Quantization with Combined Codebook for Hybrid Array Using Two-Phase-Shifter Structure

Yuyue Luo*†, J. Andrew Zhang†, Shaode Huang*, Jin Pan*, Xiaojing Huang†
*School of Electronic Science and Engineering, University of Electronic Science and Technology of China, China
†School of Electrical and Data Engineering, University of Technology Sydney, Australia
Email: Yuyue.Luo@student.uts.edu.au; {Andrew.Zhang; Xiaojing.Huang}@uts.edu.au;
shaodehuang@foxmail.com; panjin@uestc.edu.cn

***Abstract*—We propose a novel joint quantization scheme for hybrid antenna array systems using the two-phase-shifter (2-PS) structure, where two phase shifters are combined to represent one beamforming weight. Conventional quantization using a single phase shifter for each beamforming weight cannot represent the magnitude. We propose a new codebook design that combines the two codebooks of the two phase shifters in the recently proposed 2-PS structure. We also study the scaling problem of the beamforming vector and propose a low-complexity searching algorithm for finding a near-optimal scalar based on element-wise quantization. The mean squared quantization error and signal-to-noise ratio (SNR) degradation are derived analytically. Simulation results validate the accuracy of the analytical results and the effectiveness of the proposed quantization methods.**

I. Introduction

Millimeter wave (mmWave) hybrid array systems are regarded as a promising solution in 5G [1], [2]. Hybrid arrays involve baseband digital and analog Radio Frequency (RF) precoding/combining. Analog phase shifters are often used in the RF domain, with discrete phase shifting values and constant modulus. Hence RF precoding/combining practically use quantized values for the beamforming (BF) vectors.

Various quantization methods have been studied for codebook design in MIMO systems [3]–[5] and mmWave hybrid array [1], [6]. Although digital precoder can mitigate the performance degradation due to the quantization error in the RF beamforming [6]–[8], the RF beamforming forms part of the equivalent channel and quantization error still has a notable impact on the overall system performance. The quantized phase shifts and lack of amplitude adjustment for precoders/combiners can cause significant degradation in array gain and output signal-to-noise ratio (SNR), as pointed in [9], [10].

Different to conventional phase shifter structures where only one phase shift is used to represent one beamforming weight, a two-phase-shifter (2-PS) structure was recently proposed and analyzed in [10], [11]. The 2-PS structure uses two phase shifters either serially or in parallel to represent one beamforming weight. Using the 2-PS structures, RF precoder/combiner can potentially represent any precoding coefficients with very small quantization error, when the number of quantization bits is sufficiently large. Basic performance analysis for this method can be found in [10], implicitly considering quantization for each phase shifter separately in the 2-PS structure. Such separated quantization can lead to large quantization error unless the number of quantization bits in each phase shifter is very large.

In this paper, we propose a novel joint quantization method for the 2-PS structure, which can achieve small quantization error with only more than 2 quantization bits for each phase shifter. This method uses a new codebook generated by combining the codebooks of the two phase shifters. We also introduce a fixed phase shifting value into one codebook which can double the number of different codes in the combined codebook. Based on this new codebook, we propose low-complexity element-wise quantization methods with refined scaling of the beamforming vector. We propose a simple one-dimensional searching algorithm for finding the optimal scaling factor based on the improved golden section search (IGSS) method [12]. We also characterize the performance of the proposed methods and compare them with the results in [10]. We analytically evaluate the mean squared quantization error (MSQE) as well as the precoder's SNR degradation due to quantization. These analytical results are shown to match the simulated results well.

Notations: $(\cdot)^H$, $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^{-1}$, and $(\cdot)^\dagger$ denote the Hermitian transpose, conjugate, transpose, inverse, and pseudo-inverse, respectively. $|\cdots|$, $\|\cdots\|$, and $\|\cdots\|_2^2$ denote the element-wise absolute value, the norm, and the Euclidean norm, respectively. $E(\cdot)$ denotes the expected value. $\mathbf{I}_N$ is the $N \times N$ identity matrix.

II. System Model and Preliminaries

In this section, we first present the precoding system and the channel model, and then introduce the 2-PS structures for quantizing RF precoding vectors.

A. System and Channel Model

We consider a narrow-band hybrid precoding system with N_t transmit and N_r receive antennas carrying N_s data streams. Using the uniform linear array (ULA) where antenna elements are spaced at an interval of d, the array steering vector is

$$\begin{aligned}\mathbf{a}_t(\theta_t) &= [1, e^{jk\sin(\theta_t)}, \cdots, e^{jk(N_t-1)\sin(\theta_t)}]^T\\ \mathbf{a}_r(\theta_r) &= [1, e^{jk\sin(\theta_r)}, \cdots, e^{jk(N_r-1)\sin(\theta_r)}]^T,\end{aligned} \tag{1}$$

where $k = 2\pi\lambda/d$ with λ being the wavelength, θ_t and θ_r are the angle of departing (AoD) and angle of arriving

(AoA), respectively. A typical channel model with L multipath signals is used in this paper. The quasi-static physical channels between the transmitting and receiving antennas can then be represented as

$$\mathbf{H} = r\sum_{\ell=1}^{L} b_\ell \delta(t-\tau_\ell)e^{j2\pi f_{D,\ell}t}\mathbf{a}_r(\theta_{r,\ell})\mathbf{a}_t^T(\theta_{t,\ell}), \tag{2}$$

where for the ℓ-th multipath, b_ℓ is its amplitude of complex value, τ_ℓ is the propagation delay, and $f_{D,\ell}$ is the associated Doppler frequency.

