization*. SIAM, 2017, vol. 25.

[41] M. Sion, "On general minimax theorems," *Pacific J. Math.*, vol. 8, no. 1, pp. 171–176, 1958.

[42] L. Condat, "Fast projection onto the simplex and the l_1 ball," *Mathematical Programming*, vol. 158, no. 1-2, pp. 575–585, 2016.

[43] D. P. Bertsekas, A. Nedi, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.