Let the $N_s \times 1$ transmit signal be $\mathbf{s}$ and $E[\mathbf{ss}^H] = \mathbf{I}_{N_s}$. Let $\mathbf{F}_{\text{RF}}^t$, $\mathbf{F}_{\text{RF}}^r$ be the RF precoder and combiner. In this paper, we will mainly focus on the quantization methods of $\mathbf{F}_{\text{RF}}$, hence a simplified model is considered where a single RF chain is used, i.e., $N_s = N_{\text{RF}} = 1$. In this case, the received signal can be represented as

$$\mathbf{y} = \sum_{\ell=1}^{L} b_\ell e^{j2\pi f_{D,\ell}t}({\mathbf{f}_{\text{RF}}^r}^H \mathbf{a}_r(\theta_{r,\ell}))(\mathbf{a}_t^H(\theta_{t,\ell})\mathbf{f}_{\text{RF}}^t)s(t-\tau_\ell) + z(t),$$

where $\mathbf{f}_{\text{RF}}^r$ and $\mathbf{f}_{\text{RF}}^t$ are the $N_r \times 1$ and $N_t \times 1$ beamforming vectors, and $z(t)$ is AWGN with mean 0 and variance σ_n^2. Let the singular value decomposition of the channel matrix be

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V^H},$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary, of sizes $N_r \times N_r$ and $N_t \times N_t$, respectively, $\mathbf{\Sigma}$ is a diagonal matrix whose elements are in non increasing order, i.e., $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{M-1}$, where $M = \min\{N_r, N_t\}$. Let $\mathbf{v}_0$ and $\mathbf{u}_0$ be the first column vector of $\mathbf{V}$ and $\mathbf{U}$, respectively. The optimal BF vectors for precoding and combining are $\mathbf{f}_t = \mathbf{v}_0$ and $\mathbf{f}_r = \mathbf{u}_0$, respectively. Therefore, the optimal SNR can be given as

$$\text{SNR}_0 = \frac{|{\mathbf{f}_{\text{RF}}^r}^H \mathbf{H}\mathbf{f}_{\text{RF}}^t|^2}{\|\mathbf{f}_{\text{RF}}^r\|^2\|\mathbf{f}_{\text{RF}}^t\|^2}\frac{\sigma_s^2}{\sigma_n^2} = \frac{\lambda_0^2\sigma_s^2}{\sigma_n^2}, \tag{3}$$

where σ_s^2 and σ_n^2 the power of signal and noise, respectively.

B. Quantization with the 2-PS Structure

For conventional precoders using a single phase shifter to represent each BF weight, only values with unit magnitudes can be represented. The magnitude mismatches can cause inaccurate beam steering and notable sidelobe growth [9] even using phase shifters with infinite number of quantization bits. This problem can be solved by using an active array which uses a power amplifier with each phase shifter. Alternatively, it is shown in [10] that two phase shifters can be used to represent each BF weight, achieving negligible performance loss when the quantization step is sufficiently large. There are two optional structures for realizing this, as shown in Fig. 1. A complex value f_i can be represented by

$$f_i = |f_i|e^{j\psi_i} = e^{j\beta_1^{(i)}} + e^{j\beta_2^{(i)}}, \tag{4a}$$

and

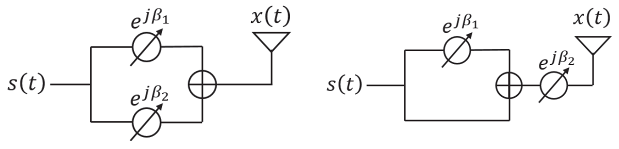
$$f_i = |f_i|e^{j\psi_i} = e^{j\beta_1^{(i)}}(1 + e^{j\beta_2^{(i)}}). \tag{4b}$$

for the parallel and serial structures, respectively. Thus, the ideal non-quantized phase values for the parallel and serial structures can be derived as

$$\beta_1^{(i)} = \psi_i + \arccos(|f_i|/2),\ \beta_2^{(i)} = \psi_i - \arccos(|f_i|/2), \tag{5a}$$

$$\beta_1^{(i)} = \psi_i - \arccos(\frac{|f_i|}{2}),\ \beta_2^{(i)} = 2\arccos(\frac{|f_i|}{2}), \tag{5b}$$

respectively.



(a) Option 1: parallel structure (b) Option 2: serial structure

Fig. 1. Optional parallel and serial structures with two phase shifters.

Let b be the number of quantization bits in each phase shifter. Assume that the discrete phase values are equally spaced over the interval of 2π with a quantization step of $\Delta = 2\pi 2^{-b}$. Hence each phase shifter has a codebook of 2^b codes. For the 2-PS options, [10] implemented a straightforward way for quantization - separately deciding the quantized precoding values through quantizing each phase shifters referring to (5). This method does not fully exploit the quantization potentials of the 2-PS structure, and can lead to large quantization error when the number of quantization bits is small. Next, we propose a novel codebook design method for the 2-PS scheme, employing quantization values jointly generated by the two phase shifters.

III. Joint Quantization Using Combined Codebook

In this section, we first introduce a combined codebook method for generating quantization values from the 2-PS structure, and then propose a quantization algorithm for quantizing the BF vector using this combined codebook.

A. Generation of Combined Codebook

We consider a pair of generalized codebooks containing the quantized phase values

$$\begin{aligned} \mathcal{B}_1 &= \{0, \Delta_{\beta_1}, 2\Delta_{\beta_1}, \ldots, (2^{b_1}-1)\Delta_{\beta_1}\}, \\ \mathcal{B}_2 &= \{\phi, \phi+\Delta_{\beta_2}, \ldots, \phi+(2^{b_2}-1)\Delta_{\beta_2}\}, \end{aligned} \tag{6}$$

where ϕ, $0 \leq \phi \leq \Delta_{\beta_2}/2$, is a constant for any fixed phase shifter. Such a constant phase shift can be realized easily by, e.g., using a fixed length of the delay line in the circuit for either structure in Fig. 2.

A combined codebook can be generated from these two codebooks. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be any quantized phase shifts in $\mathcal{B}_1$ and $\mathcal{B}_1$, respectively. A combined codebook $\mathcal{C}$ is generated by

$$\hat{c} = e^{j\hat{\beta}_1} + e^{j\hat{\beta}_2} \text{ or } \hat{c} = e^{j\hat{\beta}_1}(1 + e^{j\hat{\beta}_2}),\ c \in \mathcal{C}, \tag{7}$$

corresponding to Fig. 2(a) or 2(b). The codes in $\mathcal{C}$ do not have unit magnitude anymore. Note that the parallel and serial

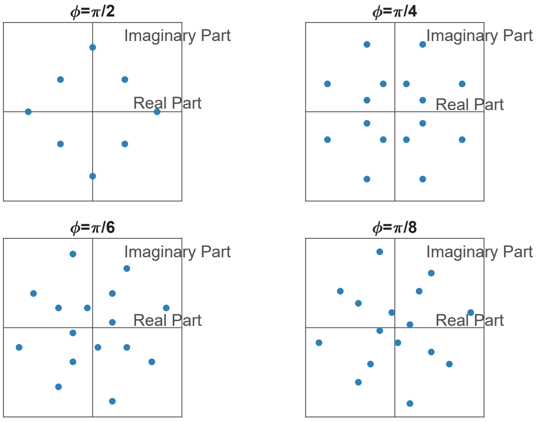
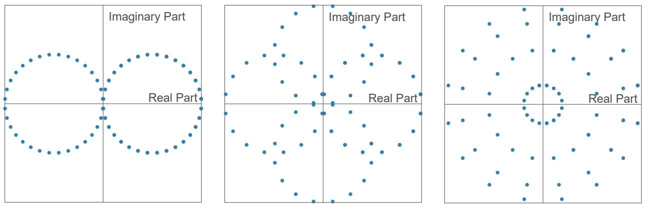


Fig. 2. Constellation of codebook $\mathcal{C}$ via ϕ when $b_1 = b_2 = 2$.



(a) $b_1 = 1$ and $b_2 = 5$. (b) $b_1 = 2$ and $b_2 = 4$. (c) $b_1 = 3$ and $b_2 = 3$.

Fig. 3. Constellation of codebook $\mathcal{C}$ via varying b_1 and b_2, when $\phi = \Delta_{\beta_2}/2$.

structures generate the same codebooks, and the two options are essentially equivalent when using the combined codebook.

This phase shift ϕ can influence the number of codes in $\mathcal{C}$ as well as the distribution of the constellation plot, as shown in Fig. 2. If $\phi = 0$, there will be 2^b repetitive values out of the total 2^{2b} codes in $\mathcal{C}$. Reduced number of distinct codes will lead to increased quantization errors. It can also be observed that when $\phi = \Delta_{\beta_2}/2$, the constellation is uniformly and symmetrically distributed over the complex plane, which can be a beneficial feature in generating a random precoding coefficient.

Besides, if the quantization bits of the two phase shifters are different, the constellations of the codebook $\mathcal{C}$ will be quite different. Fig. 3 plots the constellation for various options with $b_1 + b_2 = 6$ and $\phi = \pi/4$. It is discovered that when $b_1 \neq b_2$, the values of $\hat{c}$ tend to gather at some specific regions on the complex plane. In comparison, $\hat{c}$ has a more uniform distribution when $b_1 = b_2$. In this paper, we study two exemplified codebooks: $\mathcal{C}_1$ with $\phi = 0$ and $\mathcal{C}_2$ with $\phi = \Delta_{\beta_2}/2$, both having $b_1 = b_2 = b$. The number of different codes in $\mathcal{C}_1$ and $\mathcal{C}_2$ are $n_{c_1} = 2^{2b-1}$ and $n_{c_2} = 2^{2b}$, respectively.

Since the RF precoding vectors are often normalized to $\|\mathbf{f}_{\text{RF}}\| = 1$, the codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$ are normalized by

$$h_1 = \sqrt{\frac{N}{2^{b-1}} \sum_{i=1}^{2^{b-1}} \hat{c}_{k,i}^2} = \sqrt{2 + 2^{2-b}}\sqrt{N}, \tag{8}$$

and

$$h_2 = \sqrt{\frac{N}{2^b} \sum_{i=1}^{2^b} \hat{c}_{k,i}^2} = \sqrt{2N}, \tag{9}$$

respectively. So $E[|\hat{c}_{k,i}|^2] = 1/N$, where $\hat{c}_{k,i}$ is the i-th element in $\mathcal{C}_k$.

B. Quantization Algorithm

In this paper, we focus on studying element-wise quantization for the RF precoder vector $\mathbf{f}_{\text{RF}}$ for its simplicity and efficiency. As we will see from the simulation results in Section V, our element-wise quantization algorithm to be presented can already achieve sufficiently good performance at small number of quantization bits.

Although the codebook $\mathcal{C}_k$ is normalized to h_k, directly finding the quantization value from $\mathcal{C}_k$ for the element in $\mathbf{f}_{\text{RF}}$ may not be the best option considering the vector level quantization. This is because h_k are approximated values calculated from the statistical aspect and hence cannot guarantee the optimality for quantizing a particular RF precoder $\mathbf{f}_{\text{RF}}$. Therefore, with the goal of finding a better solution, we propose the improved golden section search-quantization (IGSS-Q) algorithm as described in Algorithm 1. It is based on the IGSS algorithm [12], which is an effective linear one-dimensional search method that relaxes the unimodal requirement for the classic golden-section search method. The IGSS-Q method solves the following problem

$$\nu_{\text{opt}} = \arg\min_{\nu} \|\nu \mathbf{f}_{\text{RF}} - \hat{\mathbf{q}}(\nu)\|_2^2 \tag{10}$$

recursively, where $\hat{\mathbf{q}}(\nu)$ is achieved by the scalar quantization with $\tilde{\mathcal{C}}_k$ and the i-th element of $\hat{\mathbf{q}}(\nu)$ is obtained by

$$\hat{q}_i = \arg\min_{\hat{c} \in \mathcal{C}_k} |\nu f_i - \hat{c}_{k,i}|^2, \tag{11}$$

where $i \in \{0, 1, \cdots, N_{\text{RF}}\}$. For a fixed ν and values of $\hat{q}_i$ obtained in (11), the quantization error function $e(\nu)$ can be expressed as

$$e(\nu) = \sum_{i=1}^{N} |\nu f_i - \hat{q}_i|^2. \tag{12}$$

The IGSS-Q method starts with seting the initial searching interval $\nu \in [a_1, a_2]$ and then define interior points x_1 and x_2 to divide the golden section in this interval. In each iteration, the IGSS-Q method finds the corresponding quantized values and compute the errors via (11) and (12) for $\nu = a_1, a_2, x_1$, and x_2. By comparing $e(a_1)$, $e(a_2)$ with $e(x_1)$ and $e(x_2)$, the searching interval is updated and narrowed gradually. Repeat this process until $e(x_1) - e(x_2)$ is smaller than a preset tiny positive threshold ϵ_0 or the maximal iteration

Algorithm 1 IGSS-Q Algorithm

```
**Input**: $a_1$, $a_2$, $L_{\max}$, $\epsilon_0$, $\rho = \frac{\sqrt{5}-1}{2}$.
1) $l = 0$, $a_1^{(0)} = a_1$, $a_2^{(0)} = a_2$, $d^{(0)} = a_2^{(0)} - a_1^{(0)}$, $x_1^{(0)} = a_1^{(0)} + (1-\rho)d^{(0)}$, $x_2^{(0)} = a_1^{(0)} + \rho d^{(0)}$; to 2).
2) $d^{(l)} = a_2^{(l)} - a_1^{(l)}$; If $l \leq L_{\max}$ & $|d^{(l)}| > \epsilon_0$, go to 3); else, go to 5).
3) Calculate $e(a_1^{(l)})$, $e(x_1^{(l)})$, $e(x_2^{(l)})$ and $e(a_2^{(l)})$ through (12); Then $[I_{\min}^{(l)}, e_{\min}] = \min\{e(a_1^{(0)}), e(x_1^{(0)}), e(x_2^{(0)}), e(a_2^{(0)})\}$, where $I_{\min}^{(l)}$ is the index value and $I_{\min}^{(l)} \in \{1,2,3,4\}$. Go to 4);
4) With the results in 3), update the values of $a_1^{(l)}$, $a_2^{(l)}$, $x_1^{(l)}$, $x_2^{(l)}$, and $l$, through the IGSS method in [12], (11) and (12). Go to 2);
5) $\nu_{\min} = \arg\min_{x_i^{(l)}}\{e(x_i^{(l)})\}, i = 1$ or 2, break.
6) Compute $\hat{q}_i$ via (11) and $\nu_{\min}$.
**Output**: $\nu_{\min}$, $\hat{\mathbf{q}} = [\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_N]^T$
```

times $L_{\max}$ is reached. The detailed process of the IGSS-Q method is provided in Algorithm 1.

The algorithm's computation complexity is relatively small, as it implements scalar quantization on top of an efficient one-dimensional search method. With the output $\hat{\mathbf{q}}$ from the algorithm, we can get the quantized precoder by

$$\hat{\mathbf{f}}_{\text{RF}} = \frac{\hat{\mathbf{q}}}{\|\hat{\mathbf{q}}\|}, \ \hat{\mathbf{q}} = [\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_N]^T. \tag{13}$$

IV. Analysis of Quantization Error and SNR Degradation

In this section, we first analyze the element-wise quantization error for two codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$. The mean squared quantization error (MSQE) used in [10] is inherited as the performance metric. Then, with the results of MSQE, we analyze the SNR degradation caused by quantization of RF precoders.

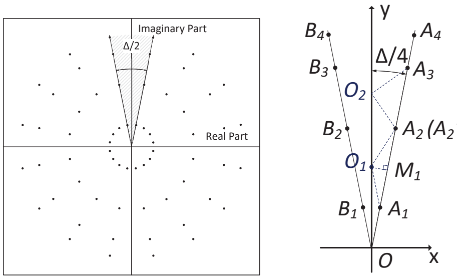
A. Quantization Error Analysis

The MSQE metric is defined as

$$\varepsilon = E[\frac{1}{N}\sum_{i=1}^{N}(|\nu_{\min} f_i - \hat{q}_i|^2)].$$

To derive a closed-form MSQE expression, we approximate the quantization error $|\delta_c|$ for each BF weight as a variable following a uniform distribution over $[0, \delta_{c,\max}]$. On the complex plane, $\delta_{c,\max}$ can be computed as the maximum of all the distances between any point $ae^{j\alpha}$ to its nearest constellation points $\hat{c}$. For simplicity, when analyzing $\delta_{c,\max}$, we only consider the case when $a \leq |\hat{c}|_{\max}$ since the probability of $a > |\hat{c}|_{\max}$ is low.

1) MSQE for Codebook $\mathcal{C}_1$: For the normalized codebook $\mathcal{C}_1$ with $\phi = 0$, $\delta_{c,\max}$ is the distance between $(0,0)$ to the nearest points, and is given by $\delta_{c,\max} = \frac{\sqrt{2-2\cos\Delta}}{h_1}$. Thus, the MSQE in this case is



(a) Constellation and the sector selected for analysis. (b) Key points in the sector.

Fig. 4. Constellation plot used for analyzing $d_{\max}$ for Codebook 2 ($b = 3$).

$$\varepsilon_{c1} = [E(|\delta_c|)]^2 + \text{Var}(|\delta_c|) = \frac{2-2\cos\Delta}{3h_1^2}. \tag{14}$$

2) MSQE for Codebook $\mathcal{C}_2$: Now we analyze $\delta_{c,\max}$ for codebook $\mathcal{C}_2$. Because the constellation of $\mathcal{C}_2$ is symmetry, we can select a circular segment (the shaded region in Fig. 4(a)) for analyzing $\delta_{c,\max}$.

In this segment, the points that can achieve the maximal distance between two adjacent constellation points are marked as $O_i, i = 1, 2, \cdots, 2^{b-1}$, as shown in Fig. 4(b). Obviously, every O_i locates on the y-axis. Note that for simplicity, in the following analysis, we temporarily let $h_2 = 1$, as its value is independent of the position relationship between constellation points. According to (4), it can be easily derived that for the constellation points A_i,

$$|A_iO| = r_{i-1} = \sqrt{2 - 2\cos[(i-1)\Delta + \frac{\Delta}{2}]}, \ i = 1, 2, \cdots \tag{15}$$

Assume that there exists O_1, making $|A_1O_1| = |A_2'O_1| = |A_1O| = r_0 = \sqrt{2-2\cos\frac{\Delta}{2}}$. Then $|A_2'O| = |OM_1| + |A_2M|$. According to the cosine rule, $|OM_1| = 2|A_1O|\cos^2(\frac{\Delta}{4})$. Thus, $|A_2'O| = |OM_1| + |A_2M| = |OM_1| + (|OM_1| - |A_1O|) = 4r_0\cos^2(\frac{\Delta}{4}) - r_0 = \sqrt{2 - 8\cos^3(\frac{\Delta}{2}) + 6\cos(\frac{\Delta}{2})} = |A_2O| = \sqrt{2-2\cos(\frac{3\Delta}{2})}$. Therefore, we can find that A_2' overlaps with A_2. This implies that in the sector A_2OB_2, $d \leq \delta_{c,\max} = r_0$.

According to the Cosine law, if $O_i(y_ie^{j2\pi})$ satisfies $|A_iO_i| = |A_{i+1}O_i|$, that is

$$r_i^2 + y_i^2 - 2r_iy_i\cos(\frac{\Delta}{4}) = r_{i+1}^2 + y_i^2 - 2r_{i+1}y_i\cos(\frac{\Delta}{4}).$$

Solving this equation, we get $y_i = \frac{r_{i+1}+r_i}{2\cos(\frac{\Delta}{4})}$. Therefore,

$$|A_iO_i|^2 = \frac{r_{i+1}^2 + r_i^2 - 2r_ir_{i+1}\cos(\frac{\Delta}{2})}{2\cos(\frac{\Delta}{2}) + 2}. \tag{16}$$

Assuming $|A_iO_i|^2 = r_0^2 = 2 - 2\cos(\frac{\Delta}{2})$, we have

$$r_{i+1}^2 + r_i^2 - 2r_ir_{i+1}\cos(\frac{\Delta}{2}) = 2 - 2\cos\Delta. \tag{17}$$

According to the Cosine law again, the left part of (17) can be seen as $|A_iB_{i+1}|$. Because of the symmetry of the constellation, $|A_iB_{i+1}|$ equals to the distance between two constellation points with similar positional relationship:

$$\begin{aligned} |A_iB_{i+1}| &= |(1+e^{j(\frac{\Delta}{2}+i\Delta)}) - (1+e^{j(\frac{\Delta}{2}+i\Delta+\Delta)})| \\ &= |e^{j\Delta}-1| = 2-2\cos\Delta. \end{aligned} \tag{18}$$

Therefore, $|A_iO_i|^2 = r_0^2$ is proven. In summary, for a $ae^{j\alpha}$ satisfying $a \leq r_{2^{b-1}}$, the maximal error distance $\delta_{c,\max} = r_0 = \frac{\sqrt{2-2\cos\frac{\Delta}{2}}}{h_2}$.

Thus, the MSQE in this case can be obtained as

$$\varepsilon_{c2} = \frac{2-2\cos\left(\frac{\Delta}{2}\right)}{3h_2}. \tag{19}$$

When Δ is small, $1-\cos\Delta \sim \frac{\Delta^2}{2}$ and $1-\cos\frac{\Delta}{2} \sim \frac{\Delta^2}{8}$, hence (14) and (19) can be approximated as

$$\varepsilon_{c1} \approx \frac{\Delta^2}{3h_1^2} = \frac{\Delta^2}{3(2+2^{2-b})M}, \tag{20}$$

and (21)

$$\varepsilon_{c2} \approx \frac{\Delta^2}{3h_2^2} = \frac{\Delta^2}{24M},$$

respectively.

From (8), we can find that for $b \geq 2$, $\sqrt{2M} < h_1 \leq \sqrt{3M}$. Therefore, ε_{c1} satisfies

$$\frac{\Delta^2}{9M} \leq \varepsilon_{c1} < \frac{\Delta^2}{6M}.$$

From [10], we can find the MSQE for using separate codebooks as: $\varepsilon_1 = \frac{\Delta^2}{6}$ and $\varepsilon_2 = \frac{(1+E[|f_i|^2])\Delta^2}{12}$, where ε_1 and ε_2 are the MSQE for the two options shown in Fig. 1. For large arrays with more than, e.g., $M = 8$ antennas, it can be readily verified that

$$\varepsilon_{c2} < \varepsilon_{c1} < \varepsilon_2 < \varepsilon_1. \tag{22}$$

This indicates that the proposed joint quantization method using the codebook $\mathcal{C}_2$ achieves the smallest quantization error.

B. SNR Degradation

Now we will compare the SNRs for the original BF vector $\mathbf{f}_{\text{RF}}$ and the quantized one $\hat{\mathbf{q}}$. For $\mathbf{f}_{\text{RF}}$, the SNR is given in (3). Assume $\hat{\mathbf{q}}_r = \nu_{\min}\mathbf{f}_{\text{RF}}^t + \mathbf{d}_{\text{RF}}^t$ and $\hat{\mathbf{q}}_t = \nu_{\min}\mathbf{f}_{\text{RF}}^t + \mathbf{d}_{\text{RF}}^t$, the overall SNR is given by

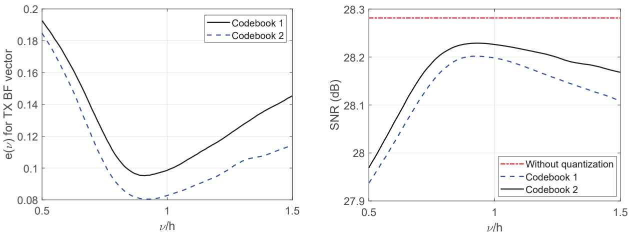
$$\text{SNR} = |\hat{\mathbf{q}}_r^H \mathbf{H} \hat{\mathbf{q}}_t^H|^2 \cdot \frac{\lambda_0^2\sigma_s^2}{\sigma_n^2},$$

where $\mathbf{d}_{\text{RF}}^t$ and $\mathbf{d}_{\text{RF}}^r$ are the quantization error vectors. Following [10], define the SNR degradation as

$$\mathcal{D} = \frac{\text{SNR}_0}{\text{SNR}} = \frac{\lambda_0^2}{\hat{\mathbf{q}}_r^H |\mathbf{H}| \hat{\mathbf{q}}_t^H}, \tag{23}$$

and the mean $\overline{\mathcal{D}}$ can be approximated as

$$\overline{\mathcal{D}} \approx 1 + E[\|\mathbf{d}_{\text{RF}}^t\|^2] + E[\|\mathbf{d}_{\text{RF}}^r\|^2]. \tag{24}$$



(a) $e(\nu)$ vs ν.

(b) Output SNR vs ν.

Fig. 5. Variation of $e(\nu)$ and the output SNR with ν when $b = 4$.

where $E[\|\mathbf{d}_{\text{RF}}^r\|^2] = N_r\varepsilon$ and $E[\|\mathbf{d}_{\text{RF}}^t\|^2] = N_t\varepsilon$, and ε can be obtained through (14) and (19) for $\mathcal{C}_1$ and $\mathcal{C}_2$. From (22), it can be found that

$$\overline{\mathcal{D}}_1 > \overline{\mathcal{D}}_2 > \overline{\mathcal{D}}_{c1} > \overline{\mathcal{D}}_{c2} \tag{25}$$

V. Simulation Results

We provide simulation results to verify the analytical results in this section. Consider the system where ULAs with $N_t = N_r = 16$ omnidirectional antennas ($d = \lambda/2$) are used for the transmitter and receiver. Note that our methods can also be applied to non-uniform antenna arrays. Assume there is a dominating LOS multipath between the transmitting and receiving nodes. All the $L = 8$ multipaths are uniformly distributed within a direction range of 14 degrees centered at the LOS direction. The mean power ratio between the LOS and the rest multipath signals is 10dB. The AoD of the LOS multipath is assumed to be uniformly distributed over $[-60^\circ, 60^\circ]$. In this paper, the results are averaged over 10^5 realizations.

Fig. 5 displays the relationship between the quantization error function $e(\nu)$ in (12) and the output SNR when $b = 4$, where $e(\nu)$ is calculated for the beamforming vector $\mathbf{f}_{RF}^t$. By comparing Fig. 5(a) and Fig. 5(b), we can see that $e(\nu)$ and the output SNR vary consistently, in the opposite direction, with the scaling value ν. This validates the reliability of the searching method used in this paper.

In Fig. 6, the output SNR for different quantization methods are presented. For comparison, we also present results for the fast block noncoherent decoding (FBND) method [5], denoted as "FBND", which is a low-complexity 1-PS vector quantization method. The method in [10] using the serial structure is denoted as "Lin2017-S2" in the legend. "Codebook 1" and "Codebook 2" denote the joint quantization method using the initial $\hat{\mathbf{f}}_{\text{RF}}$ with codebook $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, without searching for the optimal ν, while "Codebook 1-IGSS" and "Codebook 2-IGSS" represent their counterparts employing the proposed IGSS-Q searching algorithm. For the IGSS-Q method, $\epsilon_0 = 10^{-6}$ and $L_{\max} = 100$ is chosen. The figure shows that the joint quantization method proposed in this paper can significantly improve the SNR compared with the method in [10]. We can also observe that: 1) $\mathcal{C}_2$ can obtain larger SNR than $\mathcal{C}_1$ since $n_{c_2} = 2n_{c_1}$; 2) The IGSS-Q algorithm can lead to an improved SNR , at the cost of

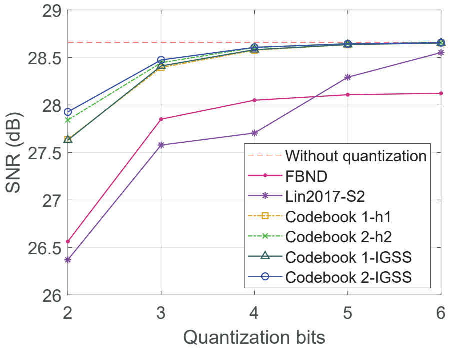


Fig. 6. Output SNR via quantization bits.

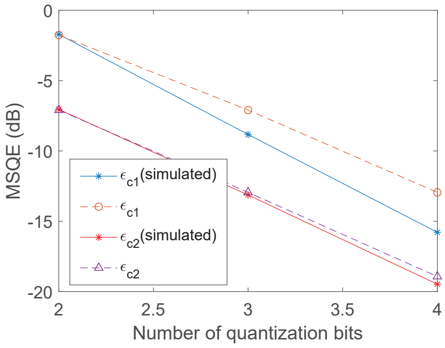


Fig. 7. MSQE versus number of quantization bits.

increased computation complexity (around 0.01s per iteration in MATLAB®). The gap is quite small when the number of quantization bits is larger than 3.

Fig. 7 demonstrates how MSQE varies with the number of quantization bits for the proposed joint quantization methods. The values of $f_i = |f_i|e^{j\psi_i}$ in $\mathbf{f}_{\text{RF}}^t$ are generated randomly with $|f_i|$ following a uniform distribution over $[0, 2)$ and with ψ_i following a uniform distribution over $[0, 2\pi)$. The vector of $\mathbf{f}_{\text{RF}}^t$ is then normalized so that its norm is 1. When $b > 2$, the simulated ε_{c1} deviates from the analytical ε_{c1} derived in Section IV-A1. This is because, for Codebook 1, many of the largest distances between any two nearest constellation points are smaller than $\delta_{c,\max}$. Therefore, the uniform distribution assumption in Section IV-A2 is not accurate enough. This may be improved via reducing $\delta_{c,\max}$ to the second largest value for Codebook 1.

In Fig. 8, we show the simulated $\overline{\mathcal{D}}$ and the analytical ones in (24) for the codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$. In accordance with (24) and (25), a smaller $\overline{\mathcal{D}}$ is achieved by codebook $\mathcal{C}_2$.

VI. Conclusions

We presented a novel and highly effective joint quantization method for hybrid arrays with the two-phase-shifter structure. By combining the codebooks of the two phase shifters, we developed the element-wise quantization method for this combined codebook, as well as an iterative IGSS-Q method for refining the normalization factor for the BF vector before quantization. Both methods are shown to achieve negligible performance loss compared to the non-quantized one in terms of BF gain, when the number of quantization bits is larger than 3. We also provided quantitative analysis for the quantization error and the degradation in SNR due to quantization, which are validated by simulation results. The work in this paper can be further improved by combining the baseband precoding [7] and vector quantization methods [13].

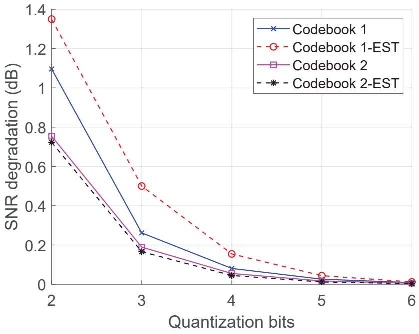


Fig. 8. Mean SNR degradation $\overline{\mathcal{D}}$ versus quantization bits.

References

[1] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave mimo systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[2] J. A. Zhang, X. Huang, V. Dyadyuk, and Y. J. Guo, "Massive hybrid antenna array for millimeter-wave cellular communications," *IEEE Wireless Communications*, vol. 22, no. 1, pp. 79–87, 2015.

[3] D. J. Ryan, I. V. L. Clarkson, I. B. Collings, D. Guo, and M. L. Honig, "QAM and PSK codebooks for limited feedback MIMO beamforming," *IEEE Transactions on Communications*, vol. 57, no. 4, 2009.

[4] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," vol. 61, no. 12, pp. 5016–5029, 2013.

[5] W. Sweldens, "Fast block noncoherent decoding," vol. 5, no. 4, pp. 132–134, 2001.

[6] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[7] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE transactions on wireless communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[8] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave mimo systems with large antenna arrays," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 998–1009, 2016.

[9] W. Jiang, Y. Guo, T. Liu, W. Shen, and W. Cao, "Comparison of random phasing methods for reducing beam pointing errors in phased array," *IEEE transactions on antennas and propagation*, vol. 51, no. 4, pp. 782–787, 2003.

[10] Y.-P. Lin, "On the quantization of phase shifters for hybrid precoding systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2237–2246, 2017.

[11] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by rf-baseband codesign for mimo systems," in *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*. IEEE, 2014, pp. 1–5.

[12] E. Höpfinger, "On the solution of the unidimensional local minimization problem," *Journal of Optimization Theory and Applications*, vol. 18, no. 3, pp. 425–428, 1976.

[13] D. J. Ryan, I. B. Collings, and I. V. L. Clarkson, "Glrt-optimal noncoherent lattice decoding," *IEEE transactions on signal processing*, vol. 55, no. 7, pp. 3773–3786, 2007